test

August 29, 2020

1 Introduction

dp: pros: hard-constraint satisfied by design (as reflected by valid recursion steps); models distribution cons: efficient dp recursion based on nesting assumption, cannot model pseudoknot; additive energy terms, hard-coded.

scfg extension: pros: constraints; models distribution trainable param cons: efficient dp recursion based on nesting assumption, cannot model pseudoknot

todo: plot explaining dp scfg approaches

nn: pros: expressive, can be trained on lots of data; cons: too many parameters, risk of overfitting if trained on dataset not diverse enough; hard constraints cannot be captured by default (some work using optimization as unrolled rnn, but no guarantee it'll converge, and even if it converges there's guarantee the output satisfies the constraints, since the original discrete variables need to be relaxed in order to run through the optimization nn, so we're no longer operating in discrete space); does not model distribution

todo: plot explaining nn approaches

In this work we are aiming at developing a model that makes use of the expressiveness of deep nn while respecting the hard constraints.

This report covers the following sections:

- predictive problem formulation
- we review the commonly used data representation for rna ss
- we propose an alternative way to represent/parameterized the ss
- under the alternative parametrization, we propose a 2-stage model, which satisfies the biological hard constraints by design
- describe stage 1 model: data, training result.
- ideas for stage2

1.1 Problem formulation

Given an RNA sequence of length L, we are interested in all possible secondary structures. To represent a specific secondary structure, there are three commonly used conventions: (1) undirected graph, where each node is a base in the sequence, and each edge represents base pairing. (2) upper triangular matrix (excluding the diagonal) of size $L \times L$ with binary values, where a value of 1 at (i, j) represents base pairing between sequence position i and j, and 0 represents no base paring.

(3) dot-bracket notation of length L where unpaired bases are denoted as dots, and paired bases are represented as left and right brackets.

As an example, for a short RNA sequence GUUGUGAAAU, one possible structure it can take on

consists of a stem and a loop, as seen in Fig **??**(a), represented by an undirected graph. Such structure can also be represented by a 10×10 upper triangular matrix with all 0's, except for positions $(1, 10), (2, 9)$ and $(3, 8)$, all being 1, as shown in Fig **??**(b). This contiguous stretch of 1's along the diagonal corresponds to the stem formed by the three base pairs: G-U, U-A and U-A. The equivalent dot-bracket notation is shown in Fig **??**(c), where the stem is represented by three pairs of left-right brackets.

1.2 Structural components

Just as sequence motif is defined in the linear sequence space, a similar notion of 'structural motif' also exists. A commonly adopted way is to break down structure into the following components:

- stem: todo
- hairpin loop: todo
- internal loop: todo
- bulge: todo
- external loop: todo
- multibranch loop: todo
- pseudoknot: todo

Such structural motifs can be easily identified on the graph representation of a structure, as shown in Fig 1, where we only annotated one instance for each structural motif class for clarity. (todo add a minimal example for pseudoknot)

In this work, we distinguish structural motifs that are 'local' and 'non-local', where local-ness is defined w.r.t. the 2D matrix representation of the structure. To illustrate this idea, we plot the 2D matrix corresponding to the above structure, as shown in Fig 2(a). For each type of structural motif, if it can be fully represented by the interaction between two (contiguous) substrings of the original sequence, it is considered a 'local' structure, and can be fully identified by a bouning box on 2D matrix representation. Note that interaction does not necessarily mean the substrings bind to each other. On the other hand, certain structure motifs can only be represented by interaction between more than two (contiguous) substrings of the original sequence, then it is a non-local structure.

Using the above defined, we drew all the local structure bounding boxes, as shown in Fig 2(b).

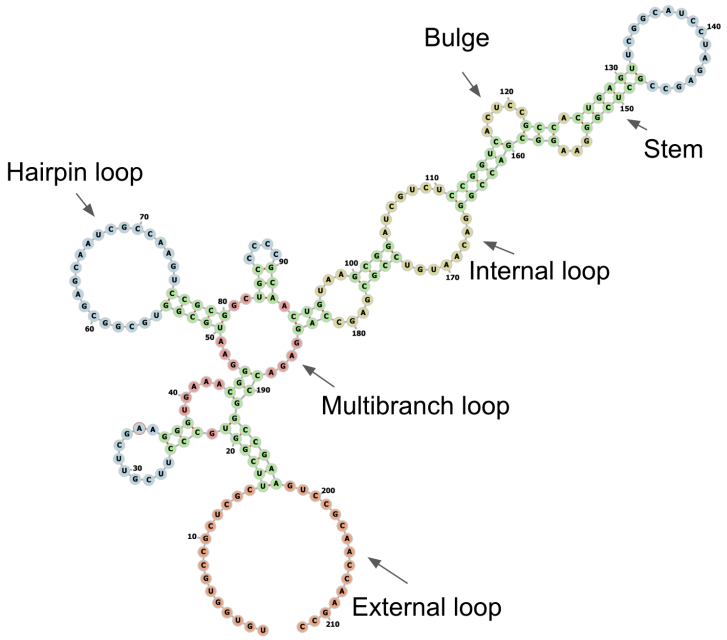

Figure 1: TODO

To precisely define what constitutes each type of local structure, we zoom in to each one annotated in Fig 1, and compare it with the corresponding zoomed-in 2D matrix, as shown in Fig 3. As we can see:

todo

Formal definition and minimal example

no need for additional parameters

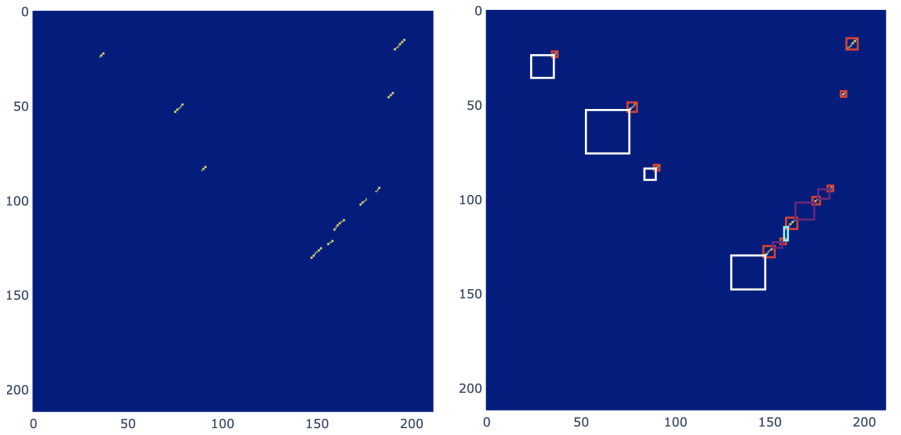

Figure 2: TODO

1.3 Related work

2 Method

2.1 Architecture

2.2 Training

2.3 Inference

3 Result

3.1 Test set performance

3.2 Structures with pesudoknot

References

[1] Ronny Lorenz, Stephan H Bernhart, Christian Höner Zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. Viennarna package 2.0. *Algorithms for molecular biology*, 6(1):26, 2011.

[2] Michael Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic acids research*, 31(13):3406–3415, 2003.

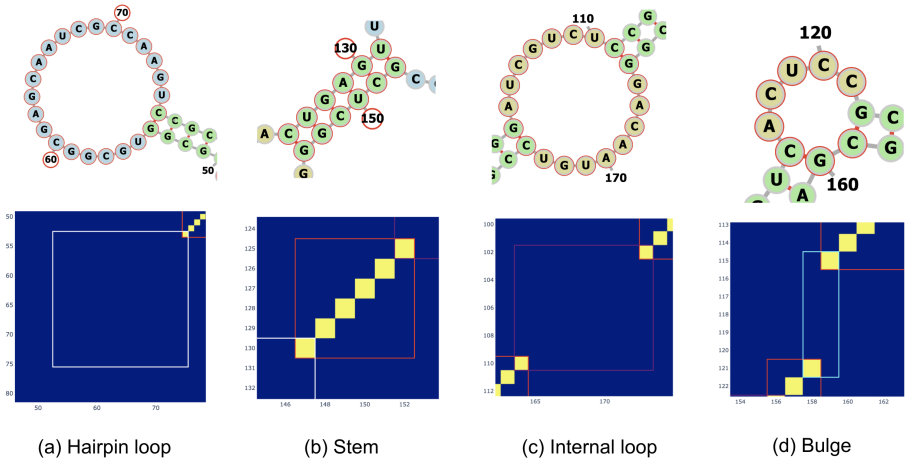

Figure 3: TODO

[3] Mirela Andronescu, Anne Condon, Holger H Hoos, David H Mathews, and Kevin P Murphy. Efficient parameter estimation for rna secondary structure prediction. *Bioinformatics*, 23(13):i19–i28, 2007.

[4] 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, 2014.

[5] Philip C Bevilacqua, Laura E Ritchey, Zhao Su, and Sarah M Assmann. Genome-wide analysis of rna secondary structure. *Annual review of genetics*, 50:235–266, 2016.

[6] Peter Kerpedjiev, Stefan Hammer, and Ivo L Hofacker. Forna (force-directed rna): simple and effective online rna secondary structure diagrams. *Bioinformatics*, 31(20):3377–3379, 2015.

[7] Michelle J Wu. Convolutional models of rna energetics. *bioRxiv*, page 470740, 2018.

[8] Devin Willmott, David Murrugarra, and Qiang Ye. State inference of rna secondary structures with deep recurrent neural networks.

[9] Hao Zhang, Chunhe Zhang, Zhi Li, Cong Li, Xu Wei, Borui Zhang, and Yuanning Liu. A new method of rna secondary structure prediction based on convolutional neural network and dynamic programming. *Frontiers in genetics*, 10, 2019.

[10] Linyu Wang, Yuanning Liu, Xiaodan Zhong, Haiming Liu, Chao Lu, Cong Li, and Hao Zhang. Dmfold: A novel method to predict rna secondary

structure with pseudoknots based on deep learning and improved base pair maximization principle. *Frontiers in genetics*, 10:143, 2019.

[11] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[12] David W Staple and Samuel E Butcher. Pseudoknots: Rna structures with diverse functions. *PLoS biology*, 3(6):e213, 2005.