RNA Secondary Structure Prediction using Deep Neural Network

TODO
Department of Electrical and Computer Engineering
University of Toronto
todo@toronto.edu

Abstract

TODO

1 Introduction

State-of-the-art methods based on dynamic programming. Basic building blocks are known local structure, with their associated free energy measured experimentally. Fixed energy parameters and hand-crafted rules.

Emerging new dataset, ... need for flexible, extensible, end-to-end model that can be trained on new types of dataset (noisy, missing value).

RNA secondary structure can be represented by a binary upper triangular matrix (excluding the diagonal).

As an example, a short RNA sequence GUUGUGAAAU of length 10 (ID `CRW_00083` TODO ref to database) takes a structure that consists of a stem and a loop, as seen in Fig 1(a). This structure can be represented by the upper diagonal 10×10 matrix with all 0's, except for positions $y_{1,10}$, $y_{2,9}$ and $y_{3,8}$, all having value 1. This contiguous stretch of 1's corresponds to the stem formed by the three base pairs: G-U, U-A and U-A. (TODO define x and y first) (TODO cite FORNA)

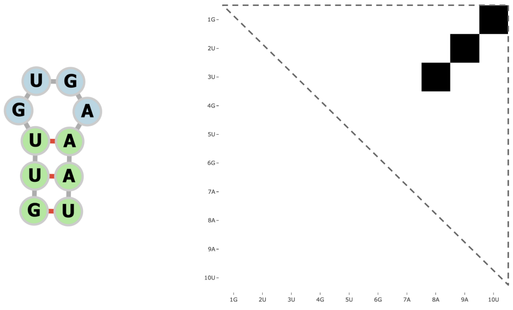

Figure 1: TODO

notation:

we use ??? to represent a binary upper triangular matrix of size LxL.

2 Problem Formulation

We formulate the predictive task as a conditional generative process? Specifically, given a input sequence with arbitrary length L: $\boldsymbol{x} = x_1, x_2, \ldots x_L$, we want to predict a distribution of structures conditioned on the sequence $P(\boldsymbol{Y} \mid \boldsymbol{x})$ (TODO use the above notation).

We factorize the generative process as follows:

$$P(\boldsymbol{Y} \mid \boldsymbol{x}) = P(\{y_{i,i+1}\}_{i=1,2,\ldots L-1} \mid \boldsymbol{x}) P(\{y_{i,i+2}\}_{i=1,2,\ldots L-2} \mid \boldsymbol{x}, \{y_{i,i+1}\}_{i=1,2,\ldots L-1}) \ldots P(y_{i,j} \mid \boldsymbol{x}, \{y_{todo}\})$$

Intuitively, this correspond to Fig ???. Binary More details.

intuition, sample local connectivity then global?

TODO plot for upper triangular generation process (just draw different slices, using the above simple example)

3 Model

TODO plot for NN architecture (high-level and low-level?)

batch, mask

Special constraints when sampling the matrix.

Describe dataset.

1-step AR.

differentialble model. use case.

4 Analysis

5 Conclusion

Speed consideration: implement split architecture, only need one forward pass for up to the last layer, then run last layer (triangular convolution) for $L - 1$ times.

Case study:

TODO compare with RNAfold performance

TODO reference on a new page