IRIS: A Method for Predicting *in vivo* RNA Secondary Structures Using PARIS Data

Supplementary information

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[8,1,2,*]

[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*

Contents

S1 Important derivation

S1.1 Determination of λ in the picking step

The parameter λ is used to scale the likelihood distribution $\mathrm{P}(\boldsymbol{Y}|\mathcal{X}_t, \boldsymbol{\alpha}_t)$ to make it comparable with the prior distribution $\mathrm{P}(\mathcal{X}_t, \boldsymbol{\alpha}_t)$. Before deriving the parameter λ, we first define the prior and likelihood distribution for a single structure from generated candidate structures as in Equations S1 and S2 according based on Equations 9 and 10:

$$\mathrm{P}(\boldsymbol{X}_c) = \frac{1}{Z}\exp\left(-\frac{1}{\beta}\mathcal{E}(\boldsymbol{X}_c)\right) \tag{S1}$$

$$\mathrm{P}(\boldsymbol{Y}|\boldsymbol{X}_c) = \lambda\exp\left(-\lambda\left(\mathcal{S}_{max} - \mathcal{S}(\boldsymbol{X}_c)\right)\right) \tag{S2}$$

In this way, the determination of parameter λ is independent of K (the number of structures in the desired ensemble) by solving the equation as in Equation S3:

$$\frac{\max\limits_c \mathrm{P}(\boldsymbol{Y}|\boldsymbol{X}_c)}{\min\limits_c \mathrm{P}(\boldsymbol{Y}|\boldsymbol{X}_c)} = \frac{\max\limits_c \mathrm{P}(\boldsymbol{X}_c)}{\min\limits_c \mathrm{P}(\boldsymbol{X}_c)} \tag{S3}$$

Then, we can easily conclude that the maximum value of $\mathrm{P}(\boldsymbol{Y}|\boldsymbol{X}_c)$ is obtained when $\mathcal{S}(\boldsymbol{X}_c) = \mathcal{S}_{max}$ and the minimum value is obtained when $\mathcal{S}(\boldsymbol{X}_c) = \mathcal{S}_{min}$, where $\mathcal{S}_{max}$ and $\mathcal{S}_{min}$ respectively denote the maximum and minimum PARIS support among the candidate structures. Finally, through logarithm transform and simple mathematical manipulation, we can get the parameter λ as in Equation S4:

$$\lambda = \frac{\max\limits_c\{\log \mathrm{P}(\boldsymbol{X}_c)\} - \min\limits_c\{\log \mathrm{P}(\boldsymbol{X}_c)\}}{\mathcal{S}_{max} - \mathcal{S}_{min}} \tag{S4}$$

S1.2 Derivation of the log-likelihood of icSHAPE data

First, the scores of icSHAPE data for paired and unpaired bases are fitted separately by using the Beta distribution implemented in SciPy as follows:

$$\begin{aligned}\mathcal{B}_{\text{paired}}(z_i) &= Beta(z_i|\alpha = 0.158, \beta = 0.846)\\ \mathcal{B}_{\text{unpaired}}(z_i) &= Beta(z_i|\alpha = 0.209, \beta = 0.365)\end{aligned} \tag{S5}$$

Then, based on the assumption that the icSHAPE scores of each base on the RNA are independent of each other, the log-likelihood of icSHAPE scores $\boldsymbol{z}$ of the RNA observed from the predicted set $\widetilde{\mathcal{X}}, \widetilde{\boldsymbol{\alpha}}$ can be decomposed by Equation S6:

$$\log \mathrm{P}\left(\boldsymbol{z}\middle|\widetilde{\mathcal{X}},\widetilde{\boldsymbol{\alpha}}\right)=\sum_{i=1}^{n}\log \mathrm{P}\left(z_i\middle|\widetilde{\mathcal{X}},\widetilde{\boldsymbol{\alpha}}\right) \tag{S6}$$

Next, the probability of observing z_i given the ensemble is a mixture of observing z_i from each structure, *i.e.*,

$$\log \mathrm{P}\left(z_i\middle|\widetilde{\mathcal{X}},\widetilde{\boldsymbol{\alpha}}\right)=\log\left(\sum_{k=1}^{K}\tilde{\alpha}_k \mathrm{P}\left(z_i\middle|\widetilde{\boldsymbol{X}}_k\right)\right) \tag{S7}$$

The probability $\mathrm{P}(z_i|\widetilde{\boldsymbol{X}}_k)$ is only dependent on whether base *i* is paired in structure *k*, that is

$$\mathrm{P}\left(z_i\middle|\widetilde{\boldsymbol{X}}_k\right)=\left(\sum_{j=1}^{n}\tilde{x}_{kij}\right)\mathcal{B}_{\mathrm{paired}}(z_i)+\left(1-\left(\sum_{j=1}^{n}\tilde{x}_{kij}\right)\right)\mathcal{B}_{\mathrm{unpaired}}(z_i) \tag{S8}$$

Note that $\sum_{j=1}^{n}\tilde{x}_{kij}$ must be 0 or 1 due to the fact that each base can either form a base pair or be unpaired. Here, we expand the base-pairing probability of base *i* defined in the main content as in Equation S9:

$$b_i=\sum_{j=1}^{n}b_{ij}=\sum_{j=1}^{n}\sum_{k=1}^{k}\tilde{\alpha}_k\tilde{x}_{kij}=\sum_{k=1}^{K}\tilde{\alpha}_k\sum_{j=1}^{n}\tilde{x}_{kij} \tag{S9}$$

Finally, we can easily derive Equation 2 by substituting equations S7, S8 and S9 into Equation S6.

S2 Parameter configuration

S2.1 Parameters for running STAR

As mentioned in the main text, because STAR was originally designed for analyzing alternative splicing in RNA-Seq data, we modify the default parameters of STAR to remove the biases that take into account certain natures of slicing as **Table S1**.

Table S1 Special parameters used in STAR for mapping PARIS reads

Parameter	Value	Parameter	Value
outSJfilterReads	Unique	outSJfilterOverhangMin	5 5 5 5
alignIntronMin	10	outSJfilterCountUniqueMin	1 1 1 1
scoreGapNoncan	0	outSJfilterCountTotalMin	1 1 1 1
scoreGapGCAG	0	outSJfilterDistToOtherSJmin	0 0 0 0
scoreGapATAC	0		

S2.2 Parameters for running IRIS

Three parameters need to be configured before running IRIS, which are the range of length k for short stems, the fraction of PARIS support in $\boldsymbol{Y}$ as the threshold for filtering stems, and the number of clusters C. Empirically, the number of clusters is set to 100 for all 11 RNAs in our benchmarking experiments. But the other two parameters vary according to the length of the input RNA. For RNAs less than 160 nt in length, k takes values from 3 to 6 and the fraction threshold is 0.5. For RNAs longer than 160 nt but shorter than 300 nt, k is from 4 to 7 and the fraction threshold is 0.75. Finally, for RNAs longer than 300 nt, the value of k is 4 to 7 and the fraction threshold is 0.90.

S3 Additional results

We show the results concerning RMRP as an additional case, as shown in **Figure S1**. From Figure S1A, we note that all structures predicted by IRIS contain a long continuous stem structure with the center located near 150 nt. The reason is that this

region has PARIS support enriched (Figure S1C), and it can be verified by evidence of evolutionary conservation (marked by the black arrow in Figure S1B).

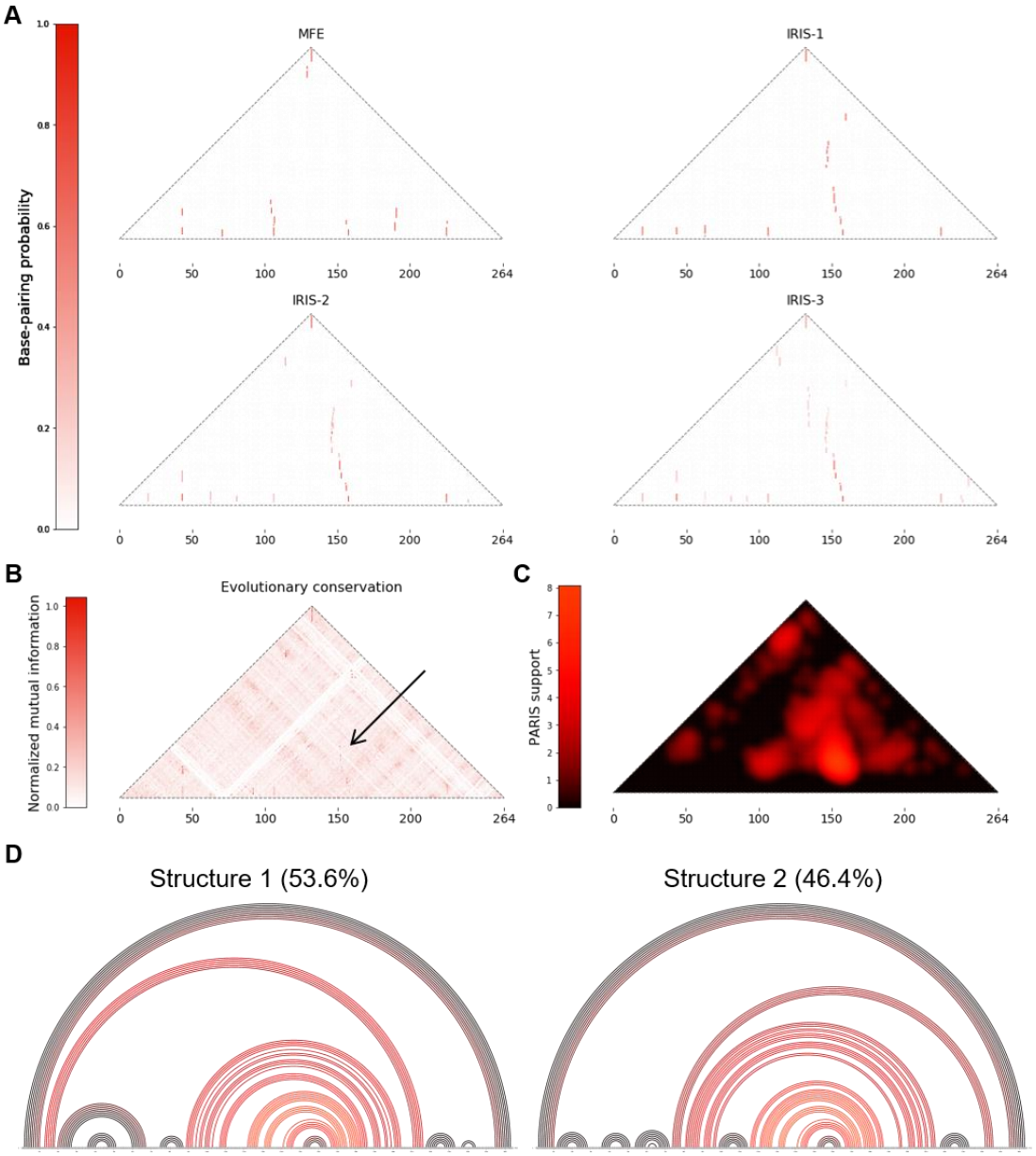

Figure S1 Results concerning the RMRP

A. The base-pairing probability matrices of the predicted ensemble from MFE, IRIS-1, IRIS-2 and IRIS-3. **B.** The matrix of normalized mutual information. **C.** The matrix of PARIS support. **D.** The ensemble of two representative structures predicted by IRIS-2.