Computational and Structural Biotechnology Journal 18 (2020) 1539–1547

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

HiPR: High-throughput probabilistic RNA structure inference

Pavel P. Kuksa [a,1], Fan Li [d,1], Sampath Kannan [b], Brian D. Gregory [c], Yuk Yee Leung [a], Li-San Wang [a,b,*]

[a] Penn Neurodegeneration Genomics Center, Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
[b] Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104, USA
[c] Department of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA
[d] Children's Hospital Los Angeles, Los Angeles, CA 90027, USA

ARTICLE INFO

Article history:
Received 28 February 2020
Received in revised form 15 May 2020
Accepted 1 June 2020
Available online 8 June 2020

Keywords:
High-throughput structure-sensitive sequencing
RNA structure inference
Probabilistic modeling
DMS-seq
DMS-MaPseq

ABSTRACT

Recent high-throughput structure-sensitive genome-wide sequencing-based assays have enabled large-scale studies of RNA structure, and robust transcriptome-wide computational prediction of individual RNA structures across RNA classes from these assays has potential to further improve the prediction accuracy. Here, we describe HiPR, a novel method for RNA structure prediction at single-nucleotide resolution that combines high-throughput structure probing data (DMS-seq, DMS-MaPseq) with a novel probabilistic folding algorithm. On validation data spanning a variety of RNA classes, HiPR often increases accuracy for predicting RNA structures, giving researchers new tools to study RNA structure.

1. Introduction

High-throughput characterization of RNA structures holds great potential for basic science and clinical applications. Recent structure-mapping methods combining biochemical assays with high-throughput-sequencing have enabled studies of RNA secondary structure on a genome-wide scale [7,12,17,25,34,40–45,48,50] and have revealed many functional and regulatory roles for RNA structure [7,34,38]. However, many of these analyses report on the global landscape of structured and unstructured regions in RNA. The utility of these high-throughput datasets in determining individual RNA structures at *single-base resolution* is relatively unexplored on a genome-wide scale. At the same time, effective and robust transcriptome-wide computational prediction of individual RNA structures across RNA classes from the high-throughput experimental assays has potential to further improve the prediction accuracy [22,27,29,30,31,34,40,45,49,50].

RNA structure prediction is often performed *in silico* by sequence-only methods based on empirical thermodynamic parameters and the minimum free energy (MFE) model [23,24,32,51], with comparative sequence-analysis methods based on the alignment of homologous RNA sequences from different organisms [11,28] providing some of the most accurate results.

To determine native RNA structures *in vivo* and improve accuracy compared to *in silico* sequence-only methods, another group of approaches [4,5,20,22,43,45,47] combine thermodynamics (MFE model) and experimental observations encoded as structural constraints (*constrained folding* approaches). These approaches include pseudo-energy methods that add pseudo-energy terms to the energy function to penalize or encourage base pairing [20,22] based on experimental observations. More recent methods minimize the discrepancy between predicted MFE structure and experimental observations by introducing new objective functions with additional error terms [43,45].

Previous studies showed these methods improve structure prediction in many cases [5,13,30,31,39,40,44,45]. Nonetheless, accurate and robust genome-wide structure prediction across diverse RNA classes is critical in the presence of noisy, sparse genome-wide measurements and experimental biases.

To address these challenges, we describe HiPR (High-throughput Probabilistic RNA structure inference), a novel method based on a probabilistic model for experimental observations in structure probing data (e.g., DMS-Seq [34], or DMS-MaPseq [50], and Markov Chain Monte Carlo (MCMC) sampling for predicting RNA secondary structure and base-pairing probabilities from these experimental observations (Materials and methods, Figs. 1, 2,

* Corresponding author at: Penn Neurodegeneration Genomics Center, Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA.
E-mail address: lswang@pennmedicine.upenn.edu (L.-S. Wang).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.csbj.2020.06.004
2001-0370/

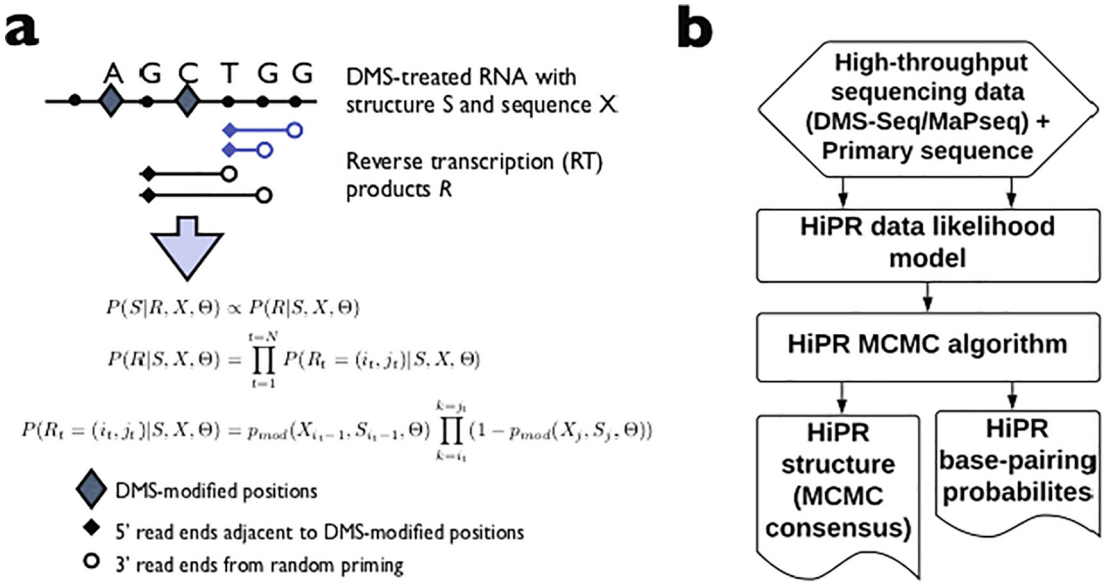

Fig. 1. Overview of structure prediction with HiPR. (a) DMS-seq and HiPR probabilistic structure and data model. HiPR assumes observed sequencing reads ***R*** corresponding to reverse transcriptase (RT) termination events and DMS-modified positions are independent events such that the overall probability of the sequencing data is a product of individual read probabilities. Corresponding HiPR model for DMS-MaPseq data is shown in Fig. 2. (b) RNA structure prediction with HiPR. HiPR RNA structure and base-pairing estimates are obtained by 1) defining a likelihood score for a structure given experimental observations and 2) using MCMC to find RNA structure maximizing the likelihood score.

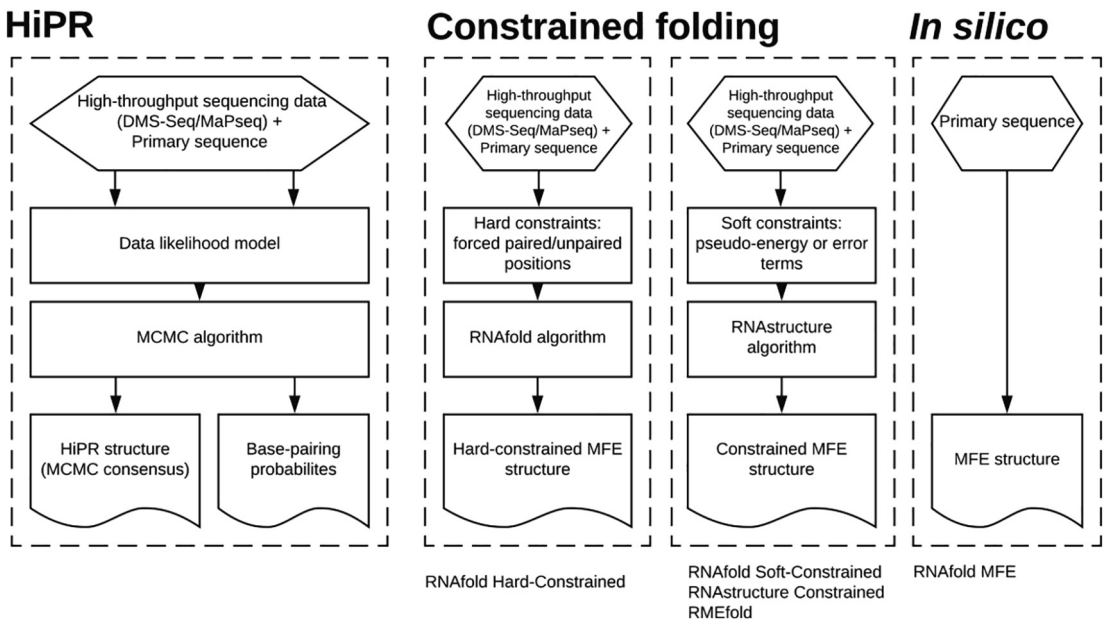

Fig. 2. Overview of strategies for *in silico* and experimental data-aided prediction of RNA structures. For each strategy, examples of the particular algorithms implementing it are listed underneath the corresponding diagram.

Supplementary Website https://www.lisanwanglab.org/HIPR; Supplementary Software https://github.com/wanglab-upenn/HiPR).

To develop and illustrate the HiPR approach, we focus on two high-throughput sequencing-based protocols for probing RNA structures, *in vivo* DMS-seq [34](reverse transcription termination-based) and DMS-MaPseq[50] (mutation-based). In the DMS-based structure probing protocol, dimethyl sulfate (DMS) methylates *in vivo* unpaired adenosine (A) and cytosine (C) residues, which causes premature termination during the following reverse transcription (RT) step in sequencing library preparation. Consequently, the 5′ ends of all cloned fragments (RT products) will fall immediately downstream of unpaired A and C nucleotides (Fig. 1a). On the other hand, the 3′ ends result from *random fragmentation* and are therefore randomly distributed. A subsequent size selection step enriches for fragments that represent premature termination and these captured fragments are then amplified and sequenced. As a result, 5′ end and all other positions in the RT products carry information on the pairing status for each nucleotide (Supplementary Note).

On the other hand, in the DMS-MaPseq structure probing protocol [50], DMS modifications of unpaired A and C residues induce RT misincorporations which are observed as mutations in the

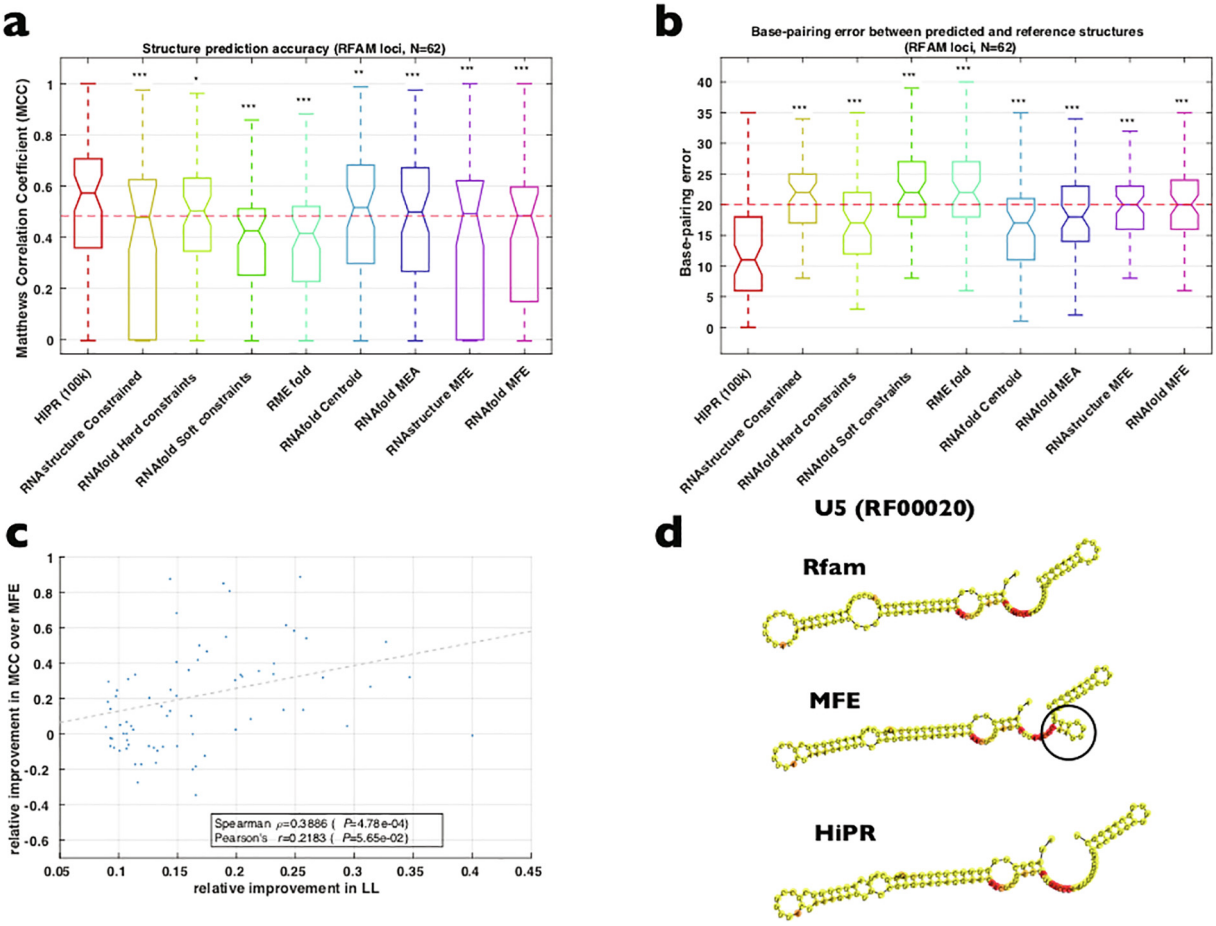

Fig. 3. RNA structure prediction with HiPR. (a) Overall structure prediction accuracy on the Rfam dataset by HiPR and other methods measured by Matthews Correlation Coefficient (MCC). DMS-Seq data (human K562 cell line) is used by HiPR and constrained folding methods that use experimental data. The *p*-values are calculated by a one-sided Wilcoxon paired signed rank test for HiPR and each of the methods (*, **, *** denote $p < .05$, $p < .01$, $p < .001$, respectively). Red dotted line corresponds to the baseline RNAfold MFE prediction. (b) Base-pairing error between predicted and reference structures for HiPR and other methods on Rfam dataset. Red dotted line corresponds to the baseline RNAfold MFE prediction. (c) HiPR likelihood scores correlate with the accuracy of structure prediction. Improvement in log-likelihood (LL) is associated with an improvement in structure accuracy (MCC). (d) An example RNA secondary structure (U5 spliceosomal RNA chr15:65588389–65588504 GRCh37/hg19). HiPR predicts structure with higher PPV and greater agreement with the structure determined by comparative sequence analysis (Rfam). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

sequencing reads [50] (Fig. 4). Compared to DMS-seq, multiple modification events can be captured in each DMS-MaPseq sequencing read.

2. Overview of HiPR method

In contrast to the previous approaches [5,29,34,43,45], HiPR explicitly models the generation of all possible reverse transcriptase (RT) fragments (observations, $\boldsymbol{R}$) from DMS-treated RNAs as a function of the underlying RNA secondary structure $\boldsymbol{S}$ and experimental conditions (model parameters, $\boldsymbol{\Theta}$) including per-nucleotide DMS modification rates, size selection, and nucleotide biases (Figs. 1a, b, 2; Materials and methods).

The HiPR structure likelihood model $L(\boldsymbol{S}, \boldsymbol{\Theta}|\boldsymbol{R})$ $P(\boldsymbol{R}|\boldsymbol{S}, \boldsymbol{\Theta})$ assumes a probabilistic process underlying DMS modifications along the RNA structure $\boldsymbol{S}$ and resulting sequence fragments $\boldsymbol{R}$ produced by the reverse transcription of these DMS-modified RNAs (Figs. 1a, 2; Materials and methods).

We use a Metropolis-Hastings random walk in the structure-parameter space $(\boldsymbol{S}, \boldsymbol{\Theta})$ to find structures maximizing the posterior $argmax_{\boldsymbol{S},\boldsymbol{\Theta}} P(\boldsymbol{S}, \boldsymbol{\Theta}|\boldsymbol{R})$. A set of basic structure editing operations is used to sample structures from the posterior distribution given the sequencing readout (Methods; Table 1).

We then use sampled structures to estimate posterior probabilities for base-pairing interactions and to construct a single consensus MCMC posterior structure (i.e. a point estimate) that serves as an output of our structure prediction method (Fig. 1b; Algorithm 1; Materials and methods).

To validate the underlying probabilistic model in HiPR, we analyzed structure probing data sets (Materials and methods and Supplementary Tables S1, S2) from *in vivo* DMS-seq [34] (reverse transcriptase (RT) stop-based assay; Fig. 1a) and DMS-MaPseq [50] (mutation-based assay; Fig. 2) using a collection of reference secondary structure models spanning a variety of RNA classes (Methods; Supplementary Figure S1). The first set of reference RNA secondary structure models used to evaluate structure prediction accuracy comprises comparative sequence analysis-based secondary structures from the Rfam database [28]('Rfam collection'). The second set of structures contains reference ribosomal RNA structures from RNA STRAND and CRW databases [1,3].

3. RNA structure prediction with HiPR

To analyze how well HiPR predicts RNA structures, we tested HiPR and existing methods (Table 2) including both *in silico* and *in vivo* experimental data-based strategies (Fig. 2).

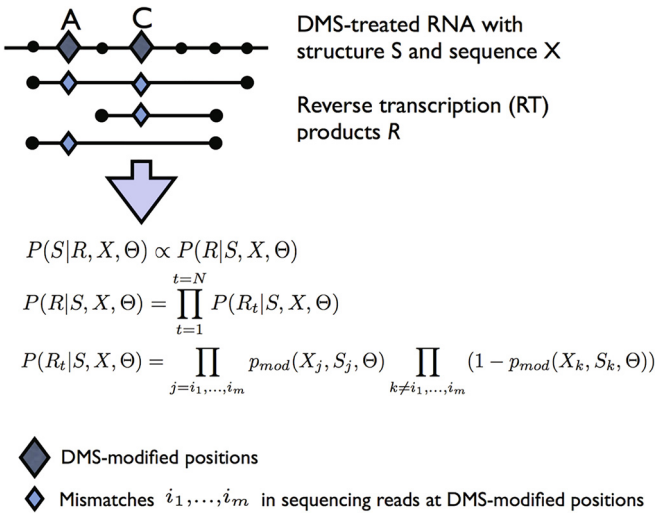

Fig. 4. DMS-MaPseq and HiPR probabilistic structure and data model. Observed sequencing reads R encode DMS-modified positions at unpaired A or C as mismatches relative to the reference RNA sequence. HiPR assumes reads are independent events such that the overall probability $\boldsymbol{P}(\boldsymbol{R}|\boldsymbol{S},\boldsymbol{X},\Theta)$ is a product of individual read probabilities.

We first examined 168 loci (Rfam[28]overlapping DMS-seq data (Supplementary Table S1, Materials and methods; Supplementary Methods) for human K562 cell line (Fig. 3a, b). In the second set of experiments (Section 'Structure prediction accuracy with DMS-MaPseq data'), we overlapped Rfam loci with DMS-MaPseq data (Materials and methods; Supplementary Methods).

3.1. Better HiPR likelihood scores are associated with more accurate structures

Using the *in vivo* DMS-seq data, we first observe that the magnitude of relative improvements in the data likelihood $\frac{P(R|S_{final})-P(R|S_{start})}{P(R|S_{start})}$ between the final structure S_{final} and the starting structure S_{start} by MCMC iterations correlated positively with relative improvements in the structure prediction accuracy (Matthews Correlation Coefficient, MCC) $\frac{\left[MCC_{S_{final}}-MCC_{S_{MFE}}\right]}{MCC_{S_{MFE}}}$ over MFE structure S_{MFE} across structures in Rfam (Fig. 3c).

This reinforces our idea of using MCMC sampling to optimize the data likelihood $P(R|S)$ in order to refine the initial estimate of the structure (S_{start}) and find a more accurate RNA structure (S_{final}) that best fits the observed sequencing data.

3.2. Structure prediction using DMS-seq data

We calculated the accuracy of the structures predicted by each of the methods (Table 2) using commonly used metrics [5,6,29,30] (Materials and methods; Supplementary Methods): Matthews Correlation coefficient (MCC) (Fig. 3a), base-pairing error (Fig. 3b), precision and sensitivity (Supplementary Fig. S2).

We assessed structure prediction accuracy using several different metrics for HiPR and other structure prediction methods including MFE, centroid, hard- and soft- constrained structures (Fig. 3a, b, and d; Supplementary Fig. S2; Supplementary Table S3). For Rfam RNA structures with sufficient sequencing coverage [34] ($\geq$ 15 read stops at A or C on average along RNA structure; Materials and methods), the improvement in the average structure prediction accuracy (MCC) was 23.85% relative to MFE and other methods (Fig. 3b). Similarly, the reduction in the average base-pairing error was 31.3% relative to MFE and other methods (Fig. 3a). Improvement in positive predictive value (PPV) was 48.94% relative to MFE and centroid, while sensitivity reduced slightly by 3.37% and 6.21% on average relative to MFE and centroid, respectively (Supplementary Fig. S2). Fig. 3d, Supplementary Figs. S4, S5 show examples of reference and predicted RNA structures [10,28].

3.2.1. Accuracy on individual classes of RNA

Specific non-coding RNA types are defined by the presence of structural motifs [8] and could therefore bias our results. To investigate any potential overconfidence in our approach, we further tested HiPR performance on individual RNA classes (Supplementary Table S3). We found the improvement in structure prediction accuracy by HiPR varies by RNA classes but is generally quite robust. For instance, the reduction in average base-pairing error was 28.86% for non-C/D box RNAs across MFE and other methods. HiPR improved average positive predictive value (PPV) by 15.48% relative to MFE and centroid, while sensitivity was reduced by 2.48% relative to MFE, and 9.47% relative to centroid.

3.2.2. HiPR accuracy improves with sequencing depth

Across RNA structures in Rfam collection, the number of RT stops at A or C was strongly correlated with the reduction in base-pairing error (Spearmanρ = 0.4819, P < 7.34e−05). The improvement in prediction accuracy (MCC) was similarly strongly correlated with the average number of RT stops per A or C position (Spearmanρ = 0.3919, P < .0018). Power analysis (Supplementary Fig. S3) also shows an increase in accuracy with the number of sequencing reads available for each locus.

3.3. Structure prediction using DMS-MaPseq data

Using *in vivo* DMS-MaPseq data sets (Supplementary Table S2), we compared accuracy of HiPR with *in silico* and DMS-MaPseq con-

Table 1
Metropolis-Hastings move set.

Move	Example	Parameter update	Constraints
Add a pairing interaction	...((.....))... ↓ .(.((.....)).).	$\boldsymbol{S} \rightarrow \boldsymbol{S}^*$	Must be a valid base pair, follow steric hindrance rules, and result in a fully nested structure.
Delete a pairing interaction	...(((.....)))... ↓ ...((.....)))....	$\boldsymbol{S} \rightarrow \boldsymbol{S}^*$	None
Select a new per-nucleotide modification rate $r^* = r + \epsilon$ $\epsilon\ U(-0.01, 0.01)$	$\{u_A, u_C, u_T, u_G\}$ ↓ $\{u_A + 0.006, u_C - 0.002, u_T - 0.005, u_G + 0.001\}$	$\boldsymbol{u} \rightarrow \boldsymbol{u}^*$ $\boldsymbol{v} \rightarrow \boldsymbol{v}^*$	All rate parameters must be in $[0,1]$,$u_x < v_x$

Table 2
RNA structure prediction methods.

Software	Strategy for experimental data	Strategy for structure prediction	Reference
HiPR	Probabilistic model	MCMC posterior	This work
RNAfold	–	Minimum free energy (MFE)	[18]
RNAfold	–	Maximum Expected Accuracy (MEA)	[21]
RNAfold	–	Centroid	[18]
RNAstructure	–	MFE	[32]
RNAstructure	Pseudo-energy	MFE	[5,32,46]
RNAfold	D/Z/W algorithms	D/Z/W	[5,19,43,47]
	W (soft-constraints)	MFE	[43]
	Hard constraints	MFE	[51]
RMEfold	pseudo-energy posterior	MEA	[45]

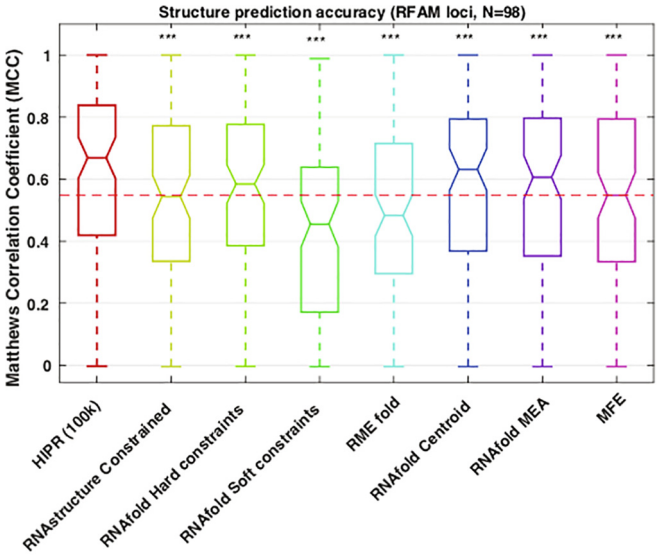

Fig. 5. Overall structure prediction accuracy using DMS-MaPseq data on the Rfam dataset by HiPR and other methods. Structure prediction accuracy is measured by Matthews Correlation Coefficient (MCC). The *p*-values are calculated by a one-sided Wilcoxon paired signed rank test for HiPR and each of the methods (*, **, *** denote $p < .05$, $p < .01$, $p < .001$, respectively).

strained folding methods (Fig. 5). For Rfam loci with sufficient DMS-MaPseq sequencing coverage (>20× mismatch coverage; Supplementary Methods), HiPR improved the structure prediction accuracy (MCC) compared to *in silico* MFE, centroid, and constrained-folding methods, with the average per-structure accuracy improvement of 14.5% relative to hard-constrained and 20% relative to soft-constrained methods.

3.4. HiPR predictions on curated structures

When applied to the challenging 18S ribosomal RNA structure, HiPR showed on both DMS-seq and DMS-MaPseq data higher overall accuracy and significantly lower false positive rates compared to *in silico* and *in vivo* constrained folding methods (DMS-seq, Supplementary Tables S4; DMS-MaPseq, Supplementary Tables S5 and Supplementary Fig. S6), with a 22% improvement in the structure accuracy (MCC) and a 54% reduction in false positives (FPR) on average over *in silico* and DMS-seq constrained folding methods (Supplementary Tables S4).

For human 5S rRNA (Supplementary Tables S6) using DMS-seq, HiPR predicted more accurate structure compared to both *in silico* and constrained folding methods (40% average increase in MCC, 58% reduction in false positives).

4. Discussion

Overall, as evidenced by application to DMS-seq and DMS-MaPseq structure probing data and analysis of Rfam non-coding RNA structures, the novel *probabilistic modeling* approach undertaken by HiPR often increases structure prediction accuracy (Figs. 3a, b, 5; Supplementary Fig. S2).

HiPR offers five main advantages. First, HiPR introduces an alternative approach to predicting RNA structure by using probabilistic modeling of the experimental structural data (Figs. 1, 2): this is achieved by 1) defining a likelihood score for a structure given experimental observations and 2) using MCMC to optimize RNA structure with respect to the likelihood score. Second, HiPR approach is not limited to a particular protocol (DMS-seq or DMS-MaPseq) and can accommodate other structure probing assays (Supplementary Note) by virtue of a modular MCMC framework (such as Structure-seq [7,33]. Third, HiPR overcomes inherent experimental biases (e.g., preferential modification of A or C nucleotides by DMS) by joint modeling of both paired bases and all four unpaired bases (cf. original DMS-Seq [34]. Fourth, HiPR base-pairing probabilities and likelihood scores may be used in many downstream analysis steps such as conservation analysis, analysis of structural motifs and substructures, or as constraints for other structure analysis methods. Fifth, HiPR posterior likelihood on a particular RNA sequence also gives us an idea of whether the method has worked well (Fig. 3c) and thereby we can compute our confidence in the output of HiPR. These advantages allow HiPR to be readily used for examining native, *in vivo* RNA structures using high-throughput structure-probing sequencing data, and enabling genome-wide or targeted RNA studies of structures and their dynamics in development, along viral infection, investigating diversity of structures across biological conditions, or structural effects of mutations [2,15,26,36,50].

We observe that the HiPR algorithm predicts fewer (**Supplementary Fig. S2b**) yet more accurate base-pairing interactions (Supplementary Fig. S2a). While the current implementation of the HiPR algorithm does not directly support integration of the data from control samples [12], control experiments can be incorporated into HiPR, e.g., by using position specific modification rates derived from control and probing experiments. We also note that HiPR candidate structure generation currently does not allow pseudo-knots. It is also important to note that protocols where only 5′ position in the read is informative (e.g., PARS [12] may limit ability of HiPR to infer additional structural information from sequencing read-out (Supplementary Note 1).

Although HiPR often achieves higher accuracy even with limited sequencing depth, our power analyses (Supplementary Fig. 3) indicate that further improvements in the quality and coverage of sequencing-based structural assays would likely lead to even higher accuracy of RNA structure prediction. An ensemble method combining data from multiple experimental protocols in our probabilistic modeling framework could further improve the accuracy of RNA structure prediction. We also note that choice of appropriate method or methods to be used in a particular study can be guided by the type of experimental data available and desirable structure prediction strategies (Table 2 and Fig. 2).

In summary, we find that the probabilistic modeling of the data from recent high-throughput structure-sensitive assays [34,50] improves the accuracy of RNA structure prediction. Application of this approach to the growing body of structure probing data [7,15,33,35,37,40], along with further experimental validations,

could lead to significant insights into the fundamental roles of this important RNA feature.

5. Materials and methods

5.1. Data preprocessing

DMS-seq and DMS-MaPseq data preprocessing are described in Supplementary Methods.

5.2. HiPR algorithm for RNA structure prediction

HiPR (High-throughput Probabilistic RNA structure inference) algorithm for predicting RNA secondary structure and base-pairing probabilities using experimental data from high-throughput structure-probing assays is based on the Bayesian Markov Chain Monte Carlo (MCMC) method.

Generally, in a Bayesian framework, we are interested in finding parameters θ such that their posterior probability $P(\theta|\boldsymbol{D})$ given observed data $\boldsymbol{D}$ is maximized. Using Bayes' rule,

$$P(\theta|\boldsymbol{D}) \propto P(\boldsymbol{D}|\theta)P(\theta)$$

we will instead optimize the likelihood function $P(\boldsymbol{D}|\theta)$ as it is proportional to our desired quantity. In the context of RNA secondary structure and our experimental structure probing data (e.g., DMS-seq, or DMS-MaPseq sequencing), the above equation can be rewritten as

$P(\boldsymbol{S},\boldsymbol{u},\boldsymbol{v},r_{min},r_{max}|\boldsymbol{X},\boldsymbol{R}) \propto P(\boldsymbol{R}|\boldsymbol{S},\boldsymbol{X},\boldsymbol{u},\boldsymbol{v},r_{min},r_{max})P(\boldsymbol{S},\boldsymbol{X},\boldsymbol{u},\boldsymbol{v},r_{min},r_{max})$ for DMS-seq and as

$P(\boldsymbol{S},\boldsymbol{u},\boldsymbol{v}|\boldsymbol{X},\boldsymbol{R}) \propto P(\boldsymbol{R}|\boldsymbol{S},\boldsymbol{X},\boldsymbol{u},\boldsymbol{v})P(\boldsymbol{S},\boldsymbol{X},\boldsymbol{u},\boldsymbol{v})$ for DMS-MaPseq, where $\boldsymbol{X}$ is the primary RNA sequence of length l, $\boldsymbol{S}$ is a secondary structure, $\boldsymbol{u}$ and $\boldsymbol{v}$ are per-nucleotide modification rates for paired and unpaired bases, r_{min} is the minimum cloneable fragment size, r_{max} is the maximum cloneable fragment size, and $R = \{R_k\}, k = 1,\ldots,m$ is our sequencing data (a set of observed sequence reads R_k). Our task is then to find a set of parameters $\{\boldsymbol{S},\boldsymbol{u},\boldsymbol{v},r_{min},r_{max}\}$, including the desired secondary structure $\boldsymbol{S}$, that maximizes the likelihood function $P(\boldsymbol{R}|\boldsymbol{S},\boldsymbol{X},\boldsymbol{u},\boldsymbol{v},r_{min},r_{max})$. For prior $P(\boldsymbol{S},\boldsymbol{X},\boldsymbol{u},\boldsymbol{v},r_{min},r_{max})$, we assume that all structures satisfying base-pairing rules have the same prior probability (see Table 1 for rules for generating plausible structures). The more likely structures will have higher posterior probabilities based on the their likelihood function given the observed read-out $\boldsymbol{R}$ and thus are more likely to be visited by our MCMC search (probabilities of all structures that violate Watson-Crick base-pairing rules are set to zero; see also Table 1 for candidate structure-generating rules). Experimental parameters $r_{min}, r_{max}$for DMS-Seq are set according to experimental protocol and are used in data likelihood computations to restrict range of possible read lengths to $[r_{min}, r_{max}]$.

At any given locus with the sequence $\boldsymbol{X}$ and structure $\boldsymbol{S}$, we assume that the observed sequencing reads $\boldsymbol{R}$ (cloned sequence fragments) are generated independently such that the overall probability of the sequencing data $\boldsymbol{R}$ is a product of individual read probabilities:

$$P(\boldsymbol{R}|\boldsymbol{S},\boldsymbol{X},\boldsymbol{u},\boldsymbol{v},r_{min},r_{max}) = \prod_{k=1}^{m} P(R_{\boldsymbol{k}}|\boldsymbol{S},\boldsymbol{X},\boldsymbol{u},\boldsymbol{v},r_{min},r_{max})$$

For RT-stop based read-out (DMS-seq), probability of a sequencing read R with endpoints $[i,j]$ is proportional to

$$P(R|\boldsymbol{S},\boldsymbol{X},\boldsymbol{u},\boldsymbol{v},r_{min},r_{max})\ Z_{i-1}(X_{i-1}) \prod_{k=i}^{j} (1 - Z_k(X_k))$$

where

$$Z_{\boldsymbol{i}}(\boldsymbol{X}_i) = \begin{cases} v(\boldsymbol{X}_i), & \textit{if base } i \textit{ is unpaired} \\ u(\boldsymbol{X}_i), & \textit{if base } i \textit{ is paired} \end{cases}$$

is the probability of DMS modification of the ith nucleotide X_i.

Intuitively, this read probability is equivalent to observing modification immediately upstream of the 5′ endpoint of the read fragment $[i,j]$, and not observing modification anywhere along the length of the read up to the last position.

To make these terms proper probabilities, these terms are normalized by the sum $\sum_{\substack{ij \\ r_{min} \le j-i+1 \le r_{max}}} P(R = [i,j]|\boldsymbol{S},\boldsymbol{X},\boldsymbol{u},\boldsymbol{v},r_{min},r_{max})$ over all possible reads $[i,j]$ with the length restricted to $[r_{min}, r_{max}]$ interval (probabilities of all reads outside the length interval are set to zero).

Similarly, for a mutation-based read-out (DMS-MaPseq), probability of a read R_k with m mismatches at positions $i_1,\cdots,i_m$ is defined as

$$P(R_k|S,X,u,v) \sim \prod_{i=i_1,\ldots,i_m} Z_i(X_i) \prod_{j\neq i_1,\ldots,i_m} (1 - Z_j(X_j))$$

This read probability for a mutation-based readout (DMS-MaPseq) is a product over positions with mismatches (the first term) and all other positions matching reference sequence (the second term).

5.3. Markov chain Monte Carlo (MCMC) for structure estimation

Using the Bayesian structure and data modeling framework described above, we now turn to the task of optimizing the likelihood function $P(\boldsymbol{R}|\boldsymbol{S},\boldsymbol{X},\Theta)$ by a random walk Metropolis-Hastings algorithm on the parameter space $\{\boldsymbol{S},\boldsymbol{u},\boldsymbol{v},r_{min},r_{max}\}$ ($\{\boldsymbol{S},\boldsymbol{u},\boldsymbol{v}\}$ for DMS-MaPseq). To accomplish this, we define a move set $\boldsymbol{M}$ that simply and comprehensively explores the entire parameter space (Table 1). Steric hindrance is ensured by enforcing minimum loop size (a parameter, set by default to 3nts or more) during candidate structure generation by checking and rejecting candidate structures violating the minimum loop size.

The ratio of the proposal density between two structures $\boldsymbol{S}$ and $\boldsymbol{S}^*$ is given by:

$$\frac{Q(\boldsymbol{S}|\boldsymbol{S}^*)}{Q(\boldsymbol{S}^*|\boldsymbol{S})} = \frac{n(\boldsymbol{M}(\boldsymbol{x}\rightarrow\boldsymbol{S}))}{n(\boldsymbol{M}(\boldsymbol{x}\rightarrow\boldsymbol{S}^*))}$$

where $n(\boldsymbol{M}(\boldsymbol{x}\rightarrow\boldsymbol{s}))$ is the number of valid structures that can yield structure $\boldsymbol{s}$ in a single move. Note however that the symmetrical nature of the move set $\boldsymbol{M}$ allows us to calculate $n(\boldsymbol{M}(\boldsymbol{x}\rightarrow\boldsymbol{S}))$ simply as the number of valid moves from the structure $\boldsymbol{S}$.

Taken together, we have the following implementation of the Metropolis-Hastings algorithm for sampling structures and experimental parameters from their posterior distribution:

Algorithm 1.. *HiPR algorithm for structure prediction. Steps* 1–8 *implement MCMC algorithm* (*Metropolis-Hastings random walk*). *The algorithm alternates between generating candidate structures in Steps* 2–4 *and updating experimental variables in Steps* 5–7. *Steps* 9 *and* 10 *compute final posterior estimates for the structure and base-pairing probabilities along the structure.*

Input: sequence $\boldsymbol{X}$, sequencing reads $\boldsymbol{R}$

Output: RNA secondary structure $\boldsymbol{S}$and base-pairing probabilities $\boldsymbol{b}$

1. Initialize $\{\boldsymbol{S},\boldsymbol{u},\boldsymbol{v},r_{min},r_{max}\}$
2. Generate candidate state $\{\boldsymbol{S}^*,\boldsymbol{u},\boldsymbol{v},r_{min},r_{max}\}$ using move set $\boldsymbol{M}$
3. Compute Metropolis-Hastings likelihood ratio:

$$\alpha_S = \frac{P(\boldsymbol{R}|\boldsymbol{S}^*,\boldsymbol{X},\boldsymbol{u},\boldsymbol{v},r_{min},r_{max})Q(\boldsymbol{S}|\boldsymbol{S}^*)}{P(\boldsymbol{R}|\boldsymbol{S},\boldsymbol{X},\boldsymbol{u},\boldsymbol{v},r_{min},r_{max})Q(\boldsymbol{S}^*|\boldsymbol{S})}$$

4. Transition from $\{\boldsymbol{S}, \boldsymbol{u}, \boldsymbol{v}, r_{min}, r_{max}\}$ to the new state $\{\boldsymbol{S}^*, \boldsymbol{u}, \boldsymbol{v}, r_{min}, r_{max}\}$ with probability $\min(1, \alpha_S)$
5. Generate candidate state $\{\boldsymbol{S}, \boldsymbol{u}^*, \boldsymbol{v}^*, r_{min}, r_{max}\}$ using move set $\boldsymbol{M}$
6. Compute Metropolis-Hastings likelihood ratio:

$$\alpha_P = \frac{P(\boldsymbol{R}|\boldsymbol{S}, \boldsymbol{X}, \boldsymbol{u}^*, \boldsymbol{v}^*, r_{min}, r_{max})}{P(\boldsymbol{R}|\boldsymbol{S}, \boldsymbol{X}, \boldsymbol{u}, \boldsymbol{v}, r_{min}, r_{max})}$$

7. Transition from $\{\boldsymbol{S}, \boldsymbol{u}, \boldsymbol{v}, r_{min}, r_{max}\}$to the new state $\{\boldsymbol{S}, \boldsymbol{u}^*, \boldsymbol{v}^*, r_{min}, r_{max}\}$ with probability $\min(1, \alpha_P)$
8. Repeat steps 2–7 until convergence or maximum number of iterations reached
9. Compute output structure $\boldsymbol{S}$ using k structures sampled in steps 2–7 by including in the output structure all base pairings (i,j) with frequency $f_{i,j} > 0.5$:

$$f_{i,j} = \frac{1}{k}\sum\nolimits_{t\in\{t_1,t_2,\cdots,t_k\}} I(\boldsymbol{S}_t(i), \boldsymbol{S}_t(j)), I(\boldsymbol{S}_t(i), \boldsymbol{S}_t(j)) = \begin{cases} 1, \boldsymbol{S}_t(i) \quad and \quad \boldsymbol{S}_t(j) \text{ are paired} \\ 0, \boldsymbol{S}_t(i) \quad and \quad \boldsymbol{S}_t(j) \text{ are unpaired} \end{cases}$$

10. Compute base-pairing probabilities b along the structure using k structures sampled in steps 2–7

$$b_i = \frac{1}{k}\sum_{t\in\{t_1,t_2,\cdots,t_k\}} I(\boldsymbol{S}_t(i)), I(\boldsymbol{S}_t(i)) = \begin{cases} 1, \boldsymbol{S}_t(i) \text{ is paired} \\ 0, \boldsymbol{S}_t(i) \text{ is unpaired} \end{cases}$$

The initial starting structures are sampled from Boltzmann distribution (Step 1) using RNAsubopt - [18] with the '−p' option. Initial DMS per-nucleotide modification rates$\boldsymbol{u}, \boldsymbol{v}$ are set to 0.01 for paired states ($\boldsymbol{u}$). For unpaired state, $\boldsymbol{v}$ was calculated for each nucleotide as the ratio of 5′ endpoints at unpaired vs all positions (DMS-Seq) and as the ratio of mismatches to the total number of reads for DMS-MaPseq.

Overall data likelihood L can easily be computed in$O(mr_{\max})$ time for both DMS-seq and DMS-MaPseq, linear in the number of sequencing reads m and the maximum read length $r_{\max}$. To speed-up evaluation of candidate structures (steps 2–4) for DMS-Seq, we used Karp-Rabin [11] fingerprint-like algorithm with running time$O(r_{min} + l(r_{max} - r_{min}))$ to compute an $l \times l$ matrix $\boldsymbol{L}$ for sequence of length l, where L_{ij} is the probability of generating the sequence fragment $[i,j]$. Each element is defined as:

$$L_{ij} = \begin{cases} P\left(R^{ij}|\boldsymbol{S}, \boldsymbol{X}, \boldsymbol{u}, \boldsymbol{v}, r_{min}, r_{max}\right), r_{min} \le (j - i + 1) \le r_{max} \\ 0, \quad otherwise \end{cases}$$

The overall data likelihood is then computed in $O(m)$ time (compare with $O(mr_{\max})$).

In effect, we now have a lookup table of read likelihoods that accounts for the all of the variables that specify the experimental setup (structure and sequence composition, per-position modification rates, and size selection). The overall running time for a single iteration of MCMC (steps 1–8) is $O(lr_{max} + m)$, linear in the number of reads m and sequence length l. On Rfam reference RNA structure dataset (168 structures × 100 MCMC chains × 100,000 iterations), the running time is 10 h on our server with 2.8 GHz E5-2680 CPU and uses no more than 16 GB RAM during the execution. As a reasonable trade-off between convergence and running time, we chose to run HiPR for 100,000 MCMC iterations.

The resultant structure samples (step 4) can be interpreted in terms of both the posterior secondary structure$\boldsymbol{S}$ and base pairing probabilities $\boldsymbol{b}$ along our structure of interest. Formally, we define a base pairing probability vector $\boldsymbol{b} = (b_1, b_2, \cdots, b_l)$ of length l with elements

$$b_i = \frac{1}{k}\sum_{t\in\{t_1,t_2,\cdots,t_k\}} I(\boldsymbol{S}_t(i)), \quad \text{where } I(\boldsymbol{S}_t(i)) = \begin{cases} 1, \boldsymbol{S}_t(i) \text{ is paired} \\ 0, \boldsymbol{S}_t(i) \text{ is unpaired} \end{cases}$$

where $I(\boldsymbol{S}_t(i))$ is a base-pairing indicator for position i on the structure $\boldsymbol{S}_t$ at step t, with MCMC steps $t_1, t_2, \cdots, t_k$ indexing sampled structures.

Similarly, we define a posterior estimate of the secondary structure $\boldsymbol{S}$ of length l as a consensus structure that includes all base-pairs (i,j) with frequency $f_{i,j}$ greater than 0.5 among the sampled structures $\boldsymbol{S}_t, t \in \{t_1, t_2, \cdots, t_k\}$:

$$f_{i,j} = \tfrac{1}{k}\sum_{t\in\{t_1,t_2,\cdots,t_k\}} I(\boldsymbol{S}_t(i), \boldsymbol{S}_t(j)),$$

$$I(\boldsymbol{S}_t(i), \boldsymbol{S}_t(j)) = \begin{cases} 1, \boldsymbol{S}_t(i) \quad and \quad \boldsymbol{S}_t(j) \text{ are paired} \\ 0, \boldsymbol{S}_t(i) \quad and \quad \boldsymbol{S}_t(j) \text{ are unpaired} \end{cases}$$

6. Reference RNA secondary structures

We obtained the first set of reference secondary structure models for human from the Rfam database [28] (version 12.1). We required the multiple sequence alignment for an Rfam family to contain at least one seed human RNA sequence. Reference genomic regions for each individual RNA sequence in the Rfam alignment were determined by mapping (GRCh37/hg19) and filtering for exact matches between the seed and genomic sequences. We required these genomic loci to overlap existing non-coding RNA annotations [14,16]. Overall, we obtained 467 loci corresponding to 264 Rfam families with their secondary structures determined by Rfam comparative sequence analysis. We used these matched genomic regions and their corresponding Rfam RNA secondary structures as a gold standard for our analysis. These Rfam structures spanned a variety of RNA classes including snoRNA, snRNA, tRNA, miRNA, and other types of RNA.

The second set of reference structures included curated mRNA structures used by the original DMS-seq study [34] for human (TFRC (NM_003234), XBP1 (NM_005080), MSRB1 (NM_016332)) and yeast (ASH1, HAC1).

To test structure prediction accuracy on the challenging ribosomal RNA class, we obtained reference 18S ribosomal RNA structures from the comparative RNA web (CRW) site [3] for human and yeast, and reference 5S ribosomal RNA structures from RNAstrand [1].

6.1. RNA structure prediction

We compared *in vivo* DMS-constrained predicted structures, *in silico* predicted structures, and HiPR predicted structures with the reference RNA structures (see **Section** "Reference RNA secondary structures").

In particular, we tested accuracies of our algorithm and two groups of methods (Table 2):

1) three *in silico* sequence-only folding algorithms: RNAfold minimum free energy (MFE) [18], RNAfold maximum expected accuracy (MEA) [18,21], and RNAfold centroid [18],
2) four constraint-based folding methods that use structural constraints determined from experimental data [5,19,43] to guide RNA folding: RNAfold hard-constrained [19], RNAfold soft-constrained [19,43], RNAstructure with pseudo-energy constraints [32,46], and the recent Restrained Max-Expect (RMEfold) algorithm [45].

The latter group of methods includes methods for folding with *hard constraints* where individual positions are forced to remain unpaired/paired, as well as a group of approaches to RNA folding

with *soft constraints* [19,43] that add additional terms to encourage or penalize base-pairing, including pseudo-energy-based methods [20,22].

In silico predicted RNA structures (minimum free energy (MFE), maximum expected accuracy (MEA), and centroid) were obtained using RNAfold software [18,19]. *In vivo DMS-constrained* RNA structures were obtained using RNAfold with hard and soft constraints [19] implementing three approaches for incorporating experimental data [5,43,47], RMEfold [45], and RNAstructure software [32].

To obtain hard-constraints and soft-constraints [19], DMS signal (number of RT stops per position for DMS-Seq and number of mismatches per position for DMS-MaPseq) was normalized using 2–8% normalization [20,45].

Hard-constraints were obtained by requiring positions with strong DMS signal (>0.2 of the maximum [34] to remain unpaired during folding (this corresponds to 'RNAfold Hard Constraints' method in figures and text).

For constrained folding with RNAstructure software [22,32], we have optimized the slope and intercept parameters (m, b) using grid-based search and the set of$n$ = 30 Rfam structures with most data before applying RNAstructure to our reference RNA structure dataset (we refer to this method as 'RNAstructure Constrained' in the figures and text). The m, b values found by grid-search (m = 0.4, b = −0.2) gave highest median MCC score of 0.48 compared to MCC = 0.42 using default $m = 1.8, b = -0.6$ parameters.

RNAfold –shapeMethod = "W" [43] was used to obtain *in vivo* DMS soft-constrained structures (this method is referred to as 'RNAfold Soft Constraints' in the figures and text) using $\tau/\sigma = 1$ as suggested by the analysis in the original RNA folding with soft constraints paper [43].

RMEfold software was run with '-d dmsseq' option to use DMS-seq optimized parameters (m, γ_1, γ_2) [45].

To compare the accuracy between methods, we used paired Wilcoxon signed-rank test unless noted otherwise.

6.2. Data availability

All data, software, predicted and reference RNA structures used in this study are freely accessible at HiPR website https://www.lisanwanglab.org/HIPR under 'Data and Results' and 'Software' tab pages, including: 1) Genome-mapped DMS-seq data (human, K562 cell line; yeast); 2) Genome-mapped DMS-MaPseq data (human); 3) Reference Rfam RNA sequences and secondary structure models; 4) Reference validated RNA sequences and secondary structures; 5) predicted RNA secondary structures for HiPR and other methods; 6) structure prediction accuracy metrics for HiPR and other methods; 7) HiPR code and software.

7. Code availability

HiPR source code is freely available at GitHub repository https://github.com/wanglab-upenn/HiPR.

CRediT authorship contribution statement

Pavel P. Kuksa: Methodology, Software, Formal analysis, Investigation, Writing - original draft. **Fan Li:** Methodology, Software, Formal analysis, Investigation, Writing - original draft. **Sampath Kannan:** Methodology, Formal analysis, Writing - original draft. **Brian D. Gregory:** Methodology, Formal analysis, Writing - original draft. **Yuk Yee Leung:** Methodology, Writing - review & editing. **Li-San Wang:** Conceptualization, Methodology, Funding acquisition, Writing - original draft, Project administration, Supervision.

Acknowledgements

This work was supported by the National Institute on Aging (U24-AG041689, U54-AG052427, U01-AG032984), National Institute of General Medical Sciences (R01-GM099962), and National Human Genome Research Institute (5T32HG000046). We thank Paul Ryvkin, Zivadin Katanic for their help and constructive input on the manuscript and website.

Author contributions

L.S.W. conceived and supervised the project. F.L and P.P.K developed algorithm, designed and performed analyses. S.K., Y.Y.L and B. D.G. participated in the model development, analysis, and interpretation. P.P.K, L.S.W. and F.L wrote the manuscript and all other authors have read and approved the manuscript.

Competing interests

The authors declare no competing interests.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2020.06.004.

References

[1] Andronescu M, Bereg V, Hoos HH, Condon A. RNA STRAND: the RNA secondary structure and statistical analysis database. BMC Bioinf 2008;9(1):340. https://doi.org/10.1186/1471-2105-9-340.
[2] Beaudoin J-D, Novoa EM, Vejnar CE, Yartseva V, Takacs CM, Kellis M, et al. Analyses of mRNA structure dynamics identify embryonic gene regulatory programs. Nat Struct Mol Biol 2018;25(8):677–86. https://doi.org/10.1038/s41594-018-0091-z.
[3] Cannone JJ, Subramanian S, Schnare MN, Collett JR, D'Souza LM, Du Y, et al. The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. BMC Bioinf 2002;3:2. , http://www.ncbi.nlm.nih.gov/pubmed/11869452.
[4] Cordero P, Kladwang W, VanLang CC, Das R. Quantitative dimethyl sulfate mapping for automated RNA secondary structure inference. Biochemistry 2012;51(36):7037–9. https://doi.org/10.1021/bi3008802.
[5] Deigan KE, Li TW, Mathews DH, Weeks KM. Accurate SHAPE-directed RNA structure determination. PNAS 2009;106(1):97–102. https://doi.org/10.1073/pnas.0806929106.
[6] Ding Y, Chan CY, Lawrence CE. RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. RNA (New York, N.Y.) 2005;11(8):1157–66. https://doi.org/10.1261/rna.2500605.
[7] Ding Y, Tang Y, Kwok CK, Zhang Y, Bevilacqua PC, Assmann SM. In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. Nature 2014;505(7485):696–700. https://doi.org/10.1038/nature12756.
[8] Hendrix DK, Brenner SE, Holbrook SR. RNA structural motifs: building blocks of a modular biomolecule. Q Rev Biophys 2005;38(3):221–43. https://doi.org/10.1017/S0033583506004215.
[9] Hooks KB, Griffiths-Jones S. Conserved RNA structures in the non-canonical Hac1/Xbp1 intron. RNA Biol 2011;8(4):552–6. https://doi.org/10.4161/rna.8.4.15396.
[10] James BD, Olsen GJ, Pace NR. Phylogenetic comparative analysis of RNA secondary structure. Methods Enzymol 1989:227–39. https://doi.org/10.1016/0076-6879(89)80104-1.
[11] Karp RM, Rabin MO. Efficient randomized pattern-matching algorithms. IBM J Res Dev 1987;31(2):249–60. https://doi.org/10.1147/rd.312.0249.
[12] Kertesz M, Wan Y, Mazor E, Rinn JL, Nutter RC, Chang HY, et al. Genome-wide measurement of RNA secondary structure in yeast. Nature 2010;467(7311):103–7. https://doi.org/10.1038/nature09322.
[13] Kladwang W, VanLang CC, Cordero P, Das R. Understanding the errors of SHAPE-directed RNA structure modeling. Biochemistry 2011;50(37):8049–56. https://doi.org/10.1021/bi200524n.
[14] Kuksa PP, Amlie-Wolf A, Katanić Ž, Valladares O, Wang L-S, Leung YY. DASHR 2.0: integrated database of human small non-coding RNA genes and mature products. Bioinformatics 2018;35(6):1033–9. https://doi.org/10.1093/bioinformatics/bty709.
[15] Lackey L, Coria A, Woods C, McArthur E, & Laederach A. (2018). Allele-specific SHAPE-MaP assessment of the effects of somatic variation and protein binding on mRNA structure. 24(4), 513–528. DOI:10.1261/rna.064469.117.

[16] Leung YY, Kuksa PP, Amlie-Wolf A, Valladares O, Ungar LH, Kannan S, et al. DASHR: database of small human noncoding RNAs. Nucl Acids Res 2016;44 (D1):D216–22. https://doi.org/10.1093/nar/gkv1188.
[17] Li F, Zheng Q, Vandivier LE, Willmann MR, Chen Y, Gregory BD. Regulatory impact of RNA secondary structure across the Arabidopsis transcriptome. Plant Cell 2012;24(11):4346–59. https://doi.org/10.1105/tpc.112.104232.
[18] Lorenz R, Bernhart SH, Höner Zu, Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. ViennaRNA Package 2.0. Algorithms Mol Biol: AMB 2011;6(1):26. https://doi.org/10.1186/1748-7188-6-26.
[19] Lorenz R, Hofacker IL, Stadler PF. RNA folding with hard and soft constraints. Algorithms Mol Biol 2016;11(1):8. https://doi.org/10.1186/s13015-016-0070-z.
[20] Low JT, Weeks KM. SHAPE-directed RNA secondary structure prediction. Methods (San Diego, Calif.) 2010;52(2):150–8. https://doi.org/10.1016/j.ymeth.2010.06.007.
[21] Lu ZJ, Gloor JW, Mathews DH. Improved RNA secondary structure prediction by maximizing expected pair accuracy. RNA 2009;15(10):1805–13. https://doi.org/10.1261/rna.1643609.
[22] Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. Proc Natl Acad Sci 2004;101(19):7287–92. https://doi.org/10.1073/pnas.0401799101.
[23] Mathews David H, Sabina J, Zuker M, Turner DH. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol 1999;288(5):911–40. https://doi.org/10.1006/jmbi.1999.2700.
[24] McCaskill JS. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers 1990;29(6–7):1105–19. https://doi.org/10.1002/bip.360290621.
[25] Merino EJ, Wilkinson KA, Coughlan JL, Weeks KM. RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE). J Am Chem Soc 2005;127(12):4223–31. https://doi.org/10.1021/ja043822v.
[26] Mizrahi O, Nachshon A, Shitrit A, Gelbart IA, Dobesova M, Brenner S, et al. Virus-induced changes in mRNA secondary structure uncover cis-regulatory elements that directly control gene expression. Mol Cell 2018;72(5):862–74. https://doi.org/10.1016/j.molcel.2018.09.003.
[27] Mustoe AM, Busan S, Rice GM, Hajdin CE, Peterson BK, Ruda VM, et al. Pervasive regulatory functions of mRNA structure revealed by high-resolution SHAPE probing. Cell 2018;173(1):181–195.e18. https://doi.org/10.1016/j.cell.2018.02.034.
[28] Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, et al. Rfam 12.0: updates to the RNA families database. Nucl Acids Res 2015;43(D1): D130–7. https://doi.org/10.1093/nar/gku1063.
[29] Ouyang Z, Snyder MP, Chang HY. SeqFold: genome-scale reconstruction of RNA secondary structure integrating high-throughput sequencing data. Genome Res 2013;23(2):377–87. https://doi.org/10.1101/gr.138545.112.
[30] Parisien M, Major F. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. Nature 2008;452(7183):51–5. https://doi.org/10.1038/nature06684.
[31] Quarrier S, Martin JS, Davis-Neulander L, Beauregard A, Laederach A. Evaluation of the information content of RNA structure mapping data for secondary structure prediction. RNA (New York, N.Y.) 2010;16(6):1108–17. https://doi.org/10.1261/rna.1988510.
[32] Reuter JS, Mathews DH, Eddy S, Mello C, Conte D, Chow J, et al. RNAstructure: software for RNA secondary structure prediction and analysis. BMC Bioinf 2010;11(1):129. https://doi.org/10.1186/1471-2105-11-129.
[33] Ritchey LE, Su Z, Tang Y, Tack DC, Assmann SM, Bevilacqua PC. Structure-seq2: sensitive and accurate genome-wide profiling of RNA structure in vivo. Nucl Acids Res 2017;45(14). https://doi.org/10.1093/nar/gkx533. e135 e135.
[34] Rouskin S, Zubradt M, Washietl S, Kellis M, Weissman JS. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. Nature 2014;505(7485):701–5. https://doi.org/10.1038/nature12894.
[35] Siegfried NA, Busan S, Rice GM, Nelson JAE, Weeks KM. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). Nat Methods 2014;11 (9):959–65. https://doi.org/10.1038/nmeth.3029.
[36] Simon LM, Morandi E, Luganini A, Gribaudo G, Martinez-Sobrido L, Turner DH, et al. In vivo analysis of influenza A mRNA secondary structures identifies critical regulatory motifs. Nucl Acids Res 2019;47(13):7003–17. https://doi.org/10.1093/nar/gkz318.
[37] Smola MJ, Rice GM, Busan S, Siegfried NA, Weeks KM. Selective 2′-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile and accurate RNA structure analysis. Nat Protoc 2015;10 (11):1643–69. https://doi.org/10.1038/nprot.2015.103.
[38] Spitale RC, Flynn RA, Zhang QC, Crisalli P, Lee B, Jung J-W, et al. Structural imprints in vivo decode RNA regulatory mechanisms. Nature 2015;519 (7544):486–90. https://doi.org/10.1038/nature14263.
[39] Sükösd Z, Swenson MS, Kjems J, Heitsch CE. Evaluating the accuracy of SHAPE-directed RNA secondary structure predictions. Nucl Acids Res 2013;41 (5):2807–16. https://doi.org/10.1093/nar/gks1283.
[40] Sun L, Fazal FM, Li P, Broughton JP, Lee B, Tang L, et al. RNA structure maps across mammalian cellular compartments. Nat Struct Mol Biol 2019;26 (4):322–30. https://doi.org/10.1038/s41594-019-0200-7.
[41] Talkish J, May G, Lin Y, Woolford JL, McManus CJ. Mod-seq: high-throughput sequencing for chemical probing of RNA structure. RNA (New York, N.Y.) 2014;20(5):713–20. https://doi.org/10.1261/rna.042218.113.
[42] Underwood JG, Uzilov AV, Katzman S, Onodera CS, Mainzer JE, Mathews DH, et al. FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. Nat Methods 2010;7(12):995–1001. https://doi.org/10.1038/nmeth.1529.
[43] Washietl S, Hofacker IL, Stadler PF, Kellis M. RNA folding with soft constraints: reconciliation of probing data and thermodynamic secondary structure prediction. Nucl Acids Res 2012;40(10):4261–72. https://doi.org/10.1093/nar/gks009.
[44] Watters KE, Yu AM, Strobel EJ, Settle AH, Lucks JB. Characterizing RNA structures in vitro and in vivo with selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). Methods 2016;103:34–48. https://doi.org/10.1016/j.ymeth.2016.04.002.
[45] Wu Y, Shi B, Ding X, Liu T, Hu X, Yip KY, et al. Improved prediction of RNA secondary structure by integrating the free energy model with restraints derived from experimental probing data. Nucl Acids Res 2015;43 (15):7247–59. https://doi.org/10.1093/nar/gkv706.
[46] Xu ZZ, & Mathews DH. (2016). Experiment-Assisted Secondary Structure Prediction with RNAstructure. In Methods in molecular biology (Clifton, N.J.) (Vol. 1490, pp. 163–176). DOI:10.1007/978-1-4939-6433-8_10.
[47] Zarringhalam K, Meyer MM, Dotu I, Chuang JH, Clote P. Integrating chemical footprinting data into RNA secondary structure prediction. PLoS One 2012;7 (10):. https://doi.org/10.1371/journal.pone.0045160e45160.
[48] Zheng Q, Ryvkin P, Li F, Dragomir I, Valladares O, Yang J, et al. Genome-wide double-stranded RNA sequencing reveals the functional significance of base-paired RNAs in Arabidopsis. PLoS Genet 2010;6(9):. https://doi.org/10.1371/journal.pgen.1001141e1001141.
[49] Zinshteyn B, Chan D, England W, Feng C, Green R, Spitale RC. Assaying RNA structure with LASER-Seq. Nucl Acids Res 2018;47(1):43–55. https://doi.org/10.1093/nar/gky1172.
[50] Zubradt M, Gupta P, Persad S, Lambowitz AM, Weissman JS, Rouskin S. DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo. Nat Methods 2017;14(1):75–82. https://doi.org/10.1038/nmeth.4057.
[51] Zuker M, Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucl Acids Res 1981;9(1):133–48. http://www.ncbi.nlm.nih.gov/pubmed/6163133.