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,e

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

doi:10.1016/j.csbj.2020.06.004

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 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.

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.
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

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
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, $R$) from DMS-treated RNAs as a function of the underlying RNA secondary structure $S$ and experimental conditions (model parameters, $H$) including per-nucleotide DMS modification rates, size selection, and nucleotide biases (Figs. 1a, b, 2; Materials and methods).

The HiPR structure likelihood model $L(S, \Theta; R) \cdot P(R|S, \Theta)$ assumes a probabilistic process underlying DMS modifications along the RNA structure $S$ and resulting sequence fragments $R$ 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 ($S, \Theta$) to find structures maximizing the posterior $\arg\max_{S, \Theta} L(S, \Theta; R)$. 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. 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 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 (US splicosomal RNA chr15:6558339–6558504 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.)
P. P. Kuksa et al. / Computational and Structural Biotechnology Journal 18 (2020) 1539–1547

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 $P(RS|X, S, \Theta) \times P(RS|X, \Theta)$ 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) over MFE structure $S_{MFE}$ across structures in RFam (Fig. 3c).

This reinforces our idea of using MCMC sampling to optimize the data likelihood $P(RS|X, S, \Theta)$ 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.

### Table 1

| Move                        | Example                      | Parameter update | Constraints |
|-----------------------------|------------------------------|------------------|-------------|
| Add a pairing interaction   | $-((.....)),$                | $S \rightarrow S'$ | Must be a valid base pair, follow steric hindrance rules, and result in a fully nested structure. |
| Delete a pairing interaction| $-((.....)),$                | $S \rightarrow S'$ | None         |
| Select a new per-nucleotide modification rate | $\epsilon U(-0.01,0.01)$ | $u \rightarrow u'$ | All rate parameters must be in $[0.1, u_i < u']$ |
|                            | $(u_a + 0.005, u_c - 0.002, u_t - 0.005, u_g + 0.001)$ | $v \rightarrow v'$ | |

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. 3a). Similarly, the reduction in the average base-pairing error was 31.3% relative to MFE and other methods (Fig. 3b). 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 (Spearman $\rho = 0.4819, P < 7.34e-05$). The improvement in prediction accuracy (MCC) was similarly strongly correlated with the average number of RT stops per A or C position (Spearman $\rho = 0.3919, P < 0.018$). 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-
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 SS 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 \( \theta \) such that their posterior probability \( P(\theta|D) \) given observed data \( D \) is maximized. Using Bayes’ rule,

\[
P(\theta|D) \propto P(D|\theta)P(\theta)
\]

we will instead optimize the likelihood function \( P(D|\theta) \) 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

\[
P(S, u, v, r_{\text{min}}, r_{\text{max}}|X, R) \propto P(R|S, X, u, v, r_{\text{min}}, r_{\text{max}})P(S, X, u, v, r_{\text{min}}, r_{\text{max}})
\]

for DMS-seq and as

\[
P(S, u, v|X, R) \propto P(R|S, X, u, v)P(S, X, u, v)
\]

for DMS-MaPseq, where \( X \) is the primary RNA sequence of length \( l \), \( S \) is a secondary structure, \( u \) and \( v \) are per-nucleotide modification rates for paired and unpaired bases, \( r_{\text{min}} \) is the minimum cloneable fragment size, \( r_{\text{max}} \) 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 \( \{S, u, v, r_{\text{min}}, r_{\text{max}}\} \), including the desired secondary structure \( S \), that maximizes the likelihood function \( P(R|S, X, u, v, r_{\text{min}}, r_{\text{max}}) \).

For prior \( P(S, X, u, v, r_{\text{min}}, r_{\text{max}}) \), 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 their likelihood function given the observed read-out \( R \) 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 \( r_{\text{min}}, r_{\text{max}} \) for DMS-Seq are set according to experimental protocol and are used in data likelihood computations to restrict range of possible read lengths to \([r_{\text{min}}, r_{\text{max}}] \).

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

\[
P(R|S, X, u, v, r_{\text{min}}, r_{\text{max}}) = \prod_{k=1}^{m} P(R_k|S, X, u, v, r_{\text{min}}, r_{\text{max}})
\]

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

\[
P(R|S, X, u, v, r_{\text{min}}, r_{\text{max}}) Z_{\theta}(X_{i-1}) \prod_{k=i}^{j-1} (1 - Z_{\theta}(X_k))
\]

where

\[
Z_{\theta}(X_i) = \begin{cases} v(X_i), & \text{if base } i \text{ is unpaired} \\ u(X_i), & \text{if base } i \text{ is paired} \\ \end{cases}
\]

is the probability of DMS modification of the \( i \)th nucleotide \( X_i \).

Intuitively, this read probability is equivalent to observing modification immediately upstream of the \( i \)th 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 of all possible reads \([i, j]\) with the length restricted to \([r_{\text{min}}, r_{\text{max}}] \) 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 \( i_1, \ldots, i_m \) is defined as

\[
P(R_k|S, X, u, v) \sim \prod_{i=i_1}^{i_m} Z_{\theta}(X_i) \prod_{j=i_1}^{i_m} (1 - Z_{\theta}(X_j))
\]

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 \( P(R|S, X, \Theta) \) by a random walk Metropolis-Hastings algorithm on the parameter space \( \{S, u, v, r_{\text{min}}, r_{\text{max}}\} \) (\{S, u, v\} for DMS-MaPseq). To accomplish this, we define a move set \( M \) 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 \( S \) and \( S' \) is given by:

\[
Q(SS') = \frac{n(M(x \rightarrow S))}{n(M(x \rightarrow S'))}
\]

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

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 \( X \), sequencing reads \( R \)

Output: RNA secondary structure \( S \) and base-pairing probabilities \( b \)

1. Initialize \( \{S, u, v, r_{\text{min}}, r_{\text{max}}\} \)
2. Generate candidate state \( \{S', u, v, r_{\text{min}}, r_{\text{max}}\} \) using move set \( M \)
3. Compute Metropolis-Hastings likelihood ratio:

\[
x_{\theta} = \frac{P(R|S', X, u, v, r_{\text{min}}, r_{\text{max}}) Q(SS')}{P(R|S, X, u, v, r_{\text{min}}, r_{\text{max}}) Q(SS)}
\]
4. Transition from \((S, u, v, r_{\text{min}}, r_{\text{max}})\) to the new state \((S', u', v', r_{\text{min}}, r_{\text{max}})\) with probability \(\min(1, x_5)\).
5. Generate candidate state \((S, u, v, r_{\text{min}}, r_{\text{max}})\) using move set \(M\).
6. Compute Metropolis-Hastings likelihood ratio:

\[
\frac{p(R, S, X, u, v, r_{\text{min}}, r_{\text{max}})}{p(R, S, X, u, v, r_{\text{min}}, r_{\text{max}})}
\]

7. Transition from \((S, u, v, r_{\text{min}}, r_{\text{max}})\) to the new state \((S', u', v', r_{\text{min}}, r_{\text{max}})\) with probability \(\min(1, x_6)\).
8. Repeat steps 2–7 until convergence or maximum number of iterations reached.
9. Compute output structure \(S\) using \(k\) structures sampled in steps 2–7 by including in the output structure all base pairings \((i, j)\) with frequency \(f_{ij} > 0.5\):

\[
f_{ij} = \frac{1}{K} \sum_{t \in \{t_1, t_2, \ldots, t_k\}} l(S_i(i), S_j(j), i, S_i(i), S_j(j))
= \begin{cases} 1, & \text{if } (i, j) \text{ is paired} \\ 0, & \text{if } (i, j) \text{ is unpaired} \end{cases}
\]
10. Compute base-pairing probabilities \(b\) along the structure using \(k\) structures sampled in steps 2–7:

\[
b_i = \frac{1}{K} \sum_{t \in \{t_1, t_2, \ldots, t_k\}} l(S_i(i), i, S_i(i)) = \begin{cases} 1, & \text{if } (i, j) \text{ is paired} \\ 0, & \text{if } (i, j) \text{ is unpaired} \end{cases}
\]

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

6. Reference RNA secondary structures

We obtained the first set of reference secondary structure models from human RNA 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), XPB1 (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
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 dm = 30 Rfam structures with most data before applying RNAstructure to our reference RNA structure dataset (we refer to this method as 'RNAstructure Constrained' method in figures and text).

During folding (this corresponds to 'RNAfold Hard Constraints' 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.

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 (5T32 HG000046). 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 the 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, Vejan 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/s41594-018-0991-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/bi200986q.
[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.0809910106.
[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/s0033583505004215.
[9] Hooke KB, Griffiths-Jones S. Conserved RNA secondary structure in the non-canonical Hac1/Xbp1 intron. RNA Biol 2011;8(4):552–6. https://doi.org/10.4161/rna.8.4.15356.
[10] James BD, Olsen GJ, Pace NR. Phylogenetic comparative analysis of RNA secondary structure. Methods Enzymol 1989;227:229–39. https://doi.org/10.1016/0076-6879(90)01041-3.
[11] Karp RM, Rabin MO. Efficient randomized pattern-matching algorithms. IBM J Res Dev 1987;31(2):249–60. https://doi.org/10.1147/JRDP.147.1124.
[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/bi200524h.
[14] Kuksa PP, Amile-Wolf A, Katanic Z, 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/bty979.
[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:10.1016/j.rna.064469.117.
