Predicting in vivo RNA Secondary Structure

by

Jiexin Gao

A thesis proposal submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Electrical Engineering
University of Toronto

Abstract

Predicting in vivo RNA Secondary Structure

Jiexin Gao
Doctor of Philosophy
Graduate Department of Electrical Engineering
University of Toronto
2019

Contents

Chapter 1

Introduction

Although once believed to be an intermediate molecule that serves as messenger between DNA and protein, RNA is now known to be involved in many aspects in gene regulation and expression. Unlike DNA which forms stable double helix, RNA predominantly stays single-stranded and folds onto itself by forming base pairs via hydrogen bonds, including Watson-Crick pair A-U, G-C and non-canonical TODO pair A-U. Nearby paired and unpaired bases and further leads to the formation TODO of hairpins, bulges, internal loops, Multi loops and pseudoknots. In addition, secondary structures within an RNA molecule can interact via non-covalent bond, to form tertiary structure.

RNA secondary and tertiary structure play key roles in regulating both coding and nocoding transcripts, and affect all steps of gene regulation including transcription, splicing, polyadenylation, translation, localization and stability[5]. Over the past few years, combination of RNA structure probing and high throughput sequencing has enabled the discovery? of genome-wide RNA structural in multiple organisms and cell types, ??? better understanding of the relationship between RNA structure and function.

1.1 RNA secondary structure and gene regulation

Polyadenylation

Ding et. al[1] used Structure-Seq (TODO review) to study the genome-wide RNA structure of *Arabidopsis thaliana* seedlings and discovered a pattern around alternative polyadenylation sites across 5959 mRNAs. As shown in Fig 1.1, region upstream of the site (-15nt to -2nt) is significantly more structured, as indicated by lower reactivity, and region downstream of the site (-1nt to +5nt) is significantly less structured, as indicated by higher reactivity.

1. Transcription
2. Splicing
3. Polyadenylation
4. Translation
5. Localization

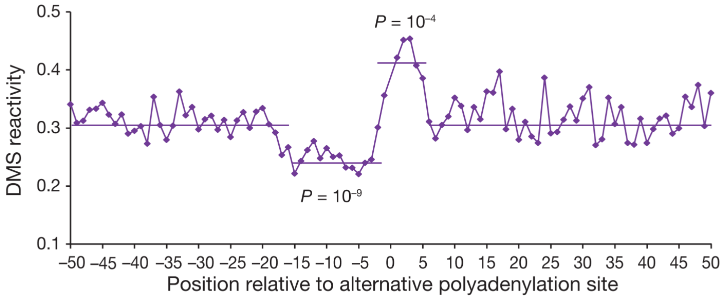

Figure 1.1: RNA structure aroung alternative polyadenylation sites in *Arabidopsis thaliana*. From Ding et. al[1].

6. Stability

1.2 High throughput probing of RNS secondary structure

It has been demonstrated in various organisms that in vivo structure can differ significantly from in vitro. In a living cell environment, the presence of proteins, ?ligand, salt, temperature?, all affects RNA structure, and he precise reconstruction of these conditions in vitro is almost impossible. Rouskin et al.[3] studied yeast RNA structure and found that RNAs are less structured in vivo than in vitro. Moreover, they observed that structures unfold when Mg^{2+} is lowered in vitro, and more structured regions emerge when cell is depleted of ATP in vivo. (which underscores the importance of in vivo probing that captures the specific physiological condition of the cell type of interest)

vast number of non coding RNAs (still being discovered)

constran prediction using data

1.3 Deep neural network

Chapter 2

Yeast Model

2.1 Training Dataset

To model in vivo RNA secondary structure, we compiled training data from [3]. In this study, yeast strain was treated with dimethyl sulphate (DMS), which reacts with unpaired adenine and cytosine bases. The pool of modified RNAs were fragmented and sequenced. Since DMS modification blocks reverse transcription, number of reads (TODO stops?) at each position is indicative of relative accessibility of that site.

Raw count data was downloaded from GSE45803 (`GSE45803_Feb13_VivoAllextra_1_15_PLUS.wig.gz` and `GSE45803_Feb13_VivoAllextra_1_15_Minus.wig.gz`). The authors aligned 25nt of each read to a non-redundant set of RefSeq transcripts, where each gene is represented by its longest protein-coding transcript. Only uniquely mapped reads with less than 2 mismatches were retained, and the authors further filtered out aligned reads whose RT stop is not A/C. The count at each position represents the combined number of RT stops at that site, across 4 biological replicates.

To construct training dataset, Saccharomyces cerevisiae assembly R61 (secCer2) RefSeq gene annotation was used to extract mRNA sequences. For each transcript, we first extract the raw read count for all adenine (A) and cytosine (C) bases (A/C positions with no RT stop coverage were set to a count of 0), and applied 90% Winsorization to remove outliers. Specifically, for each non-overlapping window of 100 A/C bases, values above the 95% percentile was set to the 95% percentile, and values below the 5% percentile was set to the 5% percentile. Then, all values within this window were divided by the max, to obtain values between 0 and 1.

2.2 Deep neural network

We construct a deep neural network to predict reactivity at single base resolution from RNA sequence context. We use an architecture similar to DenseNet[2], in which we've removed the pooling layers, to maintain the spatial resolution throughout the depth of the neural network.

As shown in Fig2.1, to make inference on a stretch of RNA sequence of length L, we need to pad the sequence with w bases on each side. (TODO explanation + how to calculate w) Input consists of the one-hot encoded, padded sequence, where A, C, G, U bases are encoded as $[1, 0, 0, 0]$, $[0, 1, 0, 0]$, $[0, 0, 1, 0]$, $[0, 0, 0, 1]$, respectively. The encoded input is then passed through multiple dense blocks, where each block consists

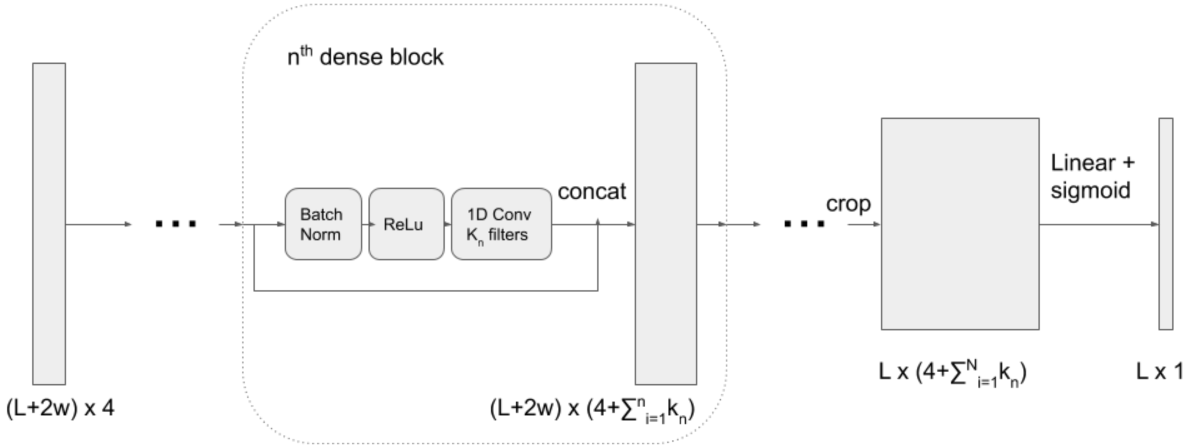

Figure 2.1: Densely connected neural network used for the yeast model

of four components:

1. Batch Normalization
2. ReLu nonlinear activation
3. 1D Convolution
4. Concatenation of the block input to the output of convolution

Block number	Number of filters	Filter width	Dilation rate
1	128	16	1
2	128	16	2
3	256	16	4
4	256	16	8
5	512	16	16

Table 2.1: Dense block parameters

We use 5 dense blocks in this work. The parameter of each layer is as shown in Table2.1. Densely connected block has the advantage that each block receives input from all preceding blocks, and passes its output to all successive blocks. The output of the last dense block essentially represents the features learnt from input at multiple resolutions.

The final dense block output is then cropped to account for the input padding, and then passed through a fully connected layer with sigmoid activation, along the feature dimension.

2.3 Training

Fold number	Chromosomes
1	chrM, chrVIII, chrII, chrXV
2	chrI, chrV, chrXIII, chrIV
3	chrVI, chrXI, chrXVI
4	chrIII, chrX, chrXII
5	chrIX, chrXIV, chrVII

Table 2.2: Chromosomes used for each fold

We use 5-fold cross validation, where the folds are splited by chromosomes, as shown in Table 2.2.

Normalized data points (between 0 and 1) are used as soft targets without being converted to binary labels, and models were trained using a masked cross-entropy loss, as described below.

Due to the nature of DMS modification, G/T bases has no coverage, thus should be excluded from the calculation of the loss and the gradient. This is achieve by first computing the per position cross-entropy loss between the prediction and the target, then multiply it with a binary mask with the same shape as the target array. Positions with G/T bases are being set to 0 in the mask, while positions with A/C bases are 1. The masked loss are then summed over positions, and minibatch dimension, to calculate the loss for the current minibatch and the gradient for back propagation.

Models were trained using fixed sequence length of 50 (before padding, sequence length at inference time can be variable), minibatch size of 10, Adam optimizer with learning rate 0.0001 and momentum 0.9. To prevent the models from overfitting, L1 and L2 regularizers with weight 0.000001 was added to the loss, and training is stopped if validation loss hasn't improved over the last 10 epochs.

We trained 5 models, each using one of the folds as validation data, and the rest as training data.

2.4 Performance

2.4.1 Cross-validation performance on training dataset

We first evaluate the model performance on training dataset. For each transcript, we used the model that wasn't trained on its chromosome to make prediction for all A/C bases. We computed the Spearman correlation between the prediction and the target for each transcript. Fig 2.2 shows the distribution of Spearman correlation across all transcripts.

2.4.2 Ribosomal RNA

Next we used our model to predict the reactivity for all A/C bases in yeast $18S$ and $15S$ ribosomal RNAs, both were never seen by the model (neither in the training nor validation set).

Raw read count data for $18S$ and $15S$ was downloaded from GSE45803, and was processed similarly to the training dataset, since the experimental protocol was identical.

Correlation between prediction and the normalized read count is shown in Fig2.3 and Fig2.4, where each data point is one A/C base in the corresponding transcript. In comparison, RNAfold (window size 50 and span 50) achieves a correlation of 0.3217 and 0.4529 for $18S$ and $15S$, respectively.

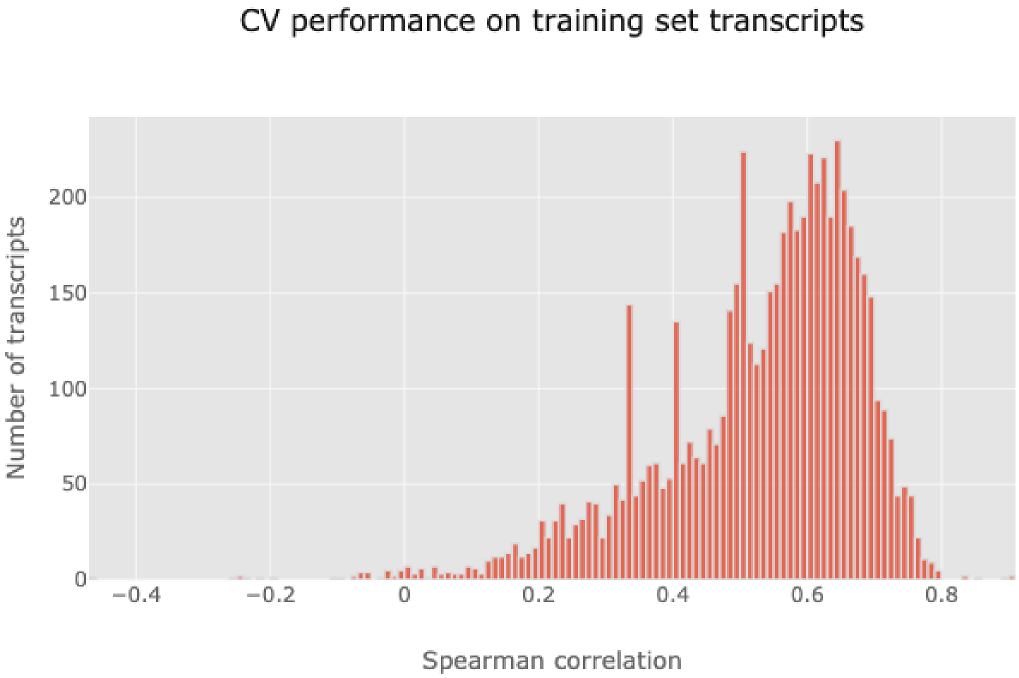

Figure 2.2: Densely connected neural network used for the yeast model

2.4.3 Noncoding RNAs

To evaluate whether the model generalizes to noncoding transcripts and different experimental protocol, we processed yeast data from the ModSeq paper[4], where yeast was treated with DMS or no-DMS (as control), and the authors identified positions that are significantly modified between treated and control, in selected noncoding and rRNA transcripts.

For each transcript, we use our models to predict on all A/C positions, and computed the au-ROC on how well the prediction distinguish the significantly modified bases from the rest. We also compare the performance of our model to that of RNAfold, as shown in Fig2.5.

2.5 Future Work

- Improve training and generalization performance, by making use of the raw sequencing data, and biological replicates. In additional to counts of RT stops, read coverage at each position can be used to infer the confidence of calling that position paired/unpaired. Transcript can be reweighted during training, according to the agreement between different biological reps.
- Multi-resolution learning.

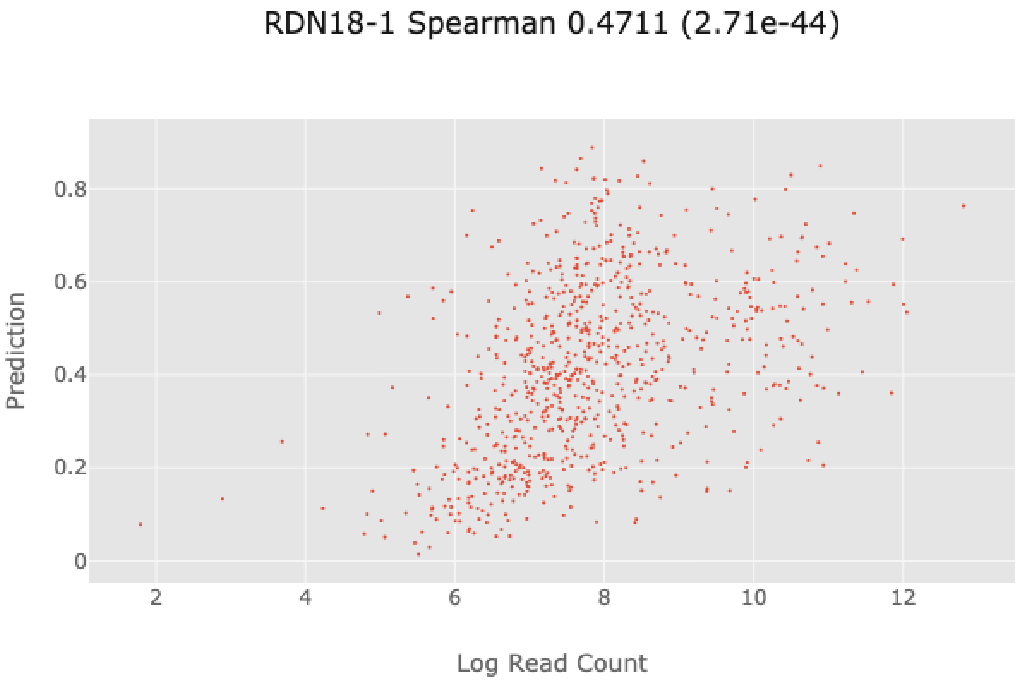

Figure 2.3: Densely connected neural network used for the yeast model

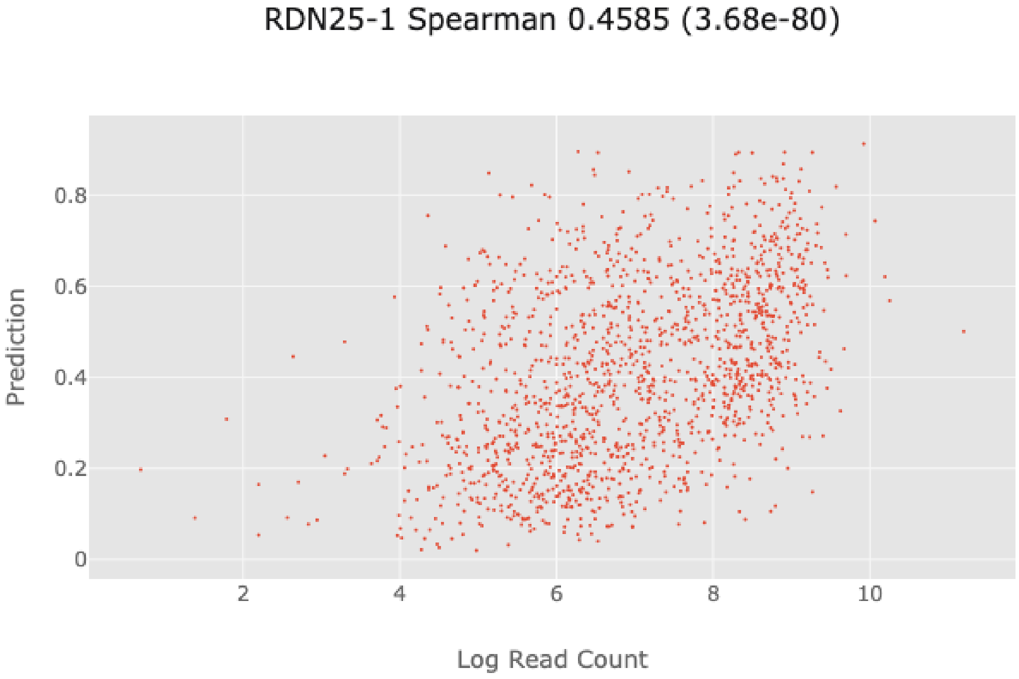

Figure 2.4: Densely connected neural network used for the yeast model

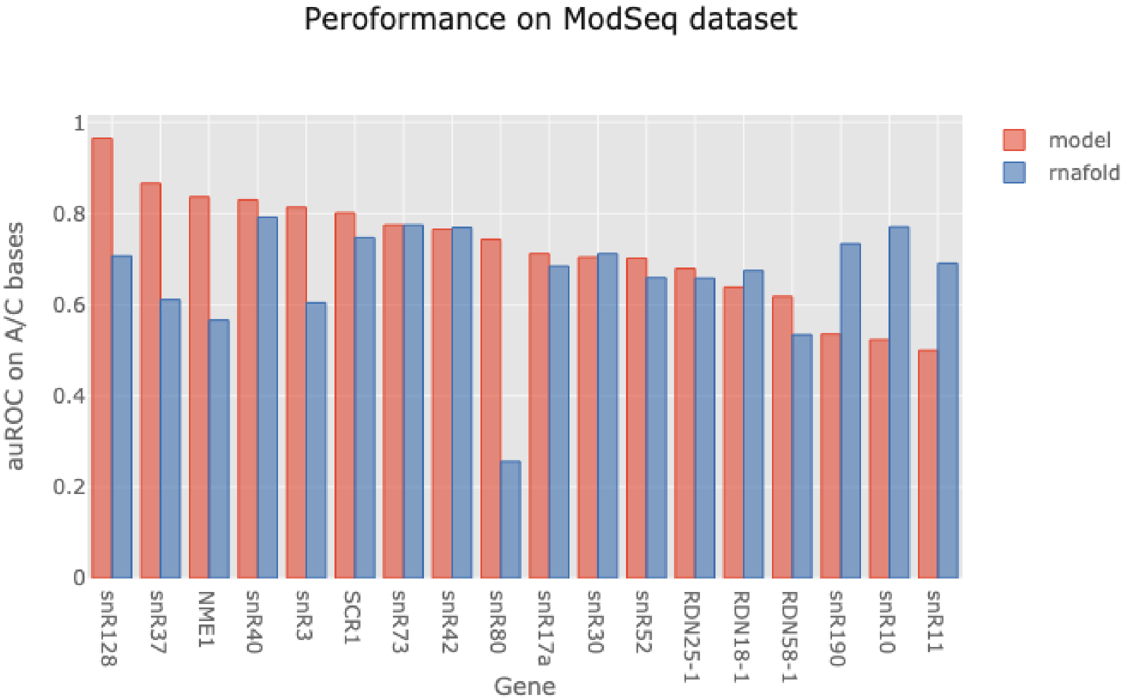

Figure 2.5: Densely connected neural network used for the yeast model

Chapter 3

Mouse Model

Chapter 4

Human Model

Chapter 5

Conclusion and future work

one dataset that has multiple mods per sequence, so we can reconstruct colleciton of structures

joint learning of accessibility and other data, e.g. chip-seq peaks

Bibliography

[1] 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*, 505(7485):696, 2014.

[2] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[3] 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*, 505(7485):701, 2014.

[4] Jason Talkish, Gemma May, Yizhu Lin, John L Woolford, and C Joel McManus. Mod-seq: high-throughput sequencing for chemical probing of rna structure. *Rna*, 20(5):713–720, 2014.

[5] 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, 2011.