Enforcing Constraints in RNA Secondary Structure Predictions: A Post-Processing Framework Based on the Assignment Problem

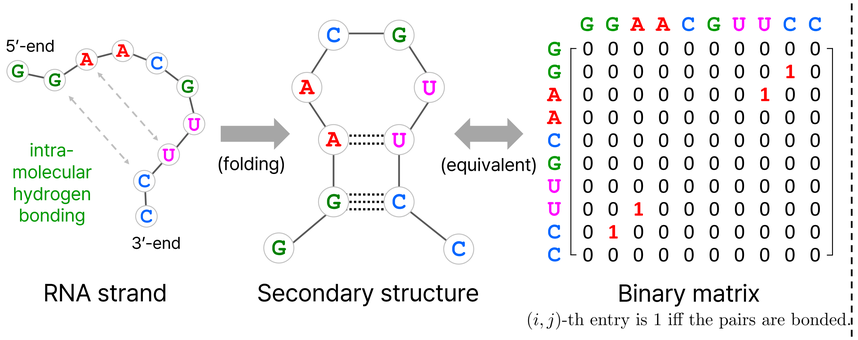
SPIDERCORE [1] KAIST [2]

Geewon Suh*[1,2] Gyeongjo Hwang*[1] Seokjun Kang[1] Doojin Baek[2] Mingeun Kang[1,2]

Correspondence to: minkang@spidercore.io

ICML International Conference On Machine Learning

RNA Secondary Structure



(i, j)-th entry is 1 iff the pairs are bonded.

Q. Do RNA structures matter? If yes, why?

A. Yes, as structure is the main determinant of the function of RNAs.

→ Knowing RNA structure is crucial for advances in biotechnology.

Problem Setting

Primary structure (as a 1D sequence of nucleotides) $\boldsymbol{x} = (x_1, \ldots, x_L)$ → ML Model → $\hat{\boldsymbol{Y}} \in \mathbb{R}^{L \times L}$ Score matrix (as a 2D matrix)

Train such a model with a collection of experimentally validated samples

$$\mathcal{D} = \{(\boldsymbol{x}^{(i)}, \boldsymbol{Y}^{(i)}\}_{i=1}^{m}$$

measured via X-ray crystallography / nuclear magnetic resonance

Considering the natural constraints governed by *physical laws:*

(C1) Binary & symmetric

$$Y_{ij} \in \{0, 1\} \text{ for } \forall i, j \text{ and } \boldsymbol{Y} = \boldsymbol{Y}^T,$$

(C2) Watson-Crick & Wobble base pairs only (Watson-Crick base pairs: AU, UA, GC, CG; Wobble base pairs: GU, UG)

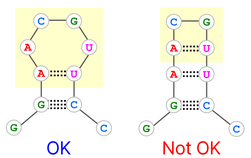
$$Y_{ij} = 0 \text{ if } x_i x_j \notin \mathcal{B} := \{\texttt{AU, UA, GC, CG, GU, UG}\}$$

(C3) No sharp loops

$$Y_{ij} = 0 \text{ if } |i - j| < 4$$

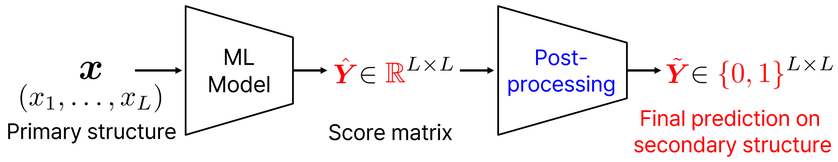
(C4) No overlap of pairs

$$\sum_{j=1}^{L} Y_{ij} \leq 1 \forall i$$



Prior Works & Challenge

E2Efold proposes a two-stage approach:



where post-processing for the constraints is presented as an optimization:

$$\begin{aligned} \max_{\tilde{\boldsymbol{Y}}} \quad & \langle \hat{\boldsymbol{Y}} - s, \tilde{\boldsymbol{Y}} \rangle \\ \text{s.t.} \quad & \tilde{Y}_{ij} \in \{0, 1\}, \quad \tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{Y}}^T, \quad \tilde{\boldsymbol{Y}}\mathbf{1} \leq \mathbf{1}, \\ & \tilde{Y}_{ij} = 0 \quad \text{if } x_i x_j \notin \mathcal{B} \text{ or } |i - j| < 4. \end{aligned}$$

Predictions made by SOTA algorithms still violate (C2) – (C4)

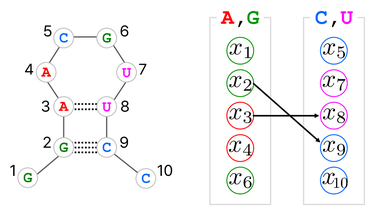
such as CNNfold, E2Efold, Ufold, REDfold

Contributions

Propose a mathematically equivalent post-process optimization based on the assignment problem

cost matrix obtained from $\hat{\boldsymbol{Y}}$ (assignment matrix)

$$\begin{aligned} \min_{\boldsymbol{Z}} \quad & \langle \boldsymbol{C}, \boldsymbol{Z} \rangle \\ \text{s.t.} \quad & \boldsymbol{Z}_{ij} \in \{0, 1\}, \\ & \boldsymbol{Z}\mathbf{1} = \mathbf{1}, \quad \boldsymbol{Z}^T\mathbf{1} = \mathbf{1}. \end{aligned}$$



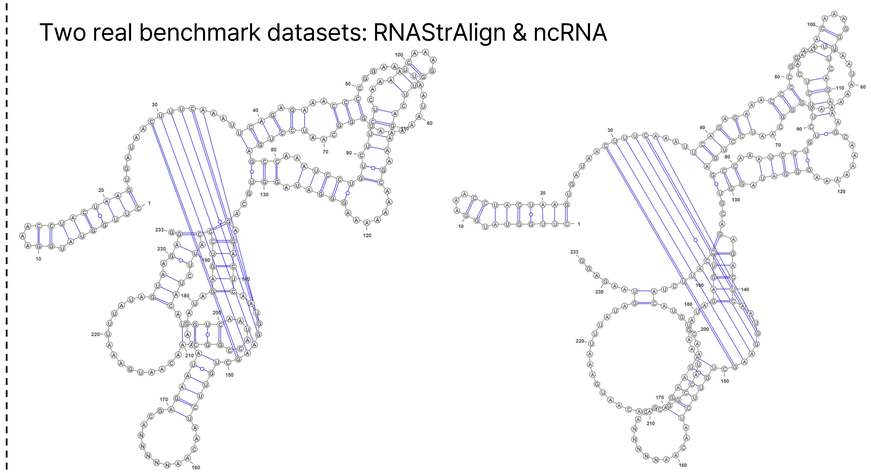
Theorem 1 (Equivalence)
Any optimal solution of each optimization can be reduced to an optimal solution of the other problem.

→ Employ off-the-shelf algorithms such as **Hungarian algorithm / Zonker-Volgenant algorithm** s.t. we obtain:

1. Model-agnostic add-on to any backbone ML model
2. Outputs completely adhering to the fundamental constraints of RNAs
3. Improved predictive performance & empirical running time

Results

Two real benchmark datasets: RNAStrAlign & ncRNA



Method	Constraint violation			Prediction performance			Run time
	(C2)	(C3)	(C4)	F1	Recall	Precision	Time (s)
RNAfold	-	-	-	0.606	0.679	0.566	-
RNAstructure	-	-	-	0.599	0.668	0.562	-
CONTRAfold	0%	8.1%	0%	0.626	0.690	0.596	-
SPOT-RNA	57.7%	30.3%	0%	0.647	0.683	0.640	-
MXfold2	-	-	-	0.631	0.687	0.608	-
E2Efold + E2E PP	0%	12.8%	58.3%	0.595	0.575	0.631	0.049
E2Efold + Blossom	73.0%	21.0%	0.3%	0.489	0.615	0.415	2.212
E2Efold + **Ours**	**0%**	**0%**	**0%**	0.608	0.602	0.622	0.308
REDfold + RED PP	0%	0.5%	0%	0.844	0.849	**0.877**	0.053
REDfold + Blossom	9.4%	0.3%	0%	0.840	**0.873**	0.838	0.378
REDfold + **Ours**	**0%**	**0%**	**0%**	**0.847**	0.867	0.858	**0.005**