Quantitative Biology 2020, 8(4): 369–381
https://doi.org/10.1007/s40484-020-0223-4

METHOD

IRIS: A method for predicting *in vivo* RNA secondary structures using PARIS data

Jianyu Zhou[1,2], Pan Li[3], Wanwen Zeng[1,4], Wenxiu Ma[5], Zhipeng Lu[6], Rui Jiang[1,7], Qiangfeng Cliff Zhang[3,*], Tao Jiang[1,2,8,*]

[1] Bioinformatics Division, BNRIST, Tsinghua University, Beijing 100084, China
[2] Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[3] MOE Key Laboratory of Bioinformatics, Beijing Advanced Innovation Center for Structural Biology, Center for Synthetic and Systems Biology, Tsinghua-Peking Center for Life Sciences, School of Life Sciences, Tsinghua University, Beijing 100084, China
[4] College of Software, Nankai University, Tianjin 300071, China
[5] Department of Statistics, University of California, Riverside, CA 92521, USA
[6] Department of Pharmacology and Pharmaceutical Sciences, University of Southern California, CA 90089, USA
[7] Department of Automation, Tsinghua University, Beijing 100084, China
[8] Department of Computer Science and Engineering, University of California, Riverside, CA 92521, USA
* Correspondence: jiang@cs.ucr.edu, qczhang@mail.tsinghua.edu.cn

Received June 21, 2020; Revised June 28, 2020; Accepted June 29, 2020

Background: **RNA secondary structures play a pivotal role in posttranscriptional regulation and the functions of non-coding RNAs, yet *in vivo* RNA secondary structures remain enigmatic. PARIS (Psoralen Analysis of RNA Interactions and Structures) is a recently developed high-throughput sequencing-based approach that enables direct capture of RNA duplex structures *in vivo*. However, the existence of incompatible, fuzzy pairing information obstructs the integration of PARIS data with the existing tools for reconstructing RNA secondary structure models at the single-base resolution.**
Methods: **We introduce IRIS, a method for predicting RNA secondary structure ensembles based on PARIS data. IRIS generates a large set of candidate RNA secondary structure models under the guidance of redistributed PARIS reads and then uses a Bayesian model to identify the optimal ensemble, according to both thermodynamic principles and PARIS data.**
Results: **The predicted RNA structure ensembles by IRIS have been verified based on evolutionary conservation information and consistency with other experimental RNA structural data. IRIS is implemented in Python and freely available at http://iris.zhanglab.net.**
Conclusion: **IRIS capitalizes upon PARIS data to improve the prediction of *in vivo* RNA secondary structure ensembles. We expect that IRIS will enhance the application of the PARIS technology and shed more insight on *in vivo* RNA secondary structures.**

Keywords: RNA secondary structure; PARIS data; *in vivo*; structure ensembles; incompatible reads

Author summary: Decoding RNA secondary structures in living cells is still a thorny problem in bioinformatics. Recently, PARIS enables the direct capture of *in vivo* RNA duplex structures in a high-throughput sequencing way. However, PARIS can only obtain low-resolution information of a mixture of alternative RNA structures. A computational method to construct the high-resolution structure ensemble is the key to exploit the full power of the PARIS technology. Here we present IRIS, a method for predicting *in vivo* RNA secondary structure ensembles base on PARIS data. We expect that IRIS will help shed more insight on *in vivo* RNA secondary structures.

INTRODUCTION

Beyond encoding proteins, RNAs play a variety of regulatory and functional roles in cells [1]. The RNA structure, in addition to its sequence, is often key to an RNA's function [2]. Driven by the intramolecular Watson-Crick base pairings (AU, GC) and the wobble base pairings (GU), RNA secondary structures form the most important step in RNA folding [3,4].

Many methods have been developed to predict RNA secondary structures from sequences [5]. Classic methods predict structures that minimize free energy based on thermodynamic parameters, including Mfold [6], ViennaRNA [7], and RNAstructure [8]. However, RNA folding is a kinetic and stochastic process, and RNA secondary structures are dynamic *in vivo* [9]. An RNA sequence can often adopt an ensemble of multiple distinct secondary structures that satisfy a thermodynamic equilibrium [10,11]. Thus, models and methods that trace kinetic folding and sample representative secondary structures have been proposed, such as barrier trees [12], basin hopping graphs [13], and non-redundant sampling [14]. These methods aim to search for local optimal secondary structures based on energy as representative structures within the kinetic folding landscape. However, the predicted structure ensembles often misrepresent the *in vivo* structures, because various environmental and trans-acting factors are not considered [9]. In addition, exploring the whole kinetic RNA folding landscape is exponential in the lengths of RNAs and therefore intractable [15].

Additional information is necessary to accurately predict RNA secondary structures *in vivo*. Comparative sequence analysis takes advantages of homologous sequences and identifies conserved base pairings, relying on the knowledge of RNA families that adopt similar structures [16]. Although thousands of such families are collected in Rfam [17,18], only a few are of high quality to construct reliable secondary structure models. RNAs structures determined by nuclear magnetic resonance (NMR), X-ray crystallography or cryogenic electron microscopy (cryo-EM) have been used as training data for machine learning [19–23], and even deep learning [24] approaches to make more predictions. Nonetheless, the scant amount of available training data, as well as the non-negligible gap between training and prediction, has limited the wide application of these methods. Most importantly, although both comparative sequence analysis and machine learning methods bring the prediction closer to the truth by introducing information from other sources, they are powerless when the target RNA structures are significantly heterogeneous, which is very common given the complicated *in vivo* situation [9].

High-throughput sequencing-based RNA structure probing techniques make it possible to directly query RNA secondary structures transcriptome-wide and *in vivo* [25–27]. These techniques can be classified into two categories according to experimental principles. Nucleotide modification-based methods, such as DMS-seq [28], Structure-seq [29] and icSHAPE [30], use small molecule modifiers to measure a score for each base that indicates whether it is paired. Many computational methods have been proposed to incorporate modification-based data for accurate RNA secondary structure prediction in more cellular states [31–34], and even to generate secondary structure ensembles of multiple conformations [35].

Crosslinking-based methods, such as SPLASH [36], LIGR-seq [37] and PARIS [38], directly capture RNA duplex structures *in vivo* [39]. For example, a PARIS data consists of a collection of sequencing reads that indicate which two regions of an RNA interact [40]. However, two major issues hamper the construction of complete RNA secondary structure models using crosslinking-based data [41]: 1) RNA structures often have alternative conformations that coexist in the population of multiple copies of a RNA molecule. These distinct conformations result in incompatible PARIS reads, *i.e.*, one region can interact with multiple other regions, 2) the resolution of PARIS reads is too low to precisely determine paired bases. Figure 1A shows a real example of PARIS reads that are mapped to the U2 snRNA [42]. The PARIS sequencing results showed that region B can pair with both region A and region C, and the regions were too wide to exactly identify paired bases. There is a desire to develop computational methods to incorporate PARIS sequencing results to construct RNA structural models that represent the multiple distinct and coexisting conformations with exact base-pairing information.

Here we propose IRIS, a method for predicting ensembles of *in vivo* RNA secondary structures using PARIS data. The ensemble is composed of a fixed number of different RNA secondary structures with corresponding proportions. IRIS outperforms minimum free energy-based prediction in terms of the agreement with evolutionary conservation and icSHAPE probing data. IRIS is implemented in Python and freely available at http://iris.zhanglab.net. We expect that IRIS will help improve the application of PARIS data and insights into RNA secondary structures *in vivo*.

RESULTS

The core concept and framework of the IRIS algorithm

IRIS uses RNA sequence with mapped PARIS reads as the input, and outputs an ensemble of representative RNA secondary structures in corresponding proportions. To

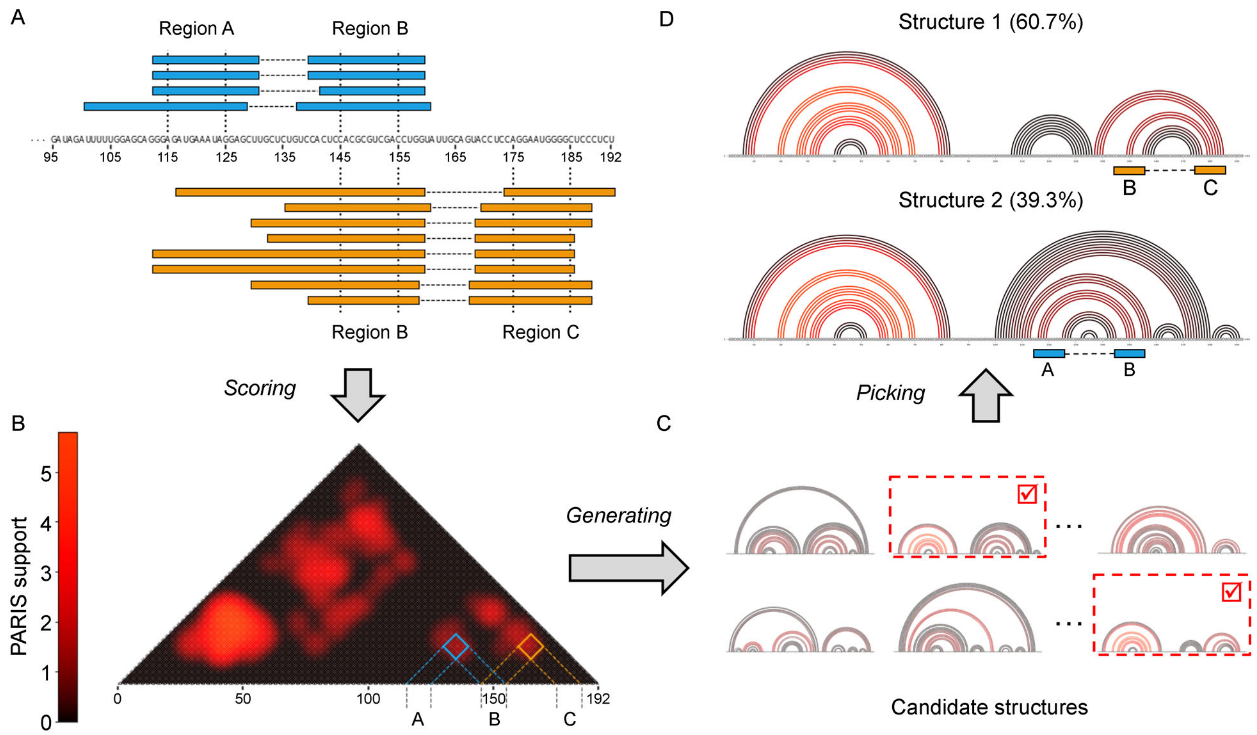

Figure 1. An example of PARIS data and the core concept of IRIS. (A) A real example of PARIS reads at the tail of the U2 snRNA. PARIS identifies 4 blue reads that support the interaction between region B (145–155 nt) and region A (115–125 nt), and 8 yellow reads support the interaction between region B and region C (175–185 nt). The interactions, however, cannot happen simultaneously. The PARIS reads do not contain the exact information of paired bases. (B) The heatmap of PARIS support of the U2 snRNA. Note that in addition to the incompatibility on regions A, B, and C, actually the entire RNA is full of incompatible interactions. (C) Candidate structures generated by IRIS. The red colors of base pairs represent the corresponding PARIS support. IRIS generates a large number of candidate structures and eventually picks two structures in this example. (D) The ensemble of two representative RNA secondary structures predicted by IRIS. The proportion of structure 1 is 60.7% which describes the source of yellow reads. Structure 2 takes 39.3% and describes the source of blue reads.

address the issues of incompatible and fuzzy PARIS reads (Fig. 1A), IRIS first converts the low-resolution reads into PARIS support to score pairwise interactions (Fig. 1B), and then generates a large number of candidate RNA secondary structures base on the PARIS support (Fig. 1C). Therefore, each candidate structure will contain precise base-pairing information supported by a subset of PARIS reads. Finally, IRIS picks out the optimal combination of candidate structures and assigns the proportions to describe the PARIS reads as well as possible (Fig.1D).

The algorithm of IRIS consists of three steps: *scoring*, *generating* and *picking*. The *scoring* step cleans mapped PARIS reads and transforms the read coverage into a matrix that represents the support from the PARIS data. The *generating* step first scans the input RNA sequence to assemble all short, theoretically possible stems with PARIS support ranking higher than a certain threshold. Compatible combinations of these stems are then set as constraints in RNA folding that generates thousands of secondary structures with locally minimum free energy. Similar structures are trimmed by clustering, and hundreds of structures are retained as candidate structures. The *picking* step first eliminates redundant candidate structures by fitting the matrix of PARIS support with the base-pairing matrices of candidate structures via LASSO regression. It then enumerates all combinations with a fixed number of structures from the remaining ones as candidate structure ensembles, assigning the proportion of each structure using linear regression. A Bayesian framework is adopted to identify the optimal secondary structure ensemble, considering both thermodynamic principles and PARIS data. Formally, the output ensemble with K representative structures can be represented as $X^*, \boldsymbol{\alpha}^*$, where $X^* = (\boldsymbol{X}_1^*, \boldsymbol{X}_2^*, \ldots, \boldsymbol{X}_K^*)^{\mathrm{T}}$ denoting the representative structures ($\boldsymbol{X}$ is the base-pairing matrix defined in Eq. (4)) and $\boldsymbol{\alpha}^* = (\alpha_1^*, \alpha_2^*, \ldots, \alpha_K^*)$ denoting their corresponding proportions. A flowchart of IRIS is shown in Fig. 2, with details given in Section “Materials and methods”.

Evaluation metrics and dataset of IRIS predictions

Since it is currently impossible to know the true *in vivo* RNA secondary structure ensemble, we use two indirect metrics to evaluate the ensemble predicted by IRIS: 1) the evidence of evolutionary conservation, 2) the consistency with icSHAPE data, an orthogonal type of *in vivo* RNA secondary structural information.

The PARIS data for evaluating the performance of IRIS was collected from our previous study in the human liver cancer cell line (Huh7) infected by the Zika virus, which also used icSHAPE to probe RNA structure in the same condition [42]. We focused on 524 human RNAs, collected from the Rfam database [18] via the bpRNA interface [43]. RNA families in Rfam offer evolutionary conservation information and bpRNA helps to curate the consensus secondary structure of each collected RNA. After mapping and cleaning PARIS data and processing the icSHAPE pipeline [44,45], 11 RNAs with over 1,000 PARIS reads mapped and with valid icSHAPE scores were retained as the test data. We then ran IRIS on these 11 RNAs with corresponding PARIS reads and predicted

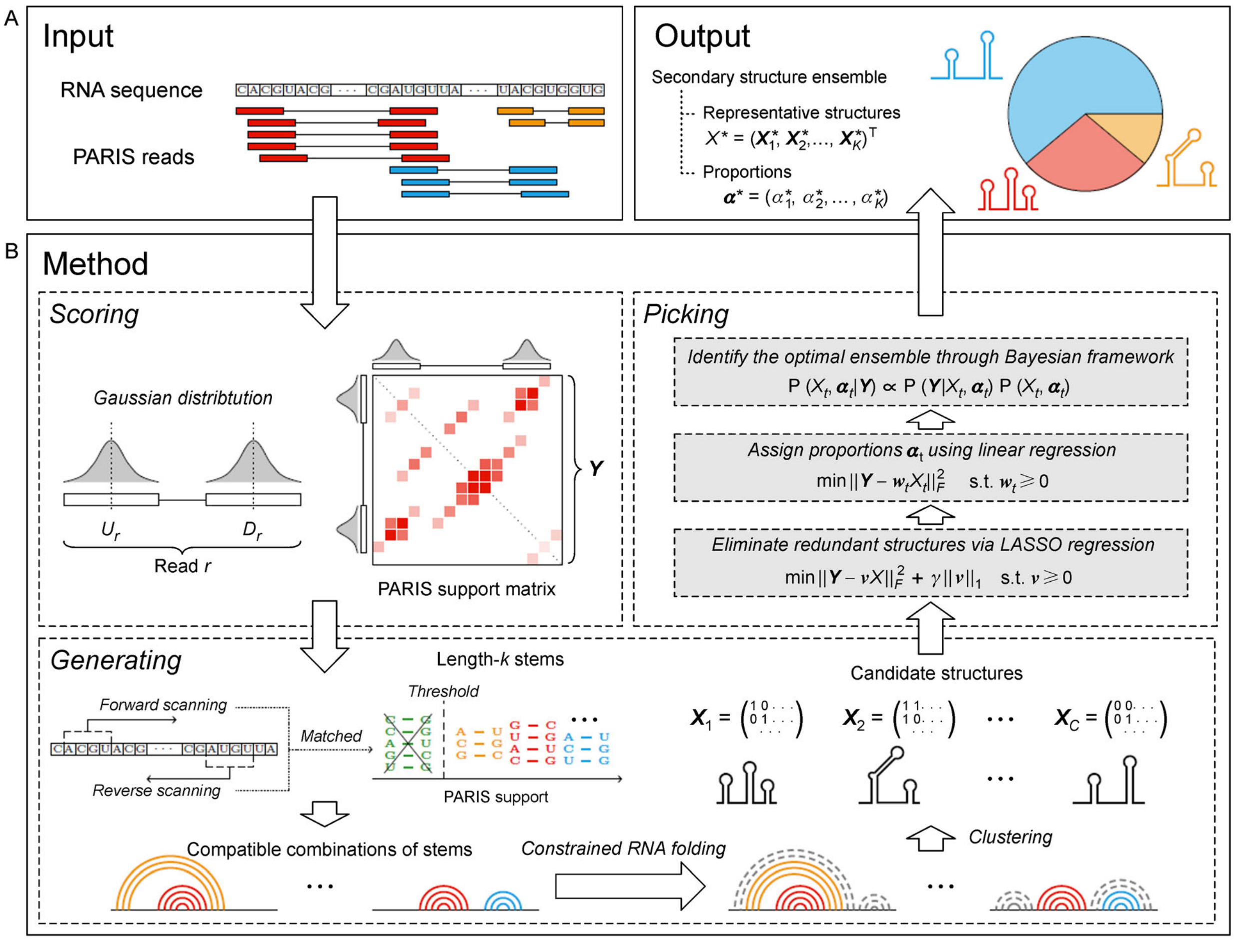

Figure 2. An overview of IRIS. (A) The input and output of IRIS. IRIS takes as the input an RNA sequence with mapped PARIS reads, and outputs an RNA secondary structure ensemble composed of K representative secondary structures with corresponding proportions. (B) The three steps of IRIS: *scoring*, *generating* and *picking*. The *scoring* step converts the read coverage through a Gaussian distribution to a matrix of PARIS support. The *generating* step first scans the RNA sequence to collect all length-k stems with PARIS support higher than a certain threshold. Next, regarding the compatible combination of stems as hard constraints, a constrained folding algorithm is applied to make up the stems into complete secondary structures, and similar structures are clustered into C candidate structures. The *picking* step eliminates redundant structures via LASSO regression, and assigns proportions using linear regression, and eventually identifies the optimal ensemble by a Bayesian model.

RNA secondary structure ensembles with 1, 2 and 3 representative structures (denoted as IRIS-1, IRIS-2 and IRIS-3 respectively). Parameters for running IRIS differed by the length of RNA (see Section S2.2 of Supplementary Materials for details). We also computed the structure with minimum free energy (denoted as MFE) as an ensemble of only one structure as the baseline.

Evaluation by evolutionary conservation

Functional RNAs often have evolutionarily conserved secondary structures, which holds true even for secondary structures that form transiently during RNA folding *in vivo* [46]. Thus, the ensemble of RNA secondary structures predicted by IRIS is expected to be evolutionarily conserved. To validate IRIS, we retrieved the multiple sequence alignment of homologous sequences in the corresponding RNA family from Rfam, computed the normalized mutual information [47] between each pair of bases using R-scape [48], and represented the result as an $n \times n$ (where n denotes the length of the RNA) matrix $\boldsymbol{M}$, which measures the degree of evolutionary conservation of base pairs. In this way, a base pair that occurs frequently among secondary structures in an ensemble should have a high score in the matrix $\boldsymbol{M}$. So, we define the base-pairing probability between bases i and j derived from the predicted ensemble $X^*, \boldsymbol{\alpha}^*$ from IRIS as $b_{ij} = \sum_{k=1}^{K} \alpha_k^* x_{kij}^*$, and we use $\boldsymbol{B}$ to represent the base-pairing probability matrix. The benchmark for testing the evolutionary conservation of the predicted ensemble is set as the Kullback-Leibler (KL) distance between $\boldsymbol{B}$ and $\boldsymbol{M}$ defined in [49] as Eq. (1).

$$D_{KL}(\boldsymbol{B},\boldsymbol{M}) = \sum_{i=1}^{n} \sum_{j=i+1}^{n} b_{ij}\log b_{ij} - b_{ij}\log m_{ij} - b_{ij} + m_{ij} \quad (1)$$

As results shown in Table 1, we noticed that ensembles with multiple representative structures predicted by IRIS-2 and IRIS-3 are all resulted in a lower KL distance compared with the structure predicted by MFE, with the exception of IRIS-2 on the small nucleolar RNA SNORD45, which is the shortest RNA in the testing data. These findings imply that IRIS can successfully utilize PARIS data to predict an ensemble of RNA secondary structures that are supported by evolutionary conservation. Moreover, IRIS-1 outperformed MFE on most of the RNAs, indicating that a single structure predicted by IRIS yields a more evolutionarily conserved structure. However, the ensemble prediction is more suitable for methods based on PARIS data.

To make the benchmarking more intuitive, we used the U2 snRNA as a proof-of-principle and plotted the matrices compared by KL distance as heatmaps (Fig.3). We established base-pairing probability matrices from ensembles predicted by MFE, IRIS-1, IRIS-2, and IRIS-3 (Fig. 3A). Focusing on the three matrices from IRIS, we noted that all structures in different ensembles are fully conserved on the heading half of the U2 snRNA as a long stem component. The alternative structures generated by IRIS-2 and IRIS-3 are all located on the tailing half of the RNA. Fig. 3B presents the matrix of normalized mutual information calculated from the multiple sequence alignment of homogenous sequences in the U2 snRNA family. The ability to accurately predict consecutive conserved base pairs contributes substantially to the increased performance of IRIS. Another example of the RMRP is shown in Section S3 of Supplementary Materials.

Evaluation on icSHAPE data

Next, we determined the consistency between IRIS-predicted RNA structure ensembles and icSHAPE data, an *in vivo* assay orthogonal to PARIS. The icSHAPE data was processed into a score ranging from 0 to 1 for each base of the RNA, indicating the probability that this base

Table 1 The KL distance between the base-pairing probability matrix and the normalized mutual information matrix

Rfam ID/Sequence accession	Name	Length	MFE	IRIS-1	IRIS-2	IRIS-3
RF00004/ABBA01028418.1	U2 snRNA	192	147.7	**98.1**	**98.0**	**93.0**
RF00030/M29212.1	RMRP	264	127.9	131.0	**96.1**	**85.3**
RF00045/L07382.1	SNORA73	207	218.2	**214.4**	**198.4**	**195.0**
RF00091/AC099332.2	SNORA62	153	327.1	**327.1**	**320.7**	**299.6**
RF00138/X72205.1	SNORD16	99	352.5	**318.3**	**315.2**	**292.3**
RF00279/AL357314.11	SNORD45	79	152.0	156.1	158.4	**142.1**
RF00478/AY077737.1	SCARNA6	275	494.1	520.7	**481.4**	**470.1**
RF00567/AL121585.22	SNORD17	237	385.3	410.9	**383.8**	**366.4**
RF00618/U62822.1	U4atac	127	53.5	**53.5**	**52.9**	**52.6**
RF01296/AF308283.1	snoU85	330	245.0	**243.5**	**234.6**	**232.2**
RF02556/ABSL01008103.1	snaR-A	115	582.8	**582.8**	**555.7**	**555.7**

Note: The scores in bold are the cases that IRIS performs better than MFE.

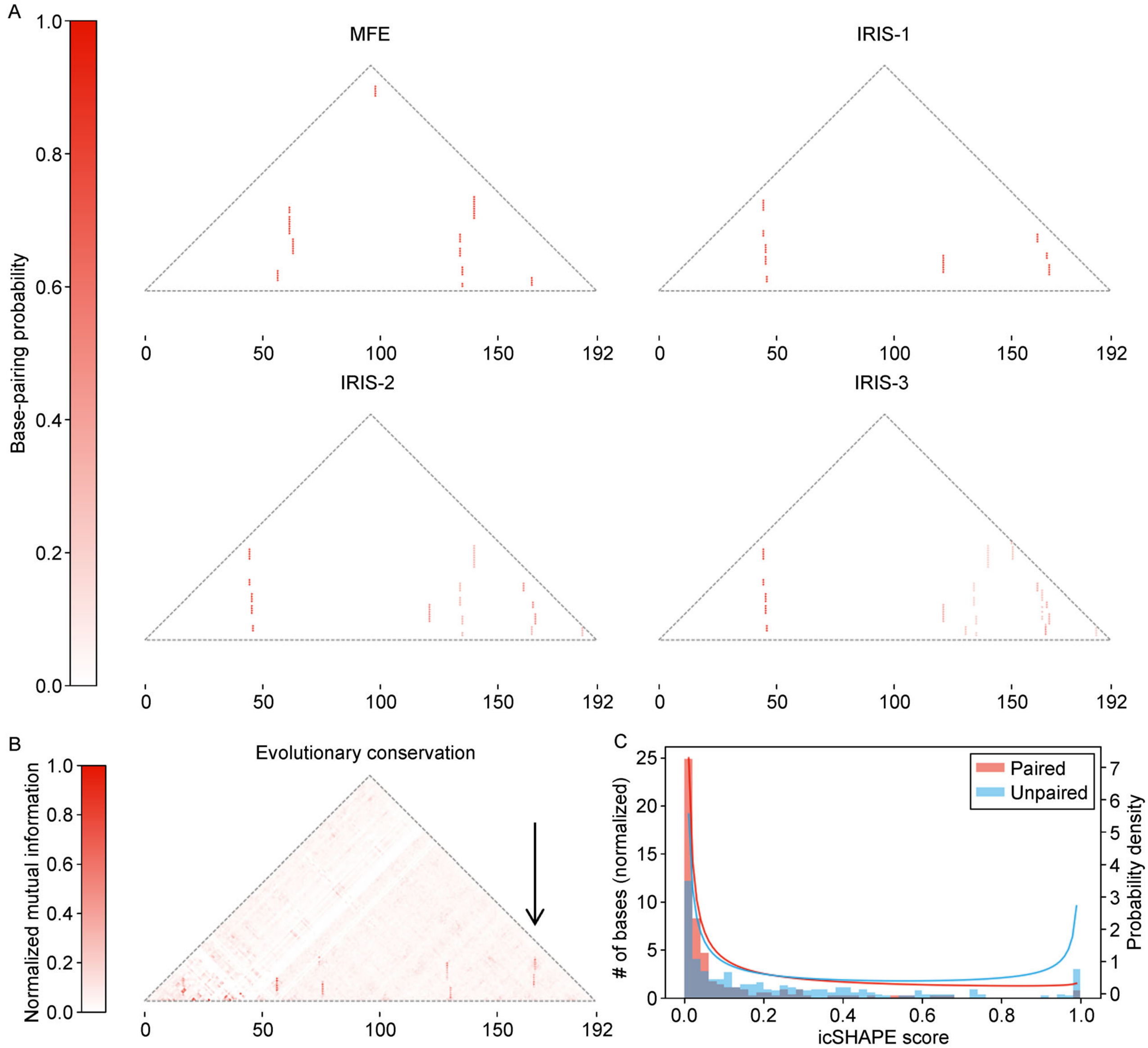

Figure 3. Evaluation results of IRIS. (A) Heatmaps of base-pairing probability matrices for the structure ensemble of the U2 snRNA predicted by MFE, IRIS-1, IRIS-2 and IRIS-3. (B) The heatmap of normalized mutual information between each pair of bases of the U2 snRNA using R-scape. The correct prediction of the pairwise interactions pointed by the black arrow is the major factor that makes IRIS perform better than MFE. (C) The histogram of icSHAPE scores of paired bases (red) and unpaired bases (blue) in the training data. The curves are the probability density function of Beta distributions for fitting icSHAPE scores.

is paired [33]. Thus, to obtain the distribution of icSHAPE scores, we focused on the RNA whose consensus secondary structure curated by bpRNA was validated by published articles, and we retained 25 RNAs whose lengths are shorter than 100 nt as the training data (considering that short, validated RNAs are unlikely to have alternative secondary structures *in vivo*). Based on validated secondary structures, we collected icSHAPE scores of 291 paired bases and 289 unpaired bases and used them to fit Beta distributions. Figure 3C shows the histogram of icSHAPE scores and the Beta distribution of scores for paired and unpaired bases.

Unlike PARIS data, icSHAPE data provide marginal information for base pairing. As a result, the benchmark for measuring the consistency between the predicted ensemble and icSHAPE data was set as the log-likelihood of observing icSHAPE scores from the predicted ensemble, which can be formulated and simplified as Eq. (2).

$$\log P(\mathbf{z}|X^*, \boldsymbol{\alpha}^*) = \sum_{i=1}^{n} \log(b_i \mathscr{B}_{\text{paired}}(z_i) + (1-b_i)\mathscr{B}_{\text{unpaired}}(z_i)) \quad (2)$$

$\mathbf{z} = (z_1, z_2, ..., z_n)$ represents the vector of icSHAPE scores, and bases whose icSHAPE scores are not available are omitted from the summation. $b_i = \sum_{j=1}^{n} b_{ij}$ represents the marginal pairing probability of base i in the predicted ensemble. $\mathscr{B}_{\text{paired}}(\cdot)$ and $\mathscr{B}_{\text{unpaired}}(\cdot)$ denote the Beta distribution of icSHAPE scores for paired bases and unpaired bases respectively. The derivation of the likelihood can be regarded as a generation model from the predicted ensemble to icSHAPE scores, and the full derivation is included in Supplementary Section S1.2.

Table 2 shows the result of the log-likelihood of observing icSHAPE scores from predicted ensembles of the test data. For most RNAs, the log-likelihood of IRIS-2 and IRIS-3 were significantly higher than MFE, which indicates the predicted ensemble from IRIS was more likely to be the *in vivo* ensemble that generated the observed icSHAPE data. In some cases (SNORA73, SNORD16 and U4atac) IRIS-2 and IRIS-3 performed slightly worse than MFE, and the log-likelihood of all methods were very close. Also, for single structure prediction, IRIS-1 performed better than MFE for more than half of the RNAs. We infer that the predicted RNA secondary structure ensembles using IRIS is consistent with icSHAPE data. Thus, IRIS can predict a physiologically relevant *in vivo* RNA secondary structure ensemble.

IRIS efficiently samples RNA structural space

The goal of IRIS is to uncover secondary structures with low free energy that are supported by PARIS data, and identify the optimal ensemble from these structures. To analyse how IRIS work, we used the *generating* step of IRIS to generate 100 candidate structures of the U2 snRNA, and compared them with another 100 secondary structures generated by the non-redundant sampling algorithm [14] (denote as NRDS). NRDS is the state-of-the-art method for exploring representative structures of the RNA folding landscape, considering only free energy. We also included the minimum free energy structure. Next, we calculated the base pair distance [50] between each pair of these 201 RNA secondary structures, and then performed multidimensional scaling [51] based on the distance matrix to visualize the relationship of the structures by embedding them into a two dimensional plane (Fig. 4). Finally, we split the plane using a Voronoi diagram [52] so that each dot representing the secondary structure occupies an area with a different color, where blue represents free energy, as shown in Fig. 4A, and green represents PARIS support, as shown in Fig. 4B.

As the results shown in Fig. 4, the secondary structures generated by IRIS are dispersedly distributed and are located in states with relatively high free energy and high PARIS support, whereas structures generated by NRDS are gathered into clusters and located near the MFE structure. These findings imply that IRIS, with the assistance of the PARIS data, can efficiently explore more space of the RNA folding landscape and avoid becoming trapped by low-free-energy basins. The Bayesian model in the *picking* step can successfully identify two distant structures, both with significantly

Table 2 The log-likelihood of observing icSHAPE data from the predicted RNA secondary structure ensemble

Rfam ID/Sequence accession	Name	MFE	IRIS-1	IRIS-2	IRIS-3
RF00004/ABBA01028418.1	U2 snRNA	551.5	**577.2**	**578.8**	**579.5**
RF00030/M29212.1	RMRP	435.4	422.4	**450.7**	**452.1**
RF00045/L07382.1	SNORA73	298.2	**298.3**	298.1	297.6
RF00091/AC099332.2	SNORA62	176.9	**176.9**	**190.1**	**187.3**
RF00138/X72205.1	SNORD16	136.4	136.1	135.1	134.7
RF00279/AL357314.11	SNORD45	25.4	**40.6**	**40.6**	**48.7**
RF00478/AY077737.1	SCARNA6	316.3	312.3	**319.3**	**317.7**
RF00567/AL121585.22	SNORD17	321.2	**322.0**	**335.6**	**335.7**
RF00618/U62822.1	U4atac	121.4	**121.4**	120.7	118.9
RF01296/AF308283.1	snoU85	468.7	467.4	**473.7**	**473.8**
RF02556/ABSL01008103.1	snaR-A	115.0	**115.0**	**120.3**	**120.3**

Note: The scores in bold are the cases that IRIS performs better than MFE.

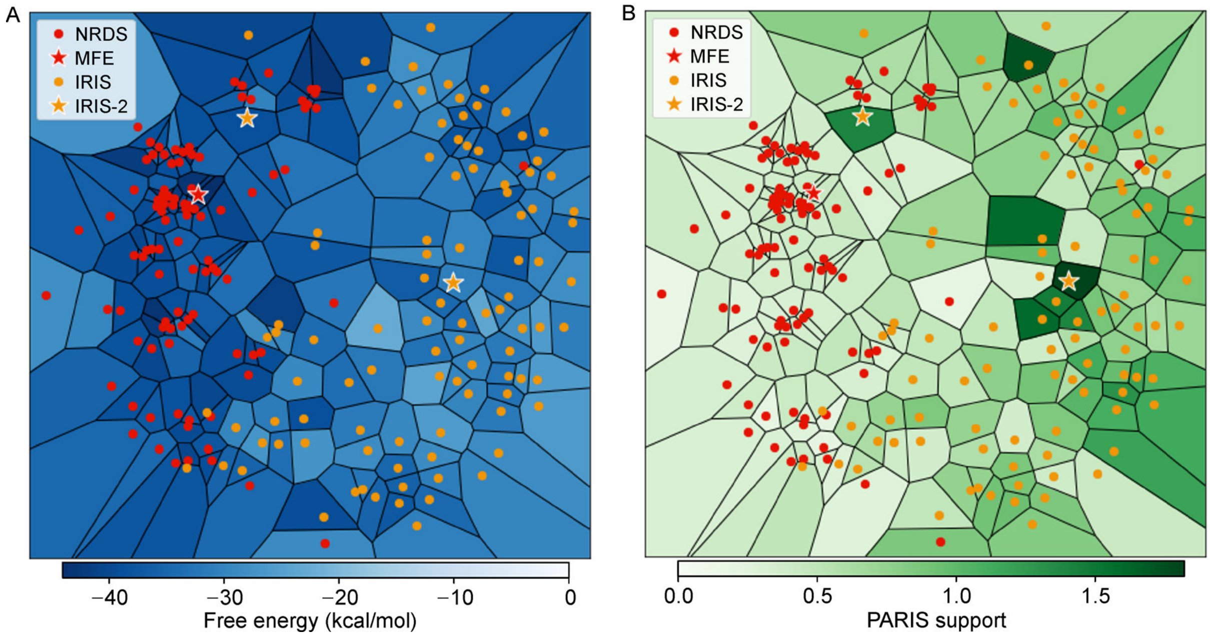

Figure 4. The distribution of candidate structures. The two-dimensional representation of RNA secondary structures embedded by the multidimensional scaling algorithm based on base pair distances. The red dots represent the structures generated by NRDS, and the red star represents the MFE structure. The yellow dots denote the structures produced by the *generating* step of IRIS, and the yellow stars denote structures elected by IRIS-2. The Voronoi diagram is applied on the plane to make each dot occupy an area with a different color, where blue indicates free energy in (A), and green represents PARIS support in (B).

high PARIS support and relatively low free energy, as representative structures in the predicted ensemble.

DISCUSSION

IRIS capitalizes upon PARIS data to improve the prediction of *in vivo* RNA secondary structure ensembles, which is an important but difficult question in the field of RNA structure. IRIS addresses two thorny issues of PARIS data, incompatible reads and low-resolution pairing information, by converting PARIS reads into PARIS supports and allocating incompatible reads to different representative secondary structures in the predicted ensemble. IRIS consists of three steps (*scoring*, *generating* and *picking*) that generate a set of candidate RNA secondary structures. With the guidance of PARIS reads, IRIS identifies the optimal ensemble through a Bayesian model that considers thermodynamic principles and PARIS data. The predicted RNA secondary structure ensemble can be validated by evolutionary conservation and icSHAPE data.

Currently, IRIS has two major limitations. Although some PARIS reads can reflect pseudoknots, IRIS can only predict representative structures without pseudoknots: the constrained folding algorithm including pseudoknots is NP-hard [53], and the number of candidate structures will dramatically increase when including pseudoknots. In addition, using PARIS data to predict an ensemble of secondary structures is not suitable for long RNAs (*e.g.*, longer than 500 nt), because the search space increases exponentially with the length of the RNA and thus becomes too large. In this scenario, a feasible approach is to use PARIS data to divide the RNA in to structural domains [42].

When evaluating the performance of IRIS, only a limited number of RNAs were considered due to the fact that a low coverage by the PARIS data would be insufficient to produce reliable predictions for lowly expressed RNAs. In addition to exploiting more powerful computational methods to mitigate low-coverage cases, a potentially more promising solution might be to conduct PARIS experiments on a small number of target RNAs of special interest. This would allow us to amplify the abundance of the target RNAs to ensure that they will have sufficient coverage in the resultant PARIS data for subsequent computational analyses.

MATERIALS AND METHODS

Scoring pairwise interactions

A PARIS data is a collection of single-end high-throughput sequencing reads that are formed by joining sequences sampled from the corresponding parts of a duplex structure. So, mapping PARIS reads to a reference RNA sequence can be accomplished by a spliced read alignment tool such as STAR [54]. Since spliced read alignment tools are originally designed for analysing RNA-Seq data and alternative splicing, their default parameters are not suitable for mapping PARIS reads, especially some biases concerning the intron length and GU-AG rules. Based on the parameters given in [38], more considerations of fairly splitting reads are set as the parameters listed in Supplementary Section S2.1. After the mapping, only reads that contain one gap and two continuously mapped sequences with lengths longer than 15 nt each are retained as purified PARIS data. Hence, each read can be represented as two intervals on the reference RNA sequence.

When PARIS captures the duplex structure of some stem, the flanking sequences of the stem are possibly included in the reads, which induces the low-resolution read issue. If flanking sequences are considered to be randomly included in the reads, it is reasonable to assume that each part of the captured stem is more likely to appear at the centre of one of the two intervals of a PARIS read. Thus, instead of counting read coverage uniformly, a score, called PARIS support, is converted from the input PARIS reads based on a Gaussian distribution, which indicates the strength of pairwise interactions. Formally, for every base pair (i,j), $i<j$, the PARIS support y_{ij} is defined as in Eq. (3):

$$y_{ij}=y_{ji}=\log\left(\sum_{r}\frac{N(i|U_r,\sigma^2)}{N(U_r|U_r,\sigma^2)}\frac{N(j|D_r,\sigma^2)}{N(D_r|D_r,\sigma^2)}+1\right) \quad (3)$$

Here, U_r and D_r represent the centre of the upstream and downstream intervals of read r, respectively. The standard deviation of the Gaussian distribution N is $\sigma=\overline{L}/6$, where $\overline{L}$ is the average interval length of all mapped reads of the RNA. This setting considers the empirical rule of Gaussian distributions, that is, about 99.7% of values are within a band around the mean with a width of six standard deviations. In this way, for each read, the centre of intervals contributes most to the PARIS support, while uncovered regions make almost no contribution. Then, the density of the distribution is normalized by the peak density, the total density is scaled by a logarithmic transformation and a constant 1 is added to ensure non-negativity. Finally, for an RNA with length n, its PARIS support y_{ij} can be represented as an $n\times n$ symmetric matrix $\boldsymbol{Y}$. Figure 1B provides an example of the PARIS support matrix of the U2 snRNA.

Generating candidate secondary structures

An RNA secondary structure consists of stems with different lengths, which are formed by stacking base pairs (AU, CG, and GU). Stems with high PARIS support are more likely to be included in *in vivo* secondary structures. Therefore, all theoretically possible length-k stems are assembled by scanning intervals of length k on the forward and reverse sequences of an RNA and checking whether two intervals from the forward and reverse sequences form a legitimate stem (with no overlap and separated by a gap big enough to form potential loops). Theoretically, multiple values of k should be considered to reduce bias and k should be small enough to achieve high sensitivity. Note that each length-k stem covers $i+1$ stems of length $k-i$. To avoid repeated counting, the covered short stems are omitted in IRIS. Then, the PARIS support for a stem is defined as the mean PARIS support of all base pairs in the stem.

Each stem with PARIS support higher than a certain fraction of nonzero elements in $\boldsymbol{Y}$ is considered as a necessary component of all *in vivo* RNA secondary structures. Treating the base pairs in such a stem as hard constraints, a constrained RNA folding algorithm developed in the ViennaRNA package [7] is applied to compute a complete secondary structure with locally minimum free energy containing this stem. In order to explore more locally optimal structures, not only is every single stem treated as constraints, compatible combinations of pairs of stems are also included as constraints, where compatibility means that two stems do not overlap or form a pseudoknot [55]. In this way, thousands of pseudoknot-free secondary structures that are at relatively stable states and contain one or two stems with high PARIS support are generated.

Although the above RNA folding was performed subject to different constraints, it is still possible that some identical or similar secondary structures were generated. So, the final step is to perform an agglomerative clustering algorithm [56] based on the base pair distance [50] between RNA secondary structures to group the generated structures into C clusters. For each cluster, the secondary structure with the lowest free energy is retained as the representative structure of the cluster. These C representative structures are considered as the candidates of *in vivo* RNA secondary structures supported by PARIS data. Formally, each candidate secondary structure can be represented by an $n\times n$ base-pairing matrix $\boldsymbol{X}_c$ with each element defined as

$$x_{cij}=\begin{cases}1, & \text{if base } i \text{ and } j \text{ are pair in the } c\text{th structure;}\\ 0, & \text{otherwise.}\end{cases} \quad (4)$$

Then, the PARIS support of the structure $\boldsymbol{X}_c$ is defined as the mean PARIS support of all base pairs as in Eq. (5):

$$S(\boldsymbol{X}_c) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} y_{ij}x_{cij}}{\sum_{i=1}^{n}\sum_{j=1}^{n} x_{cij}} \tag{5}$$

Picking the optimal ensemble

The final step of IRIS is to infer an RNA secondary structure ensemble from generated candidate structures that can best explain the PARIS data. In order to define an RNA secondary structure ensemble with K representative structures, an index vector $\boldsymbol{t} = (t_1, t_2, ..., t_K)$ is defined to denote a subset of indices of C candidate structures. Then, the structure ensemble indexed by $\boldsymbol{t}$ can be represented as a $K \times n \times n$ tensor $X_t = (\boldsymbol{X}_{t_1}, \boldsymbol{X}_{t_2}, ..., \boldsymbol{X}_{t_K})^{\mathrm{T}}$ denoting the representative structures and a vector $\boldsymbol{\alpha}_t = (\alpha_{t_1}, \alpha_{t_2}, ..., \alpha_{t_K})$ denoting their corresponding proportions. The values of α_t are determined by normalizing the coefficients of a linear regression between $\boldsymbol{Y}$ and X_t, which tries to fit the PARIS support matrix by a non-negative linear combination of base-pairing matrices of representative structures, and can be solved as in Eq. (6):

$$\begin{aligned} \min \quad & \|\boldsymbol{Y} - \boldsymbol{w_t} X_t\|_F^2 \\ \text{s.t.} \quad & \boldsymbol{w_t} \geqslant \boldsymbol{0} \end{aligned} \tag{6}$$

Here $\|\cdot\|_F$ is the Frobenius norm of a matrix, $\boldsymbol{w_t} = (w_{t_1}, w_{t_2}, ..., w_{t_K})$ is the vector of coefficients, and the proportion is defined as $\alpha_{t_i} = w_{t_i} / \sum_{j=1}^{K} w_{t_j}$.

Clearly, there are $\binom{C}{K}$ possible ensembles by picking K structures from C candidate structures, which is a considerable amount since from every ensemble, we need perform a linear regression to determine the proportion of its representative structures. Meanwhile, when enumerating all the combinations with K structures, it is common that some structures are redundant with respect to the PARIS support, which leads the simple linear regression method to incorrectly allocate proportions in this degenerative case. Therefore, a LASSO regression on all C candidate structures is applied as a quadratic programming solution as shown in Eq. (7):

$$\begin{aligned} \min \quad & \|\boldsymbol{Y} - \boldsymbol{v} X\|_F^2 + \gamma \|\boldsymbol{v}\|_1 \\ \text{s.t.} \quad & \boldsymbol{v} \geqslant \boldsymbol{0} \end{aligned} \tag{7}$$

Here, X denotes the tensor of all C candidate structures and $\boldsymbol{v}$ is the vector of all coefficients. The parameter γ is determined by iterative halving from a large initial value (*e.g.,* 1) until the resulting $\boldsymbol{v}$ contains more than K non-zero elements. Then, the scope of candidate ensembles is narrowed down by only picking structures whose corresponding coefficients in $\boldsymbol{v}$ are non-zero.

Based on ensembles filtered by LASSO regression, IRIS adopts the Bayesian framework to choose the optimal ensemble according to thermodynamic principles and the PARIS data, which can be formulated as maximizing the posterior probability of $X_t, \boldsymbol{\alpha_t}$ given the PARIS support $\boldsymbol{Y}$, as given in Eq. (8):

$$\mathrm{P}(X_t, \boldsymbol{\alpha_t} | \boldsymbol{Y}) \propto \mathrm{P}(\boldsymbol{Y} | X_t, \boldsymbol{\alpha_t}) \mathrm{P}(X_t, \boldsymbol{\alpha_t}) \tag{8}$$

Here, the prior $\mathrm{P}(X_t, \boldsymbol{\alpha_t})$ indicates the probability of observing the ensemble in the thermodynamic equilibrium. Since the proportions $\boldsymbol{\alpha_t}$ have already been determined by using the PARIS data, the prior probability can be simplified by assuming that the mean free energy of the ensemble follows the Boltzmann distribution [10], *i.e.*,

$$\mathrm{P}(X_t, \boldsymbol{\alpha_t}) = \frac{1}{Z} \exp\left(-\frac{1}{\beta} \sum_{i=1}^{K} \alpha_{t_i} E(\boldsymbol{X}_{t_i}) \right) \tag{9}$$

Here, Z represents the partition function, β represents the product of the Boltzmann factor and the thermodynamic temperature, and $E(\cdot)$ computes the free energy of a given RNA secondary structure. Then, the likelihood $\mathrm{P}(\boldsymbol{Y} | X_t, \boldsymbol{\alpha_t})$ indicates the probability of observing the PARIS support matrix from the ensemble, which can be modelled by an exponential distribution based on the mean PARIS support of representative structures as in Eq. (10):

$$\mathrm{P}(\boldsymbol{Y} | X_t, \boldsymbol{\alpha}_t) = \lambda \exp\left(-\lambda \left(S_{\max} - \sum_{i=1}^{K} \alpha_{t_i} S(\boldsymbol{X}_{t_i}) \right) \right) \tag{10}$$

Here, $S_{\max}$ denotes the maximum PARIS support among the C candidate structures, which is set as the baseline to measure the deviation in the PARIS support of the ensemble. The parameter λ allows us to tune the scale of the distribution to make it comparable to the prior distribution. The determination and derivation of λ are described in Supplementary Section S1.1. Finally, the optimal RNA secondary structure ensemble $X^*, \boldsymbol{\alpha}^*$ is the one that maximizes the posterior probability, *i.e.*,

$$X^*, \boldsymbol{\alpha}^* = \underset{X_t, \boldsymbol{\alpha_t}}{\operatorname{argmax}} \, \mathrm{P}(X_t, \boldsymbol{\alpha_t} | \boldsymbol{Y}) \tag{11}$$

Implementation

As the method described above, we have implemented IRIS in Python. The algorithms for the basic RNA secondary structure analysis such as calculating free energy, computing partition function and constrained folding are powered by ViennaRNA [7]. We utilize SciPy [57] for regression and Scikit-learn [58] to perform clustering. VARNA software [59] together with Matplotlib [60] is used to plot RNA secondary structures (*e.g.*, Fig.1C, D). When predicting an ensemble using IRIS, it is better to set the expected number K of the representative structure to a small number less than 5, to ensure that the resulting ensemble is sufficiently informative. In addition, IRIS requires three parameters, namely the range of length k for short stems, the certain fraction of PARIS support in $\boldsymbol{Y}$ as a threshold for filtering stems, and the number of clusters C.

ABBREVIATIONS

PARIS	psoralen analysis of RNA interactions and structures
icSHAPE	*in vivo* click selective 2-hydroxyl acylation and profiling experiment
MFE	minimum free energy
NRDS	non-redundant sampling algorithm
LASSO	least absolute shrinkage and selection operator
KL	distance Kullback-Leibler distance
NP-hard	non-deterministic polynomial-time hard

SUPPLEMENTARY MATERIALS

The supplementary materials can be found online with this article at https://doi.org/10.1007/s40484-020-0223-4.

AUTHOR CONTRIBUTIONS

Q.C.Z. conceived the project. T.J. and Q.C.Z. supervised the entire project. J.Z. and T.J. designed the IRIS algorithms. P.L. assisted in data collection and pre-processing and gave many critical suggestions to the methods. Q.C. Z. and T.J. proposed evaluation benchmarks. W.M. and Z.L. gave many useful suggestions. W.M. provided the support of computational resources. W.Z. and R.J. carried out a preliminary exploration of the project. All authors read and approved the final manuscript.

ACKNOWLEDGEMENTS

This work was supported by the Chinese Ministry of Science and Technology (No. 2018YFA0107603 to Q.C.Z.), the National Natural Science Foundation of China (Nos. 91740204 and 31761163007 to Q.C.Z.), the National Natural Science Foundation of China (No. 61772197 to T.J.) and the National Key Research and Development Program of China (No. 2018YFC0910404 to T.J.). Q.C.Z thanks for support from the Beijing Advanced Innovation Center for Structural Biology and the Tsinghua-Peking Joint Center for Life Sciences.

COMPLIANCE WITH ETHICS GUIDELINES

Jianyu Zhou, Pan Li, Wanwen Zeng, Wenxiu Ma, Zhipeng Lu, Rui Jiang, Qiangfeng Cliff Zhang and Tao Jiang declare that they have no conflict of interest.

The article does not contain any human or animal subjects performed by any of the authors.

REFERENCES

1. Eddy, S. R. (2001) Non-coding RNA genes and the modern RNA world. Nat. Rev. Genet., 2, 919–929
2. Cech, T. R. and Steitz, J. A. (2014) The noncoding RNA revolution-trashing old rules to forge new ones. Cell, 157, 77–94
3. Tinoco, I. Jr and Bustamante, C. (1999) How RNA folds. J. Mol. Biol., 293, 271–281
4. Fallmann, J., Will, S., Engelhardt, J., Grüning, B., Backofen, R. and Stadler, P. F. (2017) Recent advances in RNA folding. J. Biotechnol., 261, 97–104
5. Rivas, E. (2013) The four ingredients of single-sequence RNA secondary structure prediction. A unifying perspective. RNA Biol., 10, 1185–1196
6. Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res., 31, 3406–3415
7. Hofacker, I. L. (2003) Vienna RNA secondary structure server. Nucleic Acids Res., 31, 3429–3431
8. Reuter, J. S. and Mathews, D. H. (2010) RNAstructure: software for RNA secondary structure prediction and analysis. BMC Bioinformatics, 11, 129
9. Bevilacqua, P. C., Ritchey, L. E., Su, Z. and Assmann, S. M. (2016) Genome-wide analysis of RNA secondary structure. Annu. Rev. Genet., 50, 235–266
10. McCaskill, J. S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers, 29, 1105–1119
11. Chen, S.-J. (2008) RNA folding: conformational statistics, folding kinetics, and ion electrostatics. Annu. Rev. Biophys., 37, 197–214
12. Flamm, C., Hofacker, I. L., Stadler, P. F. and Wolfinger, M. T. (2002) Barrier trees of degenerate landscapes. Z. Phys. Chem., 216, 155
13. Kucharík, M., Hofacker, I. L., Stadler, P. F. and Qin, J. (2014) Basin Hopping Graph: a computational framework to characterize RNA folding landscapes. Bioinformatics, 30, 2009–2017
14. Michálik, J., Touzet, H. and Ponty, Y. (2017) Efficient approximations of RNA kinetics landscape using non-redundant sampling. Bioinformatics, 33, i283–i292
15. Hofacker, I. L., Schuster, P. and Stadler, P. F. (1998) Combinatorics of RNA secondary structures. Discrete Appl. Math., 88, 207–237
16. Rivas, E. and Eddy, S. R. (2001) Noncoding RNA gene detection using comparative sequence analysis. BMC Bioinformatics, 2, 8
17. Kalvari, I., Nawrocki, E. P., Argasinska, J., Quinones-Olvera, N., Finn, R. D., Bateman, A. and Petrov, A. I. (2018) Non-coding RNA analysis using the rfam database. Curr. Protoc. Bioinf., 62, e51
18. Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E. P., Rivas, E., Eddy, S. R., Bateman, A., Finn, R. D. and Petrov, A. I. (2018) Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Res., 46, D335–D342
19. Knudsen, B. and Hein, J. (2003) Pfold: RNA secondary structure prediction using stochastic context-free grammars. Nucleic Acids Res., 31, 3423–3428
20. Do, C. B., Woods, D. A. and Batzoglou, S. (2006) CONTRAfold: RNA secondary structure prediction without physics-based models. Bioinformatics, 22, e90–e98

21. Zakov, S., Goldberg, Y., Elhadad, M. and Ziv-Ukelson, M. (2011) Rich parameterization improves RNA structure prediction. J. Comput. Biol., 18, 1525–1542
22. Andronescu, M., Condon, A., Hoos, H. H., Mathews, D. H. and Murphy, K. P. (2007) Efficient parameter estimation for RNA secondary structure prediction. Bioinformatics, 23, i19–i28
23. Andronescu, M., Condon, A., Hoos, H. H., Mathews, D. H. and Murphy, K. P. (2010) Computational approaches for RNA energy parameter estimation. RNA, 16, 2304–2318
24. Singh, J., Hanson, J., Paliwal, K. and Zhou, Y. (2019) RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. Nat. Commun., 10, 5407
25. Kwok, C. K. (2016) Dawn of the *in vivo* RNA structurome and interactome. Biochem. Soc. Trans., 44, 1395–1410
26. Leamy, K. A., Assmann, S. M., Mathews, D. H. and Bevilacqua, P. C. (2016) Bridging the gap between *in vitro* and *in vivo* RNA folding. Q. Rev. Biophys., 49, e10
27. Strobel, E. J., Yu, A. M. and Lucks, J. B. (2018) High-throughput determination of RNA structures. Nat. Rev. Genet., 19, 615–634
28. Rouskin, S., Zubradt, M., Washietl, S., Kellis, M. and Weissman, J. S. (2014) Genome-wide probing of RNA structure reveals active unfolding of mRNA structures *in vivo*. Nature, 505, 701–705
29. Ding, Y., Tang, Y., Kwok, C. K., Zhang, Y., Bevilacqua, P. C. and Assmann, S. M. (2014) *In vivo* genome-wide profiling of RNA secondary structure reveals novel regulatory features. Nature, 505, 696–700
30. Spitale, R. C., Flynn, R. A., Zhang, Q. C., Crisalli, P., Lee, B., Jung, J.-W., Kuchelmeister, H. Y., Batista, P. J., Torre, E. A., Kool, E. T., *et al.* (2015) Structural imprints *in vivo* decode RNA regulatory mechanisms. Nature, 519, 486–490
31. Deigan, K. E., Li, T. W., Mathews, D. H. and Weeks, K. M. (2009) Accurate SHAPE-directed RNA structure determination. Proc. Natl. Acad. Sci. USA, 106, 97–102
32. Deng, F., Ledda, M., Vaziri, S. and Aviran, S. (2016) Data-directed RNA secondary structure prediction using probabilistic modeling. RNA, 22, 1109–1119
33. Wu, Y., Shi, B., Ding, X., Liu, T., Hu, X., Yip, K. Y., Yang, Z. R., Mathews, D. H. and Lu, Z. J. (2015) Improved prediction of RNA secondary structure by integrating the free energy model with restraints derived from experimental probing data. Nucleic Acids Res., 43, 7247–7259
34. Washietl, S., Hofacker, I. L., Stadler, P. F. and Kellis, M. (2012) RNA folding with soft constraints: reconciliation of probing data and thermodynamic secondary structure prediction. Nucleic Acids Res., 40, 4261–4272
35. Spasic, A., Assmann, S. M., Bevilacqua, P. C. and Mathews, D. H. (2018) Modeling RNA secondary structure folding ensembles using SHAPE mapping data. Nucleic Acids Res., 46, 314–323
36. Aw, J. G. A., Shen, Y., Wilm, A., Sun, M., Lim, X. N., Boon, K.-L., Tapsin, S., Chan, Y.-S., Tan, C.-P., Sim, A. Y., *et al.* (2016) *In vivo* mapping of eukaryotic rna interactomes reveals principles of higher-order organization and regulation. Mol. Cell, 62, 603–617
37. Sharma, E., Sterne-Weiler, T., O'Hanlon, D. and Blencowe, B. J. (2016) Global mapping of human RNA-RNA interactions. Mol. Cell, 62, 618–626
38. Lu, Z., Zhang, Q. C., Lee, B., Flynn, R. A., Smith, M. A., Robinson, J. T., Davidovich, C., Gooding, A. R., Goodrich, K. J., Mattick, J. S., *et al.* (2016) Rna duplex map in living cells reveals higher-order transcriptome structure. Cell, 165, 1267–1279
39. Gong, J., Ju, Y., Shao, D. and Zhang, Q. C. (2018) Advances and challenges towards the study of RNA-RNA interactions in a transcriptome-wide scale. Quant. Biol., 6, 239–252
40. Lu, Z., Gong, J. and Zhang, Q. C. (2018) PARIS: Psoralen analysis of RNA interactions and structures with high throughput and resolution. In: RNA Detection, pp. 59–84. Springer
41. Fischer-Hwang, I., Lu, Z., Zou, J. and Weissman, T. (2019) Cross-linked RNA secondary structure analysis using network techniques. bioRxiv, 668491
42. Li, P., Wei, Y., Mei, M., Tang, L., Sun, L., Huang, W., Zhou, J., Zou, C., Zhang, S., and Qin, C.-f. (2018) Integrative analysis of zika virus genome RNA structure reveals critical determinants of viral infectivity. Cell host & microbe. 24, 875–886. e875
43. Danaee, P., Rouches, M., Wiley, M., Deng, D., Huang, L. and Hendrix, D. (2018) bpRNA: large-scale automated annotation and analysis of RNA secondary structure. Nucleic Acids Res., 46, 5381–5394
44. Li, P., Shi, R. and Zhang, Q. C. (2019) icSHAPE-pipe: A comprehensive toolkit for icSHAPE data analysis and evaluation. Methods, 178, 96–103
45. Flynn, R. A., Zhang, Q. C., Spitale, R. C., Lee, B., Mumbach, M. R. and Chang, H. Y. (2016) Transcriptome-wide interrogation of RNA secondary structure in living cells with icSHAPE. Nat. Protoc., 11, 273–290
46. Zhu, J. Y. A., Steif, A., Proctor, J. R. and Meyer, I. M. (2013) Transient RNA structure features are evolutionarily conserved and can be computationally predicted. Nucleic Acids Res., 41, 6273–6285
47. Martin, L. C., Gloor, G. B., Dunn, S. D. and Wahl, L. M. (2005) Using information theory to search for co-evolving residues in proteins. Bioinformatics, 21, 4116–4124
48. Rivas, E., Clements, J. and Eddy, S. R. (2017) A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. Nat. Methods, 14, 45–48
49. Hamada, M. (2012) Direct updating of an RNA base-pairing probability matrix with marginal probability constraints. J. Comput. Biol., 19, 1265–1276
50. Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M., and Schuster, P. (1994) Fast folding and comparison of RNA secondary structures. Monatshefte für Chemie/Chemical Monthly. 125, 167–188
51. Cox, M. A. and Cox, T. F. (2008) Multidimensional scaling. In: Handbook of Data Visualization, pp. 315–347. Springer
52. Aurenhammer, F. (1991) Voronoi diagrams—a survey of a fundamental geometric data structure. ACM Comput. Surv., 23, 345–405
53. Lyngsø, R. B. (2004) Complexity of pseudoknot prediction in

simple models. In: International Colloquium on Automata, Languages, and Programming, pp. 919–931. Springer

54. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T. R. (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 29, 15–21
55. Lyngsø, R. B. and Pedersen, C. N. (2000) RNA pseudoknot prediction in energy-based models. J. Comput. Biol., 7, 409–427
56. Murtagh, F. (1983) A survey of recent advances in hierarchical clustering algorithms. Comput. J., 26, 354–359
57. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., *et al.* (2020) Scipy 1.0: Fundamental algorithms for scientific computing in python. Nat. Methods, 17, 261–272
58. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R. and Dubourg, V. (2011) Scikit-learn: Machine learning in python. J. Mach. Learn. Res., 12, 2825–2830
59. Darty, K., Denise, A. and Ponty, Y. (2009) VARNA: Interactive drawing and editing of the RNA secondary structure. Bioinformatics, 25, 1974–1975
60. Hunter, J. D. (2007) Matplotlib: A 2D graphics environment. Comput. Sci. Eng., 9, 90–95

