Learning RNA structure prediction from crowd-designed RNAs

RNA molecules designed by citizen scientists and probed in high-throughput experiments highlighted discrepancies among RNA folding algorithms in their ability to predict RNA structure ensembles. These datasets were used to train a new algorithm that demonstrated improved performance in a collection of independent datasets, including viral genomic RNAs and mRNAs probed in cells.

This is a summary of:
Wayment-Steele, H. K. et al. RNA secondary structure packages evaluated and improved by high-throughput experiments. *Nat. Methods* https://doi.org/10.1038/s41592-022-01605-0 (2022).

Published online:
Published online: 3 October 2022

The problem

RNA folding involves the formation of Watson–Crick–Franklin base pairs, typically referred to as secondary structure. Since the introduction of the Nussinov algorithm to enumerate RNA base-pairing states[1], algorithms to predict RNA structures have been in continuous development. Today, RNA structure algorithms are workhorses of molecular biology and biotechnology, having implications across scientific and clinical fields including gene regulation, therapeutics and diagnostics. These algorithms have traditionally been evaluated on their ability to predict single structures from databases of natural RNAs. However, RNA molecules can adopt multiple structures, a fact not captured when scoring algorithms using only single predicted structures. Being able to predict the complete set of possible molecular structures and their relative weights (termed the RNA structural ensemble) is paramount for using these algorithms in design and analysis.

The observation

We asked how a range of RNA secondary structure algorithms, varying from widely used nearest-neighbor models to recent deep-learning-based approaches, performed in their ability to predict two types of ensemble properties. The first, chemical mapping data, measures how likely a nucleotide is to be unpaired, averaged over all possible structures. The second data source came from equilibrium binding constants of synthetic riboswitch molecules that had been designed to bind a fluorescent protein, and that were synthesized and probed using the massively high-throughput RNA-MaP platform[2]. For both types of data — the chemical mapping and riboswitch affinity experiments — the thousands of RNA sequences had been designed by participants of the online RNA design project Eterna[3]. We used these data sources because they represent, to our knowledge, the largest collections of diverse RNA sequences with accompanying structure-related experimental data.

We found that, in both tasks, the package CONTRAfold[4] consistently performed best. This was a surprise, as CONTRAfold is a model that had its parameters fit by maximizing the likelihood of single structures from a database of natural RNAs. Of note, CONTRAfold does not make use of biophysical RNA thermodynamics measurements that are typically considered the gold standard for understanding RNA folding and fluctuation. We noted that CONTRAfold used a training framework that we hypothesized could be updated to 'learn' from other data sources. With this in mind, we updated CONTRAfold's code to also maximize the likelihood of the chemical mapping and riboswitch affinity data, hoping to further improve its performance on these ensemble-averaged observables (Fig. 1a). Though the two types of data had not been designed to encompass as much RNA sequence or structure space as possible, we found that performing multitask training on both these data types resulted in a model (which we term EternaFold) that demonstrated improved performance on a collection of 31 published datasets of RNA structure mapping data from other groups, including full-length RNA genomes and mRNAs probed in cells and in viral particles (Fig. 1b).

Future directions

While EternaFold is not built as an artificial neural network, its training is much closer in spirit to modern neural network approaches that learn from large data sets of crowdsourced image or text[5] than to classic biophysical approaches for improving RNA secondary structure prediction from lower throughput measurements and the intuition of a few human experts. In fact, we hope that the EternaFold model presented in this work will be readily superseded by new algorithms developed with these prediction tasks in mind that account for molecule ensemble of all possible structures.

Important improvements abound for future models, including incorporating effects of ionic conditions and temperature. Perhaps the most significant leap for RNA structure prediction will be to incorporate prediction of tertiary structure motifs into secondary structure modelling. Many state-of-the-art 3D structure prediction and structure refinement methods require accurate secondary structure predictions as a starting point. An end goal of the field is to perform end-to-end inference from sequence to atomistic structure, which training from large collections of ensemble-based measurements such as these may enable.

Hannah K. Wayment-Steele
Harvard Medical School, Boston, MA, USA.
Rhiju Das
Stanford University School of Medicine, Palo Alto, CA, USA.

EXPERT OPINION

"The manuscript by Wayment-Steele et al. performed a rigorous comparison of a diverse array of secondary structure prediction programs." **Hashim Al-Hashimi, Columbia University Irving Medical Center, New York, NY, USA.**

FIGURE

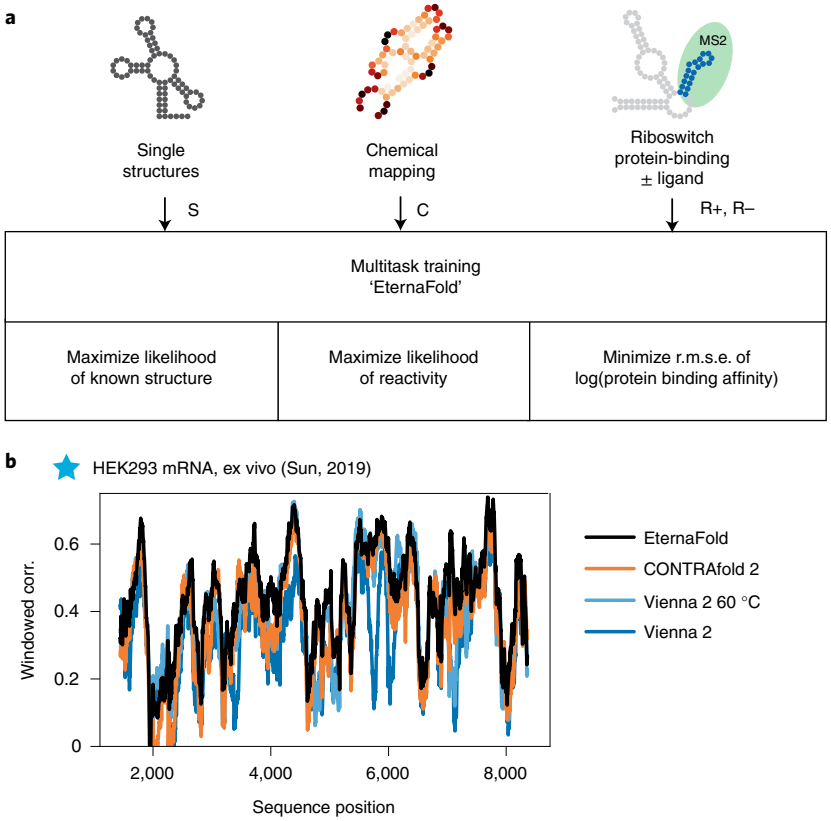

Fig. 1 | Multitask training improves prediction of ensemble-averaged base-pairing. a, Schematic of RNA data types used in multitask training of the EternaFold algorithm and loss functions used for each data type. R.m.s.e., root mean-squared error. **b**, Example prediction of mRNA for ribosomal protein S27A from HEK293 cells probed ex vivo, showing that EternaFold unpaired probabilities demonstrate higher correlation (corr.) to chemical mapping signal across sequence position than those of top-performing RNA structure prediction algorithms.

BEHIND THE PAPER

One of my first conversations with R.D. was in front of pages and pages of Eterna chemical mapping data that hung in the Stanford Biochemistry Department hallway. He pointed out eccentricities in these RNA data, collected over years of experiments, and alluded to a dream of actually inferring thermodynamics from these molecules designed by the Eterna community — molecules with names like "The Nonesuch" and "Robot Serial Killer 1." These datasets represented a massive, curiosity-driven, community labor of love. An exhilarating moment came in testing EternaFold on data from the influenza A virus and realizing it performed best on this very important RNA. We kept testing datasets from other groups to see if this was a fluke, including SARS-CoV-2 genomes (an unexpected test that emerged after EternaFold's development); 31 tests later, we concluded the model ought to be shared. **H.K.W.-S.**

REFERENCES

1. Nussinov, R. et al. Algorithms for loop matchings. *SIAM J. Appl. Math.* **35**, 68–82 (1978).
   This work presents a dynamic programming approach combined with energy weights to predict RNA base-pairing structures.
2. Andreasson, J. O. L. et al. Crowdsourced RNA design discovers diverse, reversible, efficient, self-contained molecular switches. *Proc. Natl Acad. Sci. USA* **119**, e2112979119 (2022).
   This paper describes the Eterna riboswitch experiments, which formed the basis of our RNA folding evaluation using riboswitch protein affinities.
3. Lee, J. et al. RNA design rules from a massive open laboratory. *Proc. Natl Acad. Sci. USA* **111**, 2122–2127 (2014).
   This paper describes the Eterna project and its initial results in uncovering new rules for RNA design.
4. Do, C. B. et al. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics* **22**, e90–e98 (2006).
   This work describes the CONTRAfold algorithm, a conditional log-likelihood model with thermodynamic parameters inferred from their abundance in a natural RNA dataset.
5. Gao, J. et al. A survey on deep learning for multimodal data fusion. *Neural Comput.* **32**, 829–864 (2020).
   This review discusses deep learning on combined multimodal data streams (text, images, video) and neural network architectures.

FROM THE EDITOR

"Predicting RNA secondary structure is an important problem in biophysics and is also crucially important for understanding the structure and biological function of diverse RNAs. What impressed me immediately about this work was how much could be learned by comparing the performance of available software tools for predicting RNA structures. I am also convinced that EternaFold, the newly developed prediction tool, enables improved prediction for diverse downstream applications." **Rita Strack, Senior Editor, *Nature Methods*.**