bioRxiv preprint doi: https://doi.org/10.1101/152629; this version posted July 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Learning RNA secondary structure (only) from structure probing data

Chuan-Sheng Foo*†[1] and Cristina Pop†‡[2]

[1]Department of Computer Science, Stanford University
[2]Google Inc.

1 Introduction

The structure of an RNA molecule is critical to its function. Structured regions in an RNA molecule permit or impede the binding of proteins and small molecules, resulting in downstream effects on gene expression [24, 16]. However, accurate determination of RNA structure is an involved process, requiring tedious experimentation [8, 1, 10] or careful analysis of evolutionary constraints [9]. As such, several computational methods have been developed to determine RNA structure in silico. In this work, we focus on methods that predict an RNA's secondary structure - the set of complementary Watson-Crick basepairs. These methods may be broadly classified into energy-based models [27, 14, 19], that compute a minimum free-energy secondary structure using experimentally-derived energies, and statistical models, which rely on data from a training set of sequences and their known structures in order to learn a model of secondary structure. In general, statistical methods for RNA secondary structure prediction outperform energy-based methods [20, 18]. However, their performance is found to be limited by the availability of training data [20].

The advent of high-throughput sequencing has led to the development of several genome-wide RNA structure-probing assays [26, 11, 23, 13, 5, 21] which could help plug this data gap. Such assays reveal which nucleotides are paired and which are not, but do not provide complete secondary structures. As such, existing statistical methods are unable to utilize these data for model learning. In this work, we present CONTRAfold-SE, which extends the statistical model of CONTRAfold [6] to model the structure-probing data as observations of possibly unknown secondary structures. This model can then be learned from datasets containing only structure-probing data, or a mix of known structures and probing data. We train CONTRAfold-SE on various combinations of structure probing data and complete structures and find that while genome-wide structure probing data provides modest improvement in prediction performance, with sufficiently dense probing data alone it is possible to learn a model that approaches the performance of energy-based methods. CONTRAfold-SE may be obtained from https://github.com/csfoo/contrafold-se.

2 Methods

Modelling structure-probing data as observations of secondary structure

Our model assumes that structure-probing data are available at per-base resolution – that there is a probing signal for some set of bases in a given RNA sequence. While our probabilistic framework allows the use of any distribution in the exponential family to model the (processed) probing signal at each base, we chose the Gamma distribution since it is flexible in modelling the positive, continuous, and unbounded probing signal. In addition, the Gamma distribution fits the data well, and has been previously used

*Corresponding author (csfoo@cs.stanford.edu)
†Equal contribution
‡This work was carried out while the author was at the Department of Computer Science, Stanford University

bioRxiv preprint doi: https://doi.org/10.1101/152629; this version posted July 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

to model signal from various probing assays [2]. We also assume that bases are independently modified, so that the resultant probing signals are independent of the actual location within the RNA sequence. However, as the reactivities of different bases to the probing mechanism could differ based on their identity and whether they are paired (for instance, DMS preferentially modifies unpaired adenines and cytosines), we have incorporated a separate distribution for each combination of base identity (A, C, T or G) and pairedness state (paired or unpaired) for a total of 8 separate Gamma distributions in our data model. Formally, for an RNA sequence x of length L with secondary structure y and associated structure-probing data $d = (d_1, \ldots, d_L)$, the distribution for the probing signal d_k at base k in the sequence is given by

$$d_k | x_k, y \sim \text{Gamma}(\alpha_{x_k, paired(k,y)}, \beta_{x_k, paired(k,y)}) \tag{1}$$

where $x_k \in \{A, C, T, G\}$ is the identity of base k in the sequence, $paired(k, y)$ denotes whether base k in structure y is paired, and the Gamma density is defined as $\frac{x^{\alpha-1}}{\Gamma(\alpha)\beta^\alpha} \exp(-x/\beta)$, for $x \sim \text{Gamma}(\alpha, \beta)$.

The CONTRAfold-SE model and parameter estimation

Model specification Let x be an RNA sequence of length L with structure y and S associated structure-probing datasets d. We denote by $d_k^{(j)}$ the probing signal for the jth data source at base k in the sequence. CONTRAfold-SE models the conditional joint probability of structure and probing data given sequence as

$$P(y, d | x; w, \theta) = P(y | x; w) \prod_{j=1}^{S} \prod_{k=1}^{L} P(d_k^{(j)} | x_k, y; \theta^{(j)}). \tag{2}$$

Here, the structure model $P(y|x; w)$ is given by the conditional log-linear model of CONTRAfold with parameters w, and $P(d_k^{(j)} | x_k, y; \theta^{(j)})$ is the Gamma distribution as defined in Equation 1, with $\theta^{(j)}$ being the vector of parameters for the 8 Gamma distributions for dataset j. In the absence of structure-probing data, the CONTRAfold-SE model reduces to the CONTRAfold model.

Parameter estimation Given a training set, we estimate the parameters, w and θ, by maximizing the conditional log-likelihood. Formally, for a training set $\mathcal{D} = \mathcal{D}_\mathcal{S} \cup \mathcal{D}_\mathcal{P} \cup \mathcal{D}_{\mathcal{S}+\mathcal{P}}$ of sequences with: i) only known structures and no probing data ($\mathcal{D}_\mathcal{S}$), ii) only probing data but unknown (missing) structure ($\mathcal{D}_\mathcal{P}$), and iii) both known structure and probing data ($\mathcal{D}_{\mathcal{S}+\mathcal{P}}$), we find w, θ that maximize the (regularized) conditional log-likelihood

$$\sum_{(x,y) \in \mathcal{D}_\mathcal{S}} \log P(y|x; w) + \lambda \cdot \sum_{(x,d) \in \mathcal{D}_\mathcal{P}} \log \sum_y P(y, d|x; w, \theta) + \sum_{(x,y,d) \in \mathcal{D}_{\mathcal{S}+\mathcal{P}}} \log P(y, d|x; w, \theta) + \log P(w) + \log P(\theta). \tag{3}$$

The hyperparameter λ (set by cross-validation) controls the weighting of data-only training instances as compared to instances with known structure, thus mitigating the adverse effects of noisy, partial data on model estimation; this strategy is common in the machine learning literature [17]. We used the L-BFGS algorithm [12] to find a local maximum of the likelihood. The key technical challenge is that the gradient computations for the second term in the sum (the likelihood for sequences with unknown structures) requires inference. Fortunately, the log-linear form of the structure model allows the data model to be represented as additional base-level features in the structure model. [7] independently presents a similar observation in the context of energy-based models for RNA structure, which have a similar log-linear form. We thus adapted the existing inference algorithms in CONTRAfold to efficiently compute the required gradients. While the log-likelihood for the CONTRAfold-SE model is non-convex, in practice our gradient-based parameter estimation algorithm achieves stable parameter estimates. A detailed description of the estimation procedure is found in Supplementary Methods.

bioRxiv preprint doi: https://doi.org/10.1101/152629; this version posted July 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

3 Experiments

We evaluate CONTRAfold-SE in two scenarios: 1) with training sets of known structures augmented with data from genome-wide structure probing experiments, and 2) with a training set consisting solely of probing data. In each scenario, we determine the prediction performance of CONTRAfold-SE on TestSetA and TestSetB obtained from [20], which contain structurally disjoint sets of RNA structures from the RFAM database. We constructed training and test sets to minimize structural overlap in order to more accurately assess generalization performance. Details on the construction of these training datasets is included in the Supplementary Methods.

We evaluate methods based on the following per-sequence metrics averaged over the test set: sensitivity, positive predictive value (PPV; also known as precision), F-measure (harmonic mean of sensitivity and PPV), and accuracy. As CONTRAfold-SE, like CONTRAfold, offers a sensitivity-PPV tradeoff via a hyperparameter γ, which predicts more basepairs with increasing γ, we also compare the area under the sensitivity-PPV curve (AUC); for metrics other than AUC, we show the maximum value across varying γ.

Dataset of known structures augmented with genome-wide probing-data

The training sets used in this scenario have two components making up 238 sequences: half with only known secondary structure and half with only structure-probing data. Probing-data only sequences are chosen to be the most data dense sequences across the probing datasets we use – the parallel analysis of RNA structure (PARS) [11] and DMS-seq assays [21]. We chose these genome-wide datasets as they were performed on the same set of (yeast) RNAs, allowing us to showcase the ability of CONTRAfold-SE to integrate multiple datasets, and explore the effects of such integration.

In the PARS assay, the RNA structure signal is obtained by treating RNA with enzymes that preferentially cleave either paired or unpaired nucleotides. The DMS-seq assay relies instead on the reactivity of unpaired nucleotides (primarily A and C) to the dimethyl-sulfate chemical; the DMS-seq assay was applied to both renatured RNA and live yeast. We denote these sources PARS, DMS-vitro, and DMS-vivo, respectively. We compare CONTRAfold-SE to CONTRAfold trained only on the set of sequences with known structures.

	TestSetA			**TestSetB**		
Probing data	AUC	F-measure	Accuracy	AUC	F-measure	Accuracy
None (CONTRAfold)	0.7110	0.7122	0.7610	0.6178	0.6478	0.7498
PARS	0.7209	0.7203	**0.7662**	**0.6256**	0.6554	0.7535
DMS-vitro	0.7126	0.7128	0.7616	0.6201	0.6499	0.7514
DMS-vivo	0.7096	0.7105	0.7604	0.6172	0.6483	0.7498
PARS, DMS-vitro	**0.7240**	**0.7214**	**0.7662**	0.6252	**0.6558**	**0.7551**

Table 1: Performance on probing-data augmented training dataset.

We find that while including a single source of structure-probing data has minimial effect on prediction performance compared to training on known structures alone, we obtain an improvement when using the two in-vitro data sources (Table 1).

Dataset of only probing-data from RMDB

We next explored the possibility of learning a structure model using structure-probing data alone. As we and others [25] have observed that the PARS and DMS-seq genome-wide datasets we used in the previous section are sparse and have low signal-to-noise, we used datasets obtained with the method of selective 2-hydroxyl acylation analyzed by primer extension (SHAPE) [15] for structure-probing instead. This data

bioRxiv preprint doi: https://doi.org/10.1101/152629; this version posted July 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

TestSetA	Sens	PPV	F1
CONTRAfold-SE	0.47	0.44	0.46
UNAfold	0.56	0.49	0.52

TestSetB	Sens	PPV	F1
CONTRAfold-SE	0.53	0.45	0.49
UNAfold	0.61	0.48	0.53

Table 2: Performance when model is learned only with structure probing data. We report sensitivity (Sens), positive predictive value (PPV) and the F-measure (F1).

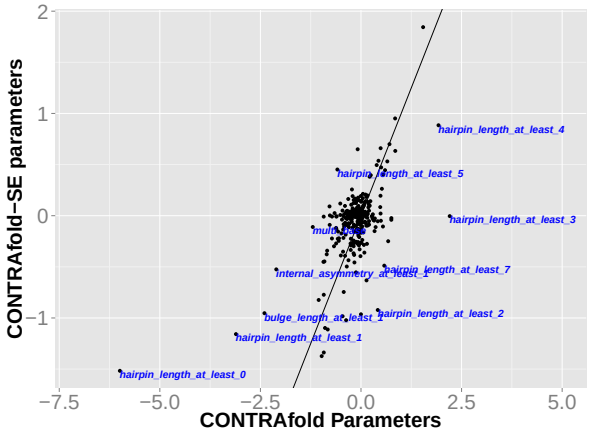


Figure 1: Comparison of parameters learned on structure-probing data alone (CONTRAfold-SE) with parameters learned on known structures (CONTRAfold).

is obtained from a single RNA species at a time, resulting in higher data density, and is also thought to be less biased [7]. We collected 17000 training sequences probed with the SHAPE method from the RNA Mapping DataBase (RMDB) [3] that had a high signal-to-noise ratio, and were not structurally similar to the test sets. We tuned the regularization hyperparameter (for L2-regularization on w) and performed early stopping using as a validation set the other test set not used for evaluation (*i.e.* TestSetA was the validation set when evaluating on TestSetB and vice versa).

While CONTRAfold-SE performs worse than UNAfold [14] (Table 2), a widely-used energy-based prediction method, this result was achieved *without the use of known structures* in the training set. In addition, comparing the learned parameters, we see that they are highly similar to those from a CONTRAfold model trained with known structures (Figure 1).

4 Discussion

We showed how structure-probing data can be used to learn models of RNA secondary structure. Our principled probabilistic framework enables easy integration of multiple probing datasets and handles situations where probing data or known structures are not available. While a similar approach has been used for improving secondary structure predictions with multiple probing datasets [22], the focus of that work was to model the characteristics of the probing data, and not to improve the underlying structure model. That CONTRAfold-SE is able to approach the performance of UNAfold when trained only on sequences with structure-probing data (without known structures) suggests that large collections of structure probing data could potentially be used to complement existing structures for training more sophisticated models. Finally, the sequences in the RMDB dataset we used were short (at most 103bp), and do not comprise a diverse set of RNAs. We speculate that with high-quality probing data from a diverse set of RNAs of various lengths, CONTRAfold-SE will be able to surpass the performance of energy-based models, and possibly that of statistical models. Incorporating the dependence of the SHAPE probing data on local structure context [4] may also help close the performance gap.

Acknowledgements

We thank Rhiju Das, Daphne Koller, members of the Koller lab, and Tom Do for useful discussions.

bioRxiv preprint doi: https://doi.org/10.1101/152629; this version posted July 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

[1] J H Cate and J A Doudna. Solving large RNA structures by X-ray crystallography. *Methods in enzymology*, 317:169–80, January 2000.

[2] Pablo Cordero, Wipapat Kladwang, Christopher C VanLang, and Rhiju Das. Quantitative dimethyl sulfate mapping for automated RNA secondary structure inference. *Biochemistry*, 51(36):7037–9, September 2012.

[3] Pablo Cordero, Julius B. Lucks, and Rhiju Das. An RNA Mapping DataBase for curating RNA structure mapping experiments. *Bioinformatics*, 28(22):3006, 2012.

[4] Fei Deng, Mirko Ledda, Sana Vaziri, and Sharon Aviran. Data-directed RNA secondary structure prediction using probabilistic modeling. *RNA*, 2016.

[5] Yiliang Ding, Yin Tang, Chun Kit Kwok, Yu Zhang, Philip C Bevilacqua, and Sarah M Assmann. In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature*, November 2013.

[6] Chuong B Do, Daniel A Woods, and Serafim Batzoglou. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, 22(14):e90–e98, July 2006.

[7] Sean R Eddy. Computational analysis of conserved RNA secondary structure in transcriptomes and genomes. *Annual Review of Biophysics*, 43:433–456, January 2014.

[8] Boris Fürtig, Christian Richter, Jens Wöhnert, and Harald Schwalbe. NMR spectroscopy of RNA. *ChemBioChem*, 4(10):936–962, October 2003.

[9] Robin R Gutell, Jung C Lee, and Jamie J Cannone. The accuracy of ribosomal rna comparative structure models. *Current opinion in structural biology*, 12(3):301–310, 2002.

[10] Ailong Ke and Jennifer A Doudna. Crystallization of RNA and RNA-protein complexes. *Methods (San Diego, Calif.)*, 34(3):408–14, November 2004.

[11] Michael Kertesz, Yue Wan, Elad Mazor, John L Rinn, Robert C Nutter, Howard Y Chang, and Eran Segal. Genome-wide measurement of RNA secondary structure in yeast. *Nature*, 467(7311):103–107, September 2010.

[12] Dong C. Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1-3):503–528, August 1989.

[13] Julius B Lucks, Stefanie A Mortimer, Cole Trapnell, Shujun Luo, Sharon Aviran, Gary P Schroth, Lior Pachter, Jennifer A Doudna, and Adam P Arkin. Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proceedings of the National Academy of Sciences of the United States of America*, 108(27):11063–8, July 2011.

[14] Nicholas R Markham and Michael Zuker. UNAFold: software for nucleic acid folding and hybridization. *Methods in molecular biology (Clifton, N.J.)*, 453:3–31, January 2008.

[15] Edward J Merino, Kevin A Wilkinson, Jennifer L Coughlan, and Kevin M Weeks. RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE). *Journal of the American Chemical Society*, 127(12):4223–31, March 2005.

[16] Stefanie A Mortimer, Mary Anne Kidwell, and Jennifer A Doudna. Insights into RNA structure and function from genome-wide studies. *Nature reviews. Genetics*, 15(7):469–79, July 2014.

[17] Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning*, 39(2-3):103–134, May 2000.

[18] Tomasz Puton, Lukasz P Kozlowski, Kristian M Rother, and Janusz M Bujnicki. CompaRNA: a server for continuous benchmarking of automated methods for RNA secondary structure prediction. *Nucleic Acids Research*, 41(7):4307–4323, April 2013.

[19] Jessica S Reuter and David H Mathews. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, 11(1):129, January 2010.

[20] Elena Rivas, Raymond Lang, and Sean R Eddy. A range of complex probabilistic models for RNA secondary structure prediction that includes the nearest-neighbor model and more. *RNA*, 18(2):193–212, February 2012.

[21] Silvi Rouskin, Meghan Zubradt, Stefan Washietl, Manolis Kellis, and Jonathan S. Weissman. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature*, December 2013.

[22] Sudhakar Sahoo, Michal P. Switnicki, and Jakob Skou Pedersen. ProbFold: a probabilistic method for integration of probing data in RNA secondary structure prediction. *Bioinformatics*, 32(17):2626, 2016.

bioRxiv preprint doi: https://doi.org/10.1101/152629; this version posted July 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[23] Jason G Underwood, Andrew V Uzilov, Sol Katzman, Courtney S Onodera, Jacob E Mainzer, David H Mathews, Todd M Lowe, Sofie R Salama, and David Haussler. FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nature Methods*, 7(12):995–1001, December 2010.

[24] Yue Wan, Michael Kertesz, Robert C Spitale, Eran Segal, and Howard Y Chang. Understanding the transcriptome through RNA structure. *Nature Reviews Genetics*, 12(9):641–655, September 2011.

[25] Yang Wu, Binbin Shi, Xinqiang Ding, Tong Liu, Xihao Hu, Kevin Y. Yip, Zheng Rong Yang, David H. Mathews, and Zhi John Lu. Improved prediction of RNA secondary structure by integrating the free energy model with restraints derived from experimental probing data. *Nucleic Acids Research*, 43(15):7247, 2015.

[26] Qi Zheng, Paul Ryvkin, Fan Li, Isabelle Dragomir, Otto Valladares, Jamie Yang, Kajia Cao, Li-San Wang, and Brian D Gregory. Genome-wide double-stranded RNA sequencing reveals the functional significance of base-paired RNAs in Arabidopsis. *PLoS genetics*, 6(9):e1001141, September 2010.

[27] Michael Zuker and Patrick Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9(1):133–148, January 1981.