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

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 pair A-U. Nearby paired and unpaired bases further leads to the formation 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[13, 8, 1]. Over the past few years, combination of RNA structure probing and high throughput sequencing has enabled the identification? of genome-wide RNA structural at single nucleotide resolution in multiple organisms and cell types, which provides new insight into the relationship between RNA structure and its function.?

1.1 High throughput probing of RNS secondary structure

Two types of reagents are being used to probe RNA structures: enzymes and chemicals. These reagents cleave or modify RNA at specific bases and/or structured regions.

Parallel analysis of RNA structures (PARS)[14] uses RNase S1 to cleave single-stranded regions, and RNase V1 to cleave double-stranded regions. Similarly, fragmentation sequencing (FragSeq)[**?**] utilizes RNase P1 to cleave single-stranded RNA. After high throughput sequencing and read alignment to the genome or transcriptome, the locations of cleavage can be determined. In PARS, the ratio between V1 and S1 is calculated (PARS score), and in FragSeq, the ratio between RNase P1 treated and untreated control is calculated. (TODO more on sequencing and meaning of score)

Dimethyl sulfide (DMS) reacts with unpaired adenine (A) and cytosine (C), and is being used in various protocols to modify RNA[10, 11]. selective 2-hydroxyl acylation analysed by primer extension (SHAPE) method uses the chemical N-methylisotoic anhydride (NMIA) and its derivatives to modifiy the ribose of nucleotides within flexible regions in RNA secondary structure, and is able to react with all four nucleotides. In both cases, the modified nucleotides result in termination of reverse transcription (RT), and can be detected via high throughput sequencing. The number of RT stops at each position is

thus indicative of the relative accessibility of that nucleotide.

In addition to measure the accessibility of a single nucleotide, effort is underway to probe the actual base-pairing, either intramolecular, or intermolecular. Three methods have been proposed to use psoralens for cross-linking the duplex regions of RNA: psoralen analysis of RNA interactions and structures (PARIS)[**?**], sequencing of psoralen crosslinked, ligated, and selected hybrids (SPLASH)[**?**], ligation of interacting RNA and high-throughput sequencing (LIGR-seq)[**?**], Sequencing of the fragmented and ends-ligated reads provides information on direct base-pairing in the secondary structure, and *trans* RNA-RNA interactions. TODO limitation

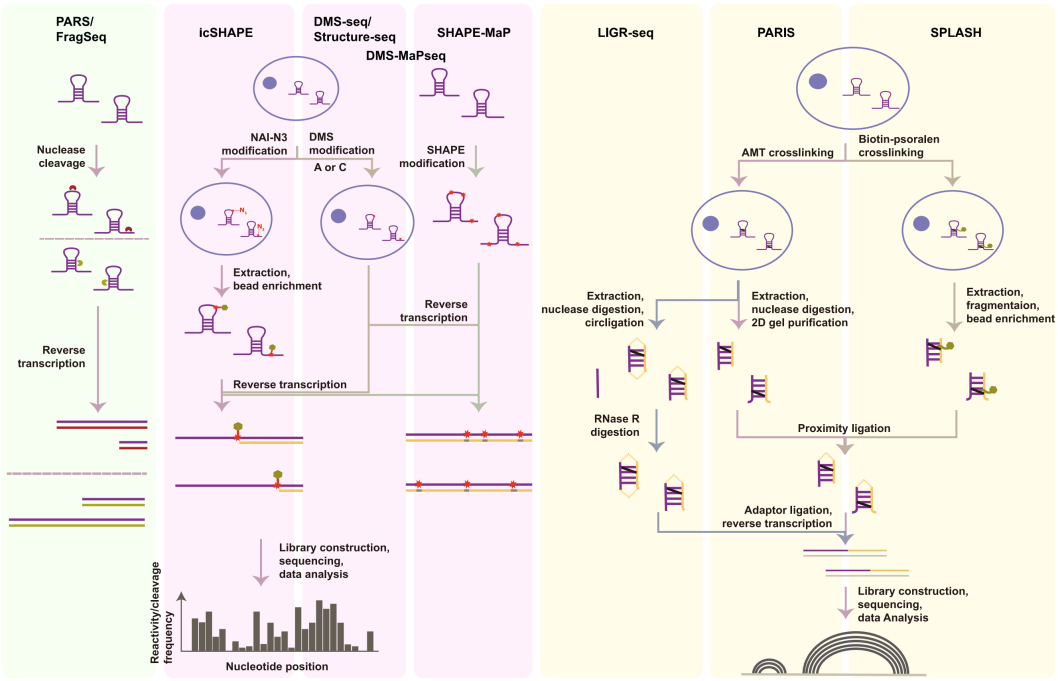

Figure 1.1: RNA structure probing methods. From Piao et. al[9].

Different probing methods have their own limitations. Enzymes are too big to permeate through cell membranes, thus can not be used *in vivo*. Certain enzymes, e.g. RNase V1, requires Mg^{2+} concentration much higher than in physiological condition to be active?, and Mg^{2+} is known to promote RNA folding. On the other hand, although Dimethyl sulfide (DMS) works *in vivo*, it only detects unstructured A/C bases. TODO cross-link methods

Furthermore?, although *in vitro* data provides invaluable insight into the genome-wide organization of RNA 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 the precise reconstruction of these conditions in vitro is almost impossible. Rouskin et al.[10] 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)

1.2 RNA secondary structure and gene regulation

Genome-wide RNA structure probing has enabled the analysis of structure and function at a global scale, which uncovers new properties and relationship that was never discovered from previous studies.

Transcription

Kertesz et. al[7] performed in vitro profiling of the budding yeast (*Saccharomyces cerevisiae*) RNA structure using PARS (TODO). Averaging PARS scores across more than 3000 mRNAs revealed a unique pattern across the transcript (Fig 1.2): the UTRs are less structured than the CDS, both start and stop codon are significantly more accessible, and the coding region exhibits a three nucleotide periodical pattern, where the first nucleotide is more accessible and the second one is less accessible.

In contrary, although RNAs in human show a similar start/stop codon accessibility and CDS periodicity, it was observed that UTRs are only slightly less structured than CDS[14]. (TODO, other paper says more structured?)

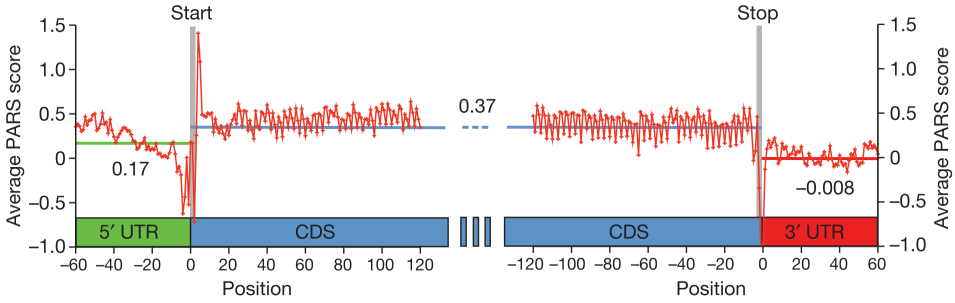

Figure 1.2: Yeast RNA structure differ in CDS and UTR. From Kertesz et. al[7].

Splicing

Wan et. al[14] analyzed human lymphoblastoid cell lines from a parent-offspring trio by PARS (TODO review), and observed less structure at AG dinucleotide in the upstream exon donor site, and more structure at the first nucleotide in the downstream acceptor site, as shown in Fig 1.3. Potential role in efficient splicesome assembly.

Polyadenylation

Ding et. al[2] used Structure-Seq (TODO review) to study the genome-wide RNA structure of *Arabidopsis thaliana* seedlings *in vivo* and discovered a pattern around alternative polyadenylation sites across 5959 mRNAs. As shown in Fig 1.4, 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. This

Translation

Kertesz et. al[7] reported a small but significant anti-correlation between PARS scores 10bp upstream of the start codon and ribosome density throughout the transcript. In addition, genes where 5'UTRs are

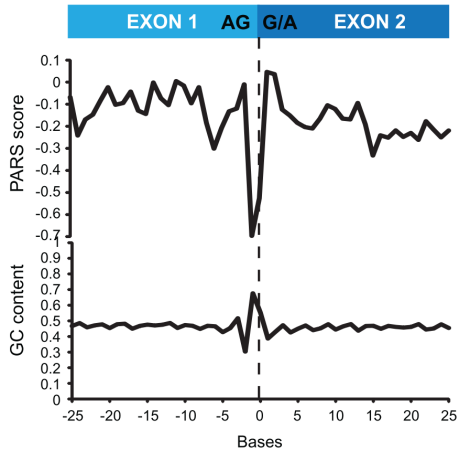

Figure 1.3: RNA structure at exon-exon junction in human lymphoblastoid cell lines. From Wan et. al[14].

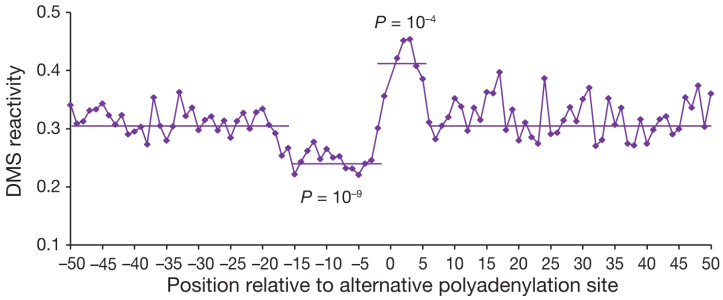

Figure 1.4: RNA structure around alternative polyadenylation sites in *Arabidopsis thaliana*. From Ding et. al[2].

less structured than the beginning of the coding region also show a tendency towards higher ribosome density.

Localization

Kertesz et. al[7] discovered increased structure in coding region for genes whose encoded proteins localize to distinct cellular domains or function in specific metabolic pathways. On the other hand, ribosomal transcripts show less structure in UTR and CDS (Fig 1.5).

Stability

Kertesz et. al[7] discribed how RNA structure in UTR regulates gene expression during heat shock in order to conserve energy. Certain class of genes (e.g. ribosomal encoding RPL1A) with less UTR structure becomes unfolded with increased temperature, which allows degradation by exosome, thus tuning down translation. On the other hand, genes with more UTR structure like chaperones and unfolded response proteins (e.g. HAC1 and PTC2) remain stable and expressed (Fig 1.6).

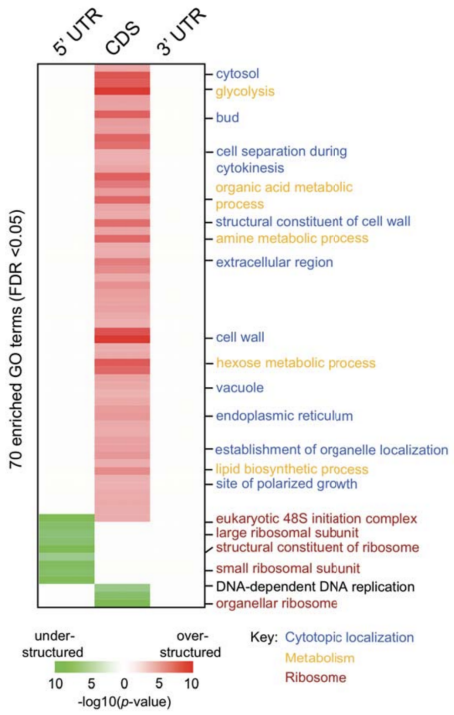

Figure 1.5: RNA structure in CDS and UTR affects localization in yeast. From Kertesz et. al[7].

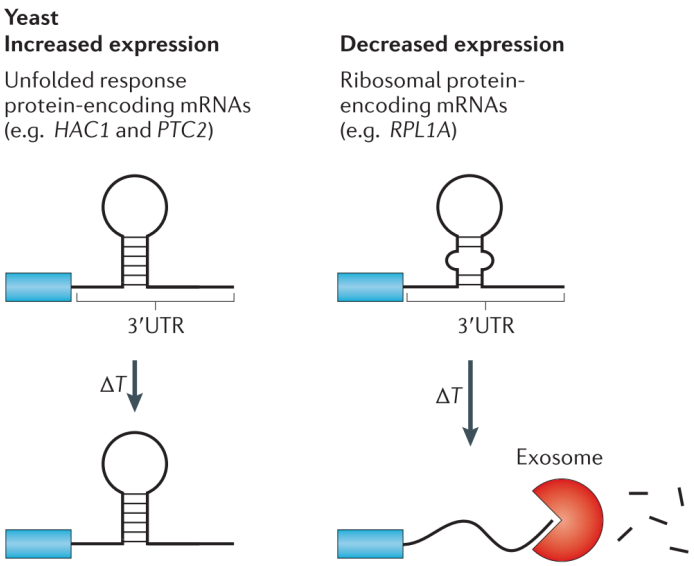

Figure 1.6: Yeast gene expression regulation via RNA structure during heat shock. From Mortimer et. al[8].

1.3 Deep neural network for sequence modelling

Predicting RNA structure is similar in its nature to many applications in computer vision (CV) and natural language processing(NLP), where the input is a variable sized sequence, and the output is a

sequence of the same size as the input. Here we review the recent breakthroughs in deep learning and discuss their application in genomics. TODO 2D MAP

Recurrent neural network and LSTM

Recurrent neural networks(RNNs), especially long short-term memory (LSTM)[4], have been successfully applied to many sequence labelling tasks Although LSTM mitigates the vanishing and exploding gradient problem in the vanilla RNNs, and has the nice property of being able to model long range dependency of any distance (at least in theory), in practise, it falls short due to the computational time, which scales linearly with the length of input for both training and inference, thus renders it difficult to scale up to long input sequence lengths.

Dilated convolutional neural network, residual function, skip connection and attention mechanism

Convolutional neural network (CNN) has been applied successfully in modelling transcription factor (TF) and RNA binding protein (RBP) binding[**?**, **?**], alternative splicing[**?**], and genome accessibility[**?**]. One limitation of the classical CNN architecture is that the receptive field size grows linearly with the number of layers, so in order to construct a large receptive field for modelling long range dependencies, the network is typically very deep, but classical CNNs are known to be difficult to optimize as the number of layers grows.

Instead, researchers has proposed improved architectures to model long range dependencies while still addressing the computational time problem that is intrinsic to all recurrent neural networks, including LSTM.

Yu et. al[15] proposed a variation of the convolutional neural network (CNN), by introducing 'holes' in convolution filters, such that the receptive filed in a single layer is n times of the original, where $n-1$ is the number of 'holes' between adjacent connections in the filter. Compared to non-dilated CNN with the same number of parameters, dilated CNN achieves exponential growth of the receptive filed, while maintaining the same resolution and coverage.

He et. al[3] proposed a neural network architecture where they reformulated each layer to, instead of learning the direct mapping $\mathcal{H}(x)$, learns the residual function $\mathcal{H}(x) - x$. They showed that very deep neural network (up to 1000 layers) can be optimized with no difficulties. Combining these residual functions with the above mentioned dilated convolution, Jaganathan et. al[6] successfully trained a 32-layer deep neural network to predict the location of splice sites from primary sequence.

DenseNet, proposed by Huang et. al[5], is another architecture to mitigate the difficulty in training deep CNN. In order to improve information flow (in both forward and backward direction), shorter path between layers is necessary. This is achieved by connecting all pairs of layers, i.e. every layer receives as input concatenation of all previous layers, and passes its output to all subsequent layers. Such dense connection not only alleviates the vanishing gradient problem, but also enables feature reuse which makes the network more parameter-efficient.

Attention mechanism (TODO ref) is another technique to model long range dependencies, without limitation by distances. Vaswani et. al[12] proposed the Transformer Network, which is solely based on attention mechanism to model global dependencies between input and output. In each block, the network applies a self-attention mechanism which directly models relationships between all positions in

the input. For each position, the attention scores for other positions are normalized, which are used as weights to combine input from all positions to construct a new representation of the current position. These networks are highly parameterizable, achieves ?? performance, and have the added benefit that mechanistic insight can be drawn from the attention weights. (TODO example) One drawback of these networks is that substantial memory is required for computing and storing the attention matrix. (TODO example, typical size)

TODO plots for NN

1.4 Related work

Most state-of-the-art RNA structure prediction algorithms, such as ViennaRNA[**?**] and Mfold[**?**], are based on the fundamental property of base-pairing. Each type of base-pairing, as well as each type of local structure, has its own associated free energy that is measured experimentally. Total free energy is the sum of all local free energy, and can be minimized efficiently using dynamic programming.

Researchers have also applied deep learning to predict RNA structure from sequence directly, as described in DMfold[**?**]. The authors trained a LSTM to classify each nucleotide in one of the seven categories, corresponding to each unique symbol in the dot-bracket notation. There are two potential limitations of this model. First, training and validation set were splited randomly, while the dataset only consists of a handful of RNA families, this means the tranining and validation set contain highly similar sequences, so the performance reported in the work does not necessarily reflect the generalization performance to unseen RNA sequences or families. Second, the output class labels are not guaranteed to be compatible throughout the sequence, in fact, the authors need to apply heuristics to post-process the neural network output into compatible labels.

In terms of prediction using RNA probing data, so far the only attempt was to use the measured accessibility to guide dynamic programming algorithm like ViennaRNA[**?**], by introducing extra penalty term when computing the free energy.

(usefulness of the model) - predict on novel transcript, mutation - low abundance and repeated region - help with other prediction tasks

1.5 Proposed thesis work

In this thesis, we propose to work on the following directions:

- Train deep learning models to predict in vivo accessibility from sequence, for multiple species and cell types
- analyze feature map, cross-tissue, cross-species? conservation?
- eval? disease? mutation?
- Improve TF/RBP binding and splicing model using RNA secondary structure model

list of in vivo dataset
list of validation data
vast number of non coding RNAs (still being discovered)
disease

Chapter 2

Yeast Model

2.1 Training Dataset

To model in vivo RNA secondary structure, we compiled training data from [10]. 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[5], 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

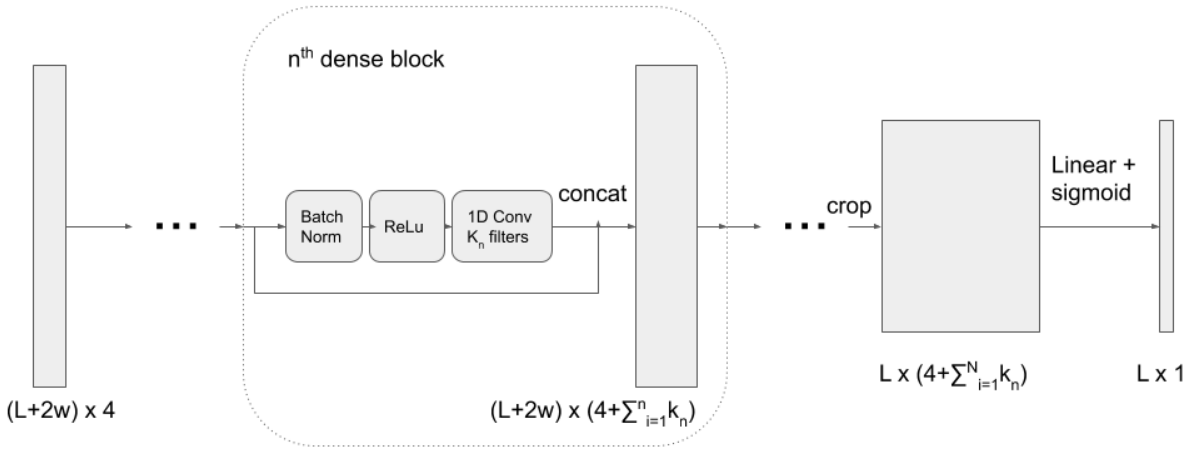

Figure 2.1: Densely connected neural network used for the yeast model

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.

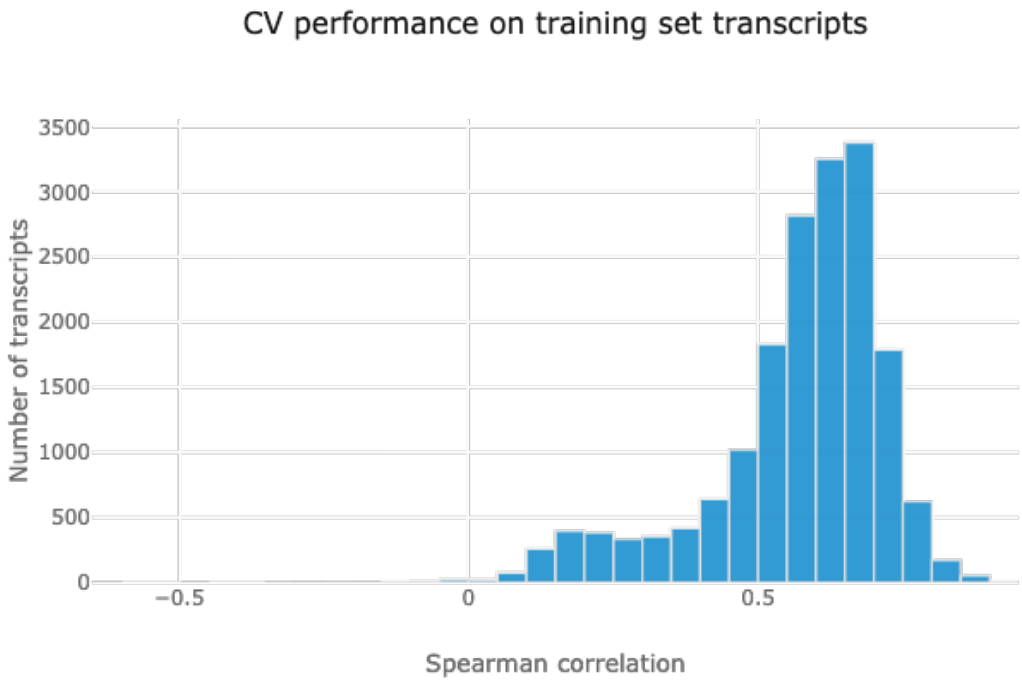

Figure 2.2: Densely connected neural network used for the yeast model

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[11], 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.

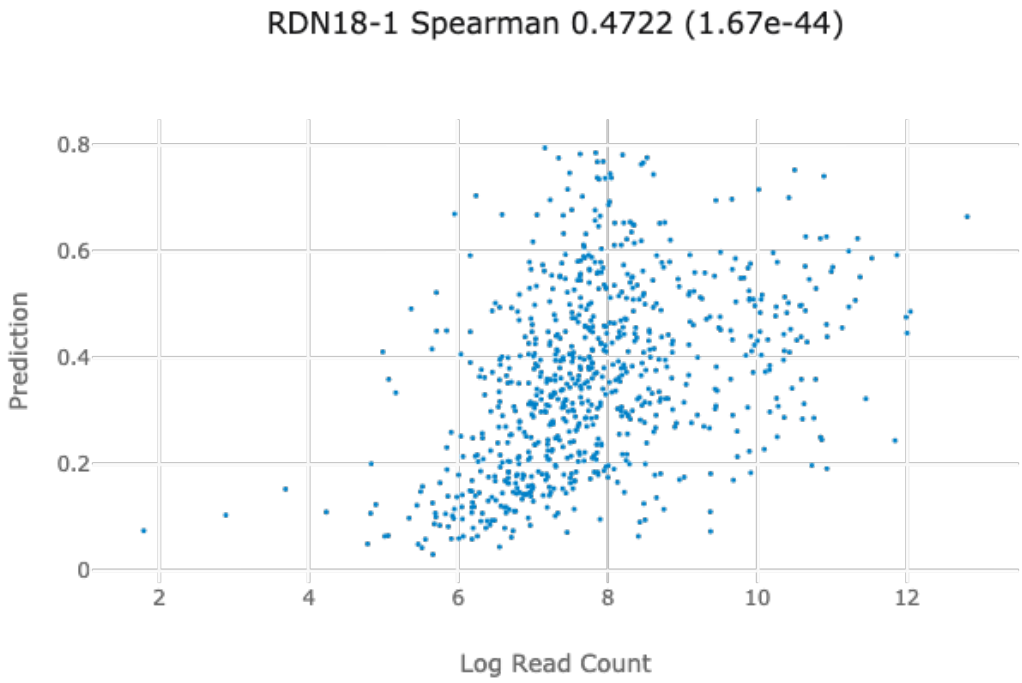

Figure 2.3: Densely connected neural network used for the yeast model

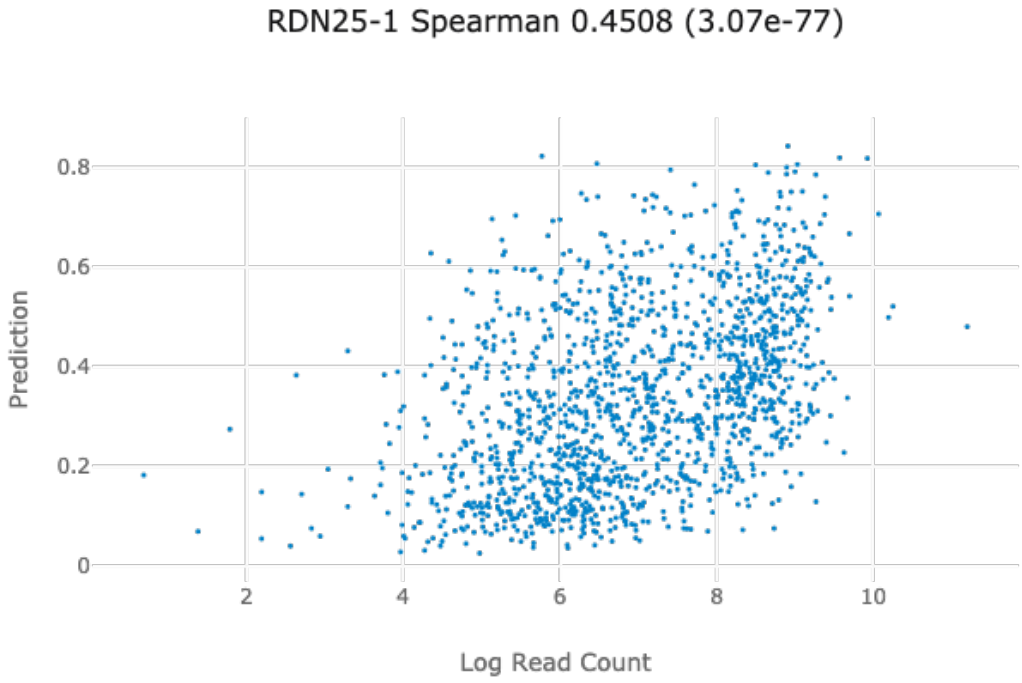

Figure 2.4: Densely connected neural network used for the yeast model

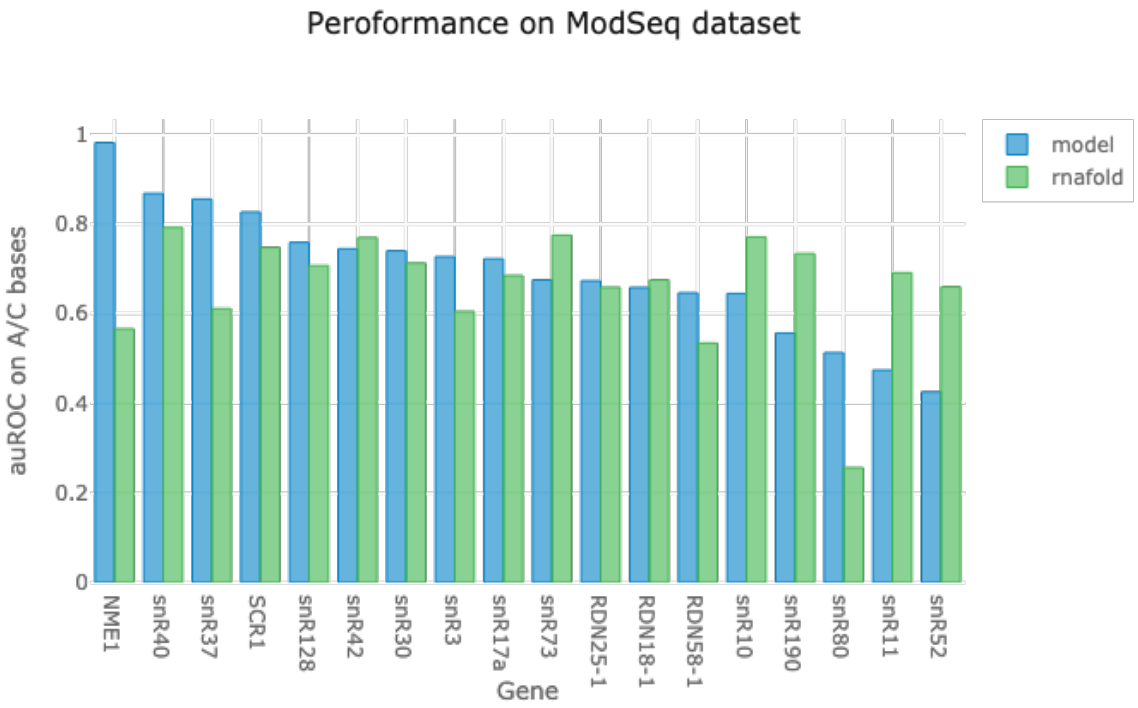

Figure 2.5: Densely connected neural network used for the yeast model

Chapter 3

Conclusion

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] 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.

[2] 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.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[4] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[5] 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.

[6] Kishore Jaganathan, Sofia Kyriazopoulou Panagiotopoulou, Jeremy F McRae, Siavash Fazel Darbandi, David Knowles, Yang I Li, Jack A Kosmicki, Juan Arbelaez, Wenwu Cui, Grace B Schwartz, et al. Predicting splicing from primary sequence with deep learning. *Cell*, 176(3):535–548, 2019.

[7] 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, 2010.

[8] 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.

[9] Meiling Piao, Lei Sun, and Qiangfeng Cliff Zhang. Rna regulations and functions decoded by transcriptome-wide rna structure probing. *Genomics, proteomics & bioinformatics*, 15(5):267–278, 2017.

[10] 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.

[11] 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.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[13] 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.

[14] Yue Wan, Kun Qu, Qiangfeng Cliff Zhang, Ryan A Flynn, Ohad Manor, Zhengqing Ouyang, Jiajing Zhang, Robert C Spitale, Michael P Snyder, Eran Segal, et al. Landscape and variation of rna secondary structure across the human transcriptome. *Nature*, 505(7485):706, 2014.

[15] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.