Predicting in vivo RNA Secondary Structure

by

Jiexin (Alice) Gao

Thesis proposal for the degree of Doctor of Philosophy
Graduate Department of Electrical Engineering
University of Toronto

Contents

Chapter 1

Introduction

Once believed to be an intermediate molecule that serves as a 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 exists as single-stranded and folds onto itself by forming base pairs via hydrogen bonds, including Watson-Crick pairs A-U, G-C and non-canonical pairs such as A-U. Nearby paired and unpaired bases further assemble into hairpins, bulges, internal loops, multi loops and pseudoknots, as shown in Fig 1.1. In addition, secondary structures within an RNA molecule can interact via non-covalent bond, to form tertiary structure.

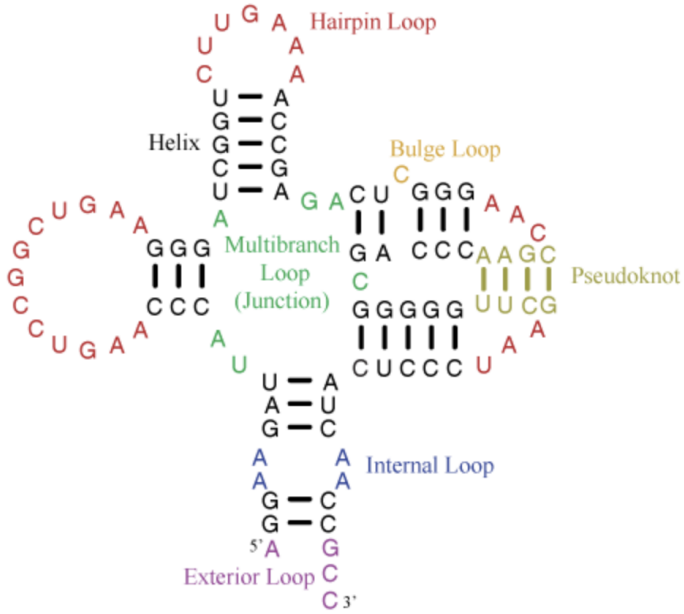

Figure 1.1: RNA secondary structure components. Courtesy of https://x3dna.org/articles/exterior-loop-in-rna-secondary-structure.

RNA secondary and tertiary structures play key roles in regulating both coding and nocoding genes, affecting all steps including transcription, splicing, polyadenylation, translation, localization and stability[1, 2, 3]. Over the past few years, combination of RNA structure probing and high throughput sequencing has enabled the measurement 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 RNA 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)[4] uses RNase S1 to cleave single-stranded regions, and RNase V1 to cleave double-stranded regions. Similarly, fragmentation sequencing (FragSeq)[5] 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.

Dimethyl sulfide (DMS) reacts with unpaired adenine (A) and cytosine (C), and is being used in various protocols to modify RNA[6, 7]. 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 measuring 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)[8], sequencing of psoralen crosslinked, ligated, and selected hybrids (SPLASH)[9], and ligation of interacting RNA and high-throughput sequencing (LIGR-seq)[10]. In all three methods, sequencing of the fragmented and ends-ligated reads provides information on direct base-pairing in secondary structure and *trans* RNA-RNA interactions.

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 requires non-physiological condition to be acitve, for example, RNase V1 needs much higher Mg^{2+}, while Mg^{2+} is known to promote RNA folding. Chemicals such as Dimethyl sulfide (DMS) works *in vivo*, but it only detects unstructured A/C bases. Cross-link methods were proposed very recently, and more work needs to be done to fully understand the bias and error involved in experimental protocol and read processing. For example, it has been hypothesized that perfectly ordinary duplexes such as those involving G-C base pairs might not be detectable due to the bias of cross-linking pyrimidines[12].

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 and ligands, concentration of salt, temperature and other factors defines a complex environment, which in turn affects RNA structure, and the precise reconstruction of these conditions *in vitro* is almost impossible. Rouskin et al.[6] 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 highlights the effect of environment on RNA structure.

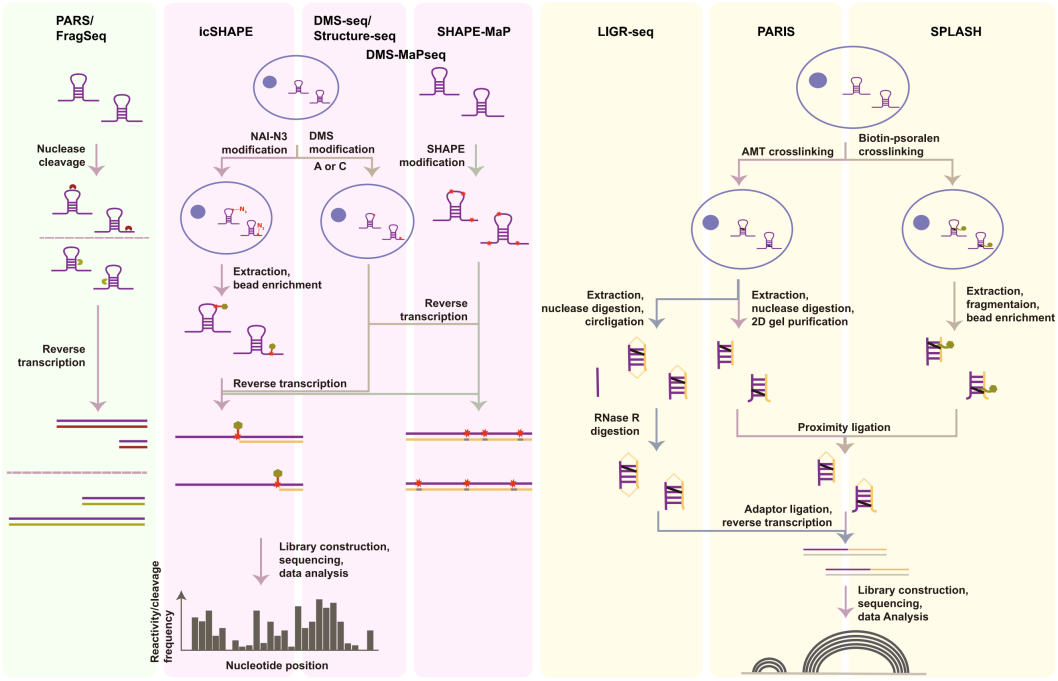

Figure 1.2: RNA structure probing methods. From Piao et. al[11].

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[13] performed in vitro profiling of the budding yeast (*Saccharomyces cerevisiae*) RNA structure using PARS. Averaging PARS scores across more than 3000 mRNAs revealed a unique pattern across the transcript (Fig 1.3): 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[4].

Splicing

Wan et. al[4] analyzed human lymphoblastoid cell lines from a parent-offspring trio by PARS, 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 1.4. This suggests a potential regulatory role of RNA structure in efficient splicesome assembly.

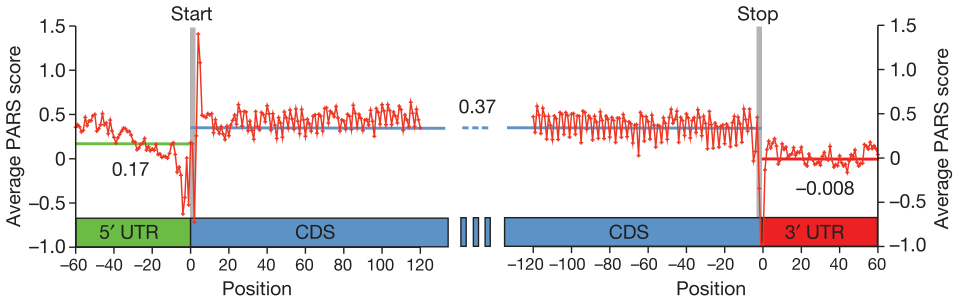

Figure 1.3: Yeast RNA structure differ in CDS and UTR. From Kertesz et. al[13].

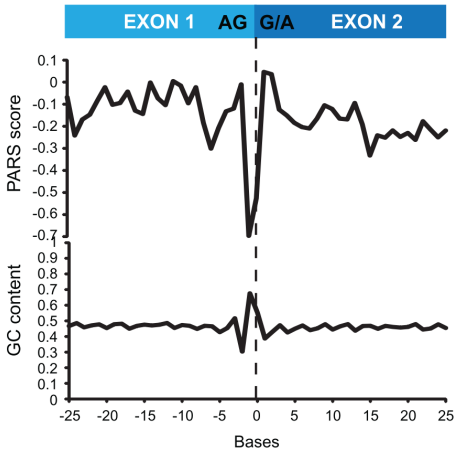

Figure 1.4: RNA structure at exon-exon junction in human lymphoblastoid cell lines. From Wan et. al[4].

Polyadenylation

Ding et. al[14] used Structure-Seq 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 1.5, 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.

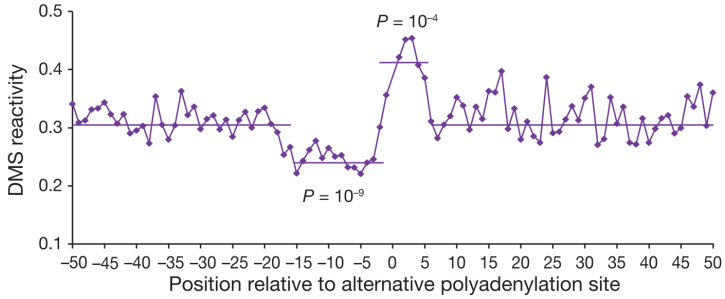

Figure 1.5: RNA structure around alternative polyadenylation sites in *Arabidopsis thaliana*. From Ding et. al[14].

Translation

Kertesz et. al[13] 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 whose 5'UTRs are less structured than the beginning of the coding region also show a tendency towards higher ribosome density.

Localization

Kertesz et. al[13] 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 1.6).

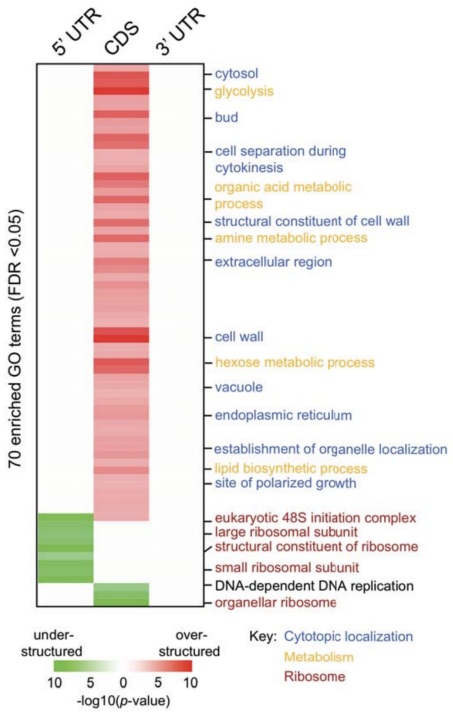

Figure 1.6: RNA structure in CDS and UTR affects localization in yeast. From Kertesz et. al[13].

Stability

Kertesz et. al[13] 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 1.7).

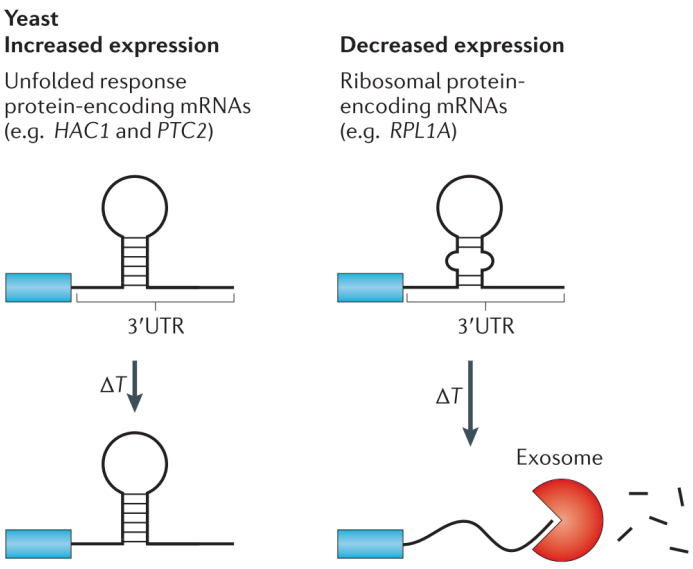

Figure 1.7: Yeast gene expression regulation via RNA structure during heat shock. From Mortimer et. al[2].

1.3 Deep neural network for sequence modelling

The task of predicting RNA secondary structure is similar to many applications in computer vision (CV) and natural language processing (NLP). The input is a variable sized sequence, and the output is a vector of the same length as the input (one dimensional output, e.g. accessibility of each base), or an array whose size is a deterministic function of the input length (two dimensional output, e.g. probability of all base pairs). Here we review the recent breakthroughs in deep learning related to sequence modeling and discuss their application in genomics.

Recurrent neural network and LSTM

Recurrent neural networks (RNNs), especially long short-term memory (LSTM)[15], have been successfully applied to many sequence labelling tasks. Although LSTM mitigates the vanishing and exploding gradient problem of the vanilla RNNs, and is capable of modeling long range dependency 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 (CNNs) have been applied successfully in modelling transcription factor (TF) and RNA binding protein (RBP) binding[16, 17], alternative splicing[18], alternative polyadenylation[19], and genome accessibility[20]. One limitation of the classical CNN architecture is that the receptive field size grows linearly with the number of layers. In order to construct a large receptive field for modelling long range dependencies, the network is typically very deep, which does not work in practise, since classical CNNs are known to be difficult to optimize as the number of layers grows.

Instead, researchers have 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[21] proposed a variation of the convolutional neural network (CNN), by introducing dilation a.k.a. '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[22] 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[23] successfully trained a 32-layer deep neural network to predict the location of splice sites from primary sequence.

DenseNet, proposed by Huang et. al[24], 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 the 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 is another technique to model long range dependencies, without limitation by distances. Vaswani et. al[25] proposed the Transformer Network, which is solely based on attention mechanism to model global dependencies. 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, achieve state-of-the-art performance, and have the added benefit that mechanistic insight can be drawn from the attention weights. One drawback of these networks is that substantial memory is required for computing and storing the attention matrix.

1.4 Related work

Most state-of-the-art RNA structure prediction algorithms, such as ViennaRNA[26] and Mfold[27], 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, as shown in Fig 1.8, and can be minimized efficiently using dynamic programming. Although people have worked out a large set of local structure types and their associated formulation for free energy calculation, there is no guarantee that it's either accurate or complete. There has been effort to fine tune the free energy parameters by training on experimentally solved structures[28], but it's still limited by the known set of local structure types.

Researchers have also applied deep learning to predict RNA structure from sequence directly, as described in DMfold[29]. The authors trained a LSTM to classify each nucleotide in one of the seven

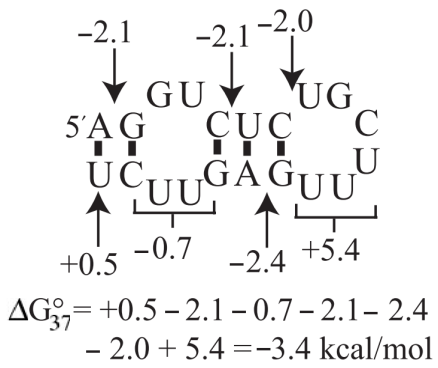

Figure 1.8: Total free energy is the sum of local free energy. From Andronescu et. al[28].

categories, corresponding to each unique symbol in the dot-bracket notation[1]. There are two potential limitations of this model. First, training and validation set were splited randomly, even though the dataset only consists of a handful of RNA families, this means the training 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[26], by introducing extra penalty terms when optimizing the total free energy.

A computational model for in vivo RNA accessibility is beneficial in the following aspects:

- Dynamic programming models have reasonable performance for *in vitro* RNA folding but perform poorly for *in vivo* folding.

- Incorporating *in vivo* probing data into dynamic programming models is not trivial, since it's unclear whether there exists an optimal way to combine the known free energy and the measured accessibility. Furthermore, high throughput probing data in general contains significant amount of noise, which might result in inaccurate result if not being handled properly.

- RNA folding *in vivo* is affected by many factors, most of which are not fully understood yet. Dynamic programming models might be limited in their ability to generalize to such conditions due to the fixed set of known local structures.

- Certain transcripts are difficult to probe due to either low abundance or repeated regions, a model can provide insight on these transcript using information learned from other transcripts in the same cell type.

- Structure change due to mutation can be predicted without the need to carry out experiments, which makes it possible to design synthetic novel RNA sequence with certain structural properties.

[1]For a comprehensive explanation on dot-bracket notation, as well as other ways to represent RNA structure, see https://www.tbi.univie.ac.at/RNA/tutorial/

- Since RNA structure influences almost every step in gene regulation, the features learned by such model will also be useful in modelling other cellular processes.

1.5 Proposed thesis work

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

- Develop neural network architecture that incorporates prior knowledge on the biological nature of RNA folding. We will explore various architectures and evaluate their potential in modelling the fundamental building block of RNA secondary structure: base pairing. **In Chapter 2, we present preliminary result on a simplified toy example of base pairing.**

- Model architecture developed above will be utilized in deep learning models to predict *in vivo* RNA secondary structure from sequence, for multiple species and cell types. *In vitro* data can also be used to help with learning useful low level features. **In Chapter 3, we present preliminary result on modelling yeast *in vivo* accessibility.**

- Fine tuning model on mutation dataset. Wan et. al[4] identified 'riboSNitches' (mutations that affect structure) by performing PARS on lymphoblastoid cells of a family trio. We can fine tune the model to improve its sensitivity to single nucleotide mutations.

- Improving TF/RBP binding and splicing models using RNA secondary structure model. Binding affinity or splicing pattern in specific tissue or cell type might be affected by the local RNA structure as a result of distal sequence factor which is not being captured by most models. Using RNA structure prediction as input features, or model structure and binding/splicing jointly can potentially improve tissue or cell type specific models.

Chapter 2

Toy Example

A fundamental component of RNA secondary structure is base pairing. In this chapter, we explore neural network architectures suitable for modelling this fundamental building block of RNA secondary structure.

2.1 Training Dataset

We construct a toy dataset for the following task:

Given a RNA sequence of length $2M+1$: $x_1, x_2, \cdots, x_{2M+1}$, where M is a positive integer (so the sequence length is an odd number), for each base x_i, predict the probability that it is paired with the middle base x_M in the equilibrium ensemble.

Such task is a simplified version of the more difficult one, where one needs to predict the pairing probability upper triangular matrix from the sequence.

We used RNAfold from ViennaRNA package[26] to generate the dataset. 100000 sequences of length 51 were generated, where each base was sampled independently and uniformly from $\{A, C, G, U\}$. We ran RNAfold on each sequences, and extracted the M^{th} row from the pairing probability matrix, corresponding to the middle position[1].

2.2 Model architecture

As shown in Fig 2.1, in order to help the neural network capture base pair information, we run convolution layers in parallel, one set on the original sequence, and the other set on the reversed sequence. After each layer of convolution, the activation from the reverse sequence is cropped to retain only the unit corresponding to the filter centered at the middle position, then a dot product is computed between this and the forward activation at all positions. Dot product is one way to measure similarity, and was also used in one variation of Attention-based Neural Machine Translation model[30]. Finally, the dot product vectors from all layers are concatenated and passed through a fully connected layer (along the concatenated feature dimension) with sigmoid activation to predict the base pair probability.

[1] To be precise, since RNAfold outputs the upper triangle matrix, we extracted the M^{th} row and M^{th} column in the upper triangle, then combined into an array of length $2M+1$

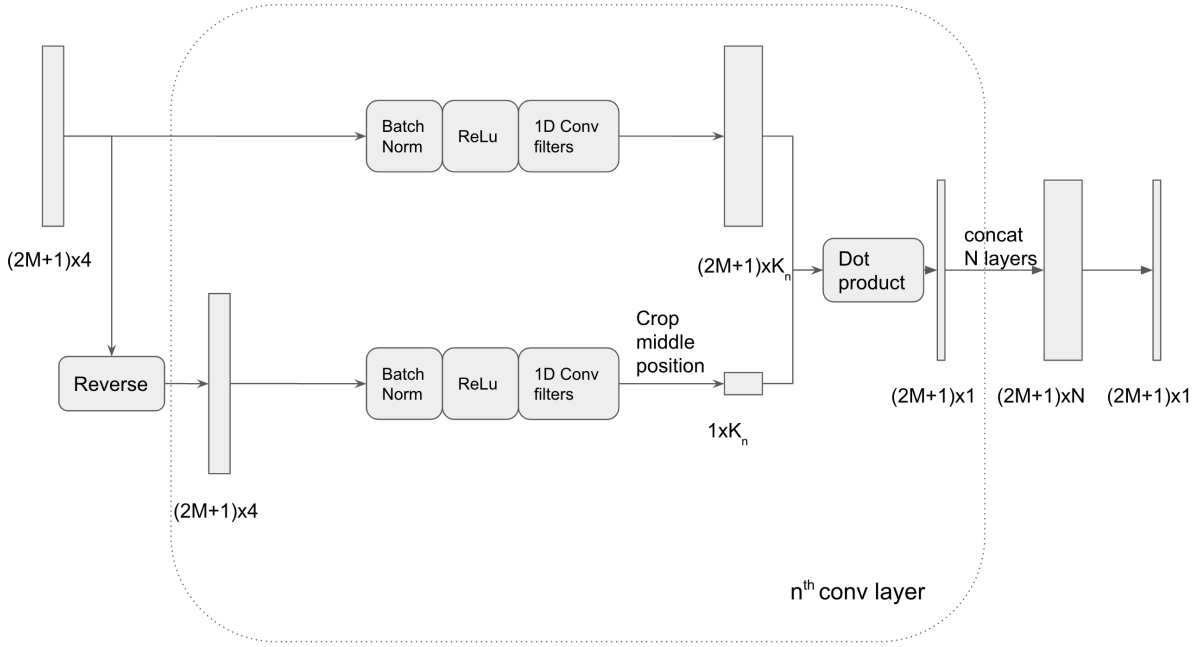

Figure 2.1: Neural network for base pair modelling.

2.3 Training

We used 80000 sequences as training data and 20000 for validation. For this toy dataset, we used 5 layers of convolution, with 64 filters in all layers, and filter widths $7, 3, 3, 5, 9$.

2.4 Result

2.4.1 Validation data performance

We first evaluate our model's performance on validation set. We focus on 75406 out of 20000 validation sequences, where there is at least one positions with pair probability larger than 0.5. We calculated top-k accuracy for each sequence, where a sequence is considered being correctly predicted if the position with largest target value is in the top k positions with highest predictions. The result is listed in Table 2.1.

k	Accuracy
1	0.66
2	0.85
3	0.92
4	0.95
5	0.97

Table 2.1: Validation set top-K accuracy.

2.4.2 Gradient ascent for output maximization

One benefit of having a differentiable model like neural network is that we can answer interesting questions like: given my current input sequence, what (small) changes can I make to maximize the pairing probability of a specific base?

We illustrate this using the following example sequence:

`GAAUGGGUUAAAAGGGGGGCGCAUUGGUACCUGCUAUUAGGGAUCAAUCGG`

Our model predict the 44^{th} (0-based) base C is the most likely one to be paired up with the 25^{th} center base G with a probability of 0.8383031. Right now, the 10^{th} base A has only a probability of $3.9591327e-05$ to be paired with the center base, which is not surprising since A and G don't form pair in typical cases.

In order to find small, local changes to be made on this sequence, we computed the gradient of the 10^{th} output node (a scalar) with respect to all the input nodes (51×4 array). We ran 200 gradient ascent iterations with step size 0.01, in each step, after adding the gradient (times step size) onto the input, we re-normalize the input so the feature dimension sum up to 1.

Fig 2.2, 2.3, 2.4 and 2.5 shows the gradient ascent progression. It can be observed that the 10^{th} base has been gradually mutated from A to C, which is expected since C pairs with G. Interestingly, a couple of other bases have also been mutated significantly, potentially to help with pairing the 10^{th} and 25^{th} bases.

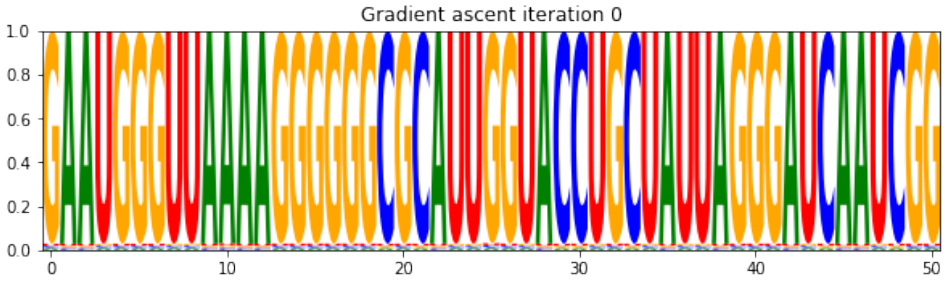

Figure 2.2: Gradient ascent on input.

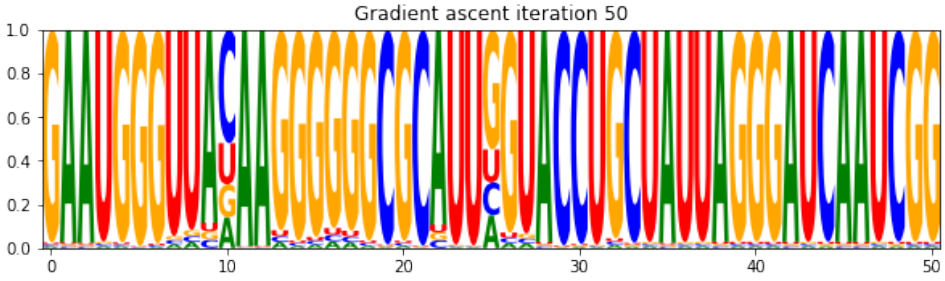

Figure 2.3: Gradient ascent on input.

As a further validation, we used RNAfold to compute the minimum free energy structure for the original and mutated sequence. For the mutated sequence, we used the bases with maximum value at each position after gradient ascent:

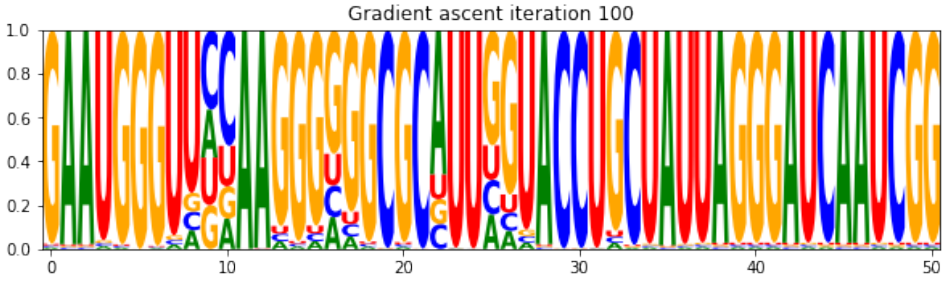

Figure 2.4: Gradient ascent on input.

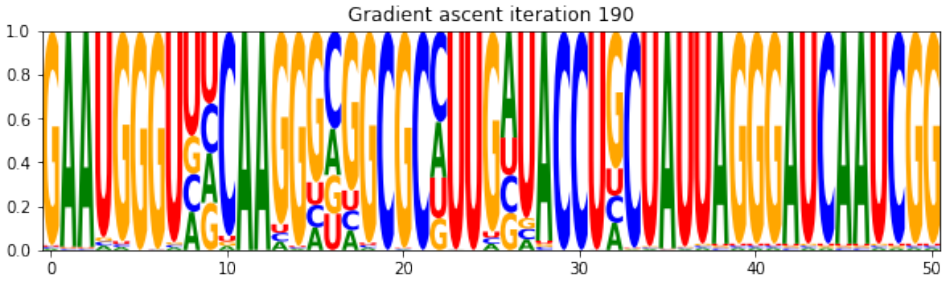

Figure 2.5: Gradient ascent on input.

```
position          0         10             25                      50
                  |         |              |                       |
original          GAAUGGGUUAAAAGGGGGGCGCAUUGGUACCUGCUAUUAGGGAUCAAUCG
                  |||||||||..|||||.|||||.|||.|||||||||||||||||||||||
mutated           GAAUGGGUUUCAAGGGCGGCGCCUUGAUACCUGCUAUUAGGGAUCAAUCG
```

(included position and original sequence for comparison)

RNAfold predicted structure for the original sequence is in Fig 2.6(a), and sequence for the mutated sequence is in Fig 2.6(b), both are consistent with our model prediction.

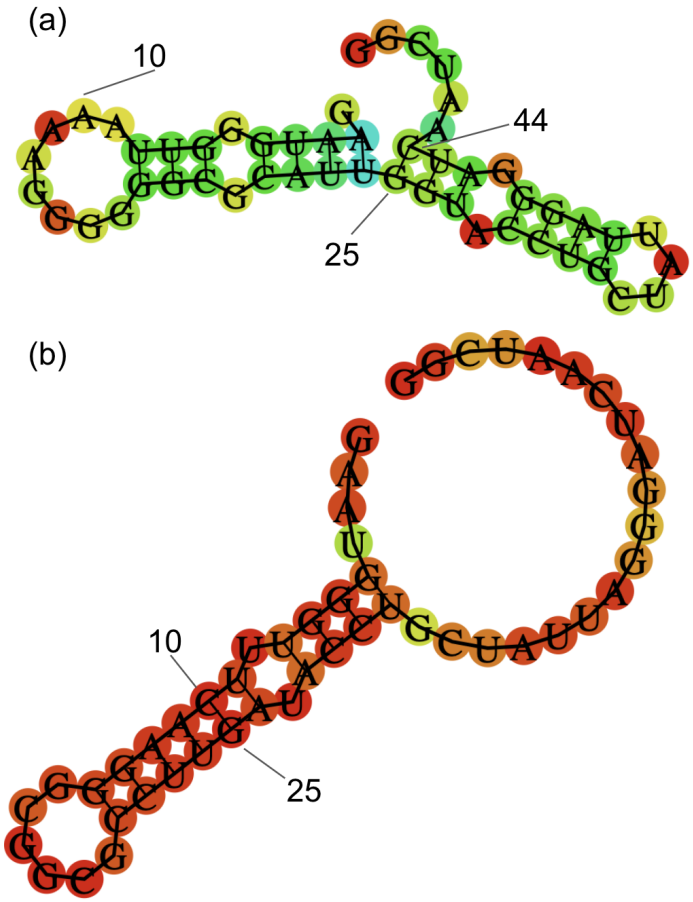

Figure 2.6: Structure predicted by RNAfold.

Chapter 3

In vivo accessibility model

In this chapter, we present preliminary result on modeling yeast in vivo RNA secondary structure. We're using generic neural network architecture for now, which will be replaced by the one developed in Chapter 2 once fully developed.

3.1 Training Dataset

To model in vivo RNA secondary structure, we compiled training data from [6]. In this study, yeast strain was treated with dimethyl sulphate (DMS), which reacts with unpaired adenine (A) and cytosine (C) bases. The pool of modified RNAs were fragmented and sequenced. Since DMS modification blocks reverse transcription (RT), number of reads with RT stop 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 at adenine or cytosine. 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 and cytosine 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 normalized values between 0 and 1.

3.2 Model architecture

We construct a deep neural network to predict reactivity at single base resolution from RNA sequence context. We use an architecture similar to DenseNet[24], in which we've removed the pooling layers, to maintain the spatial resolution throughout the depth of the neural network.

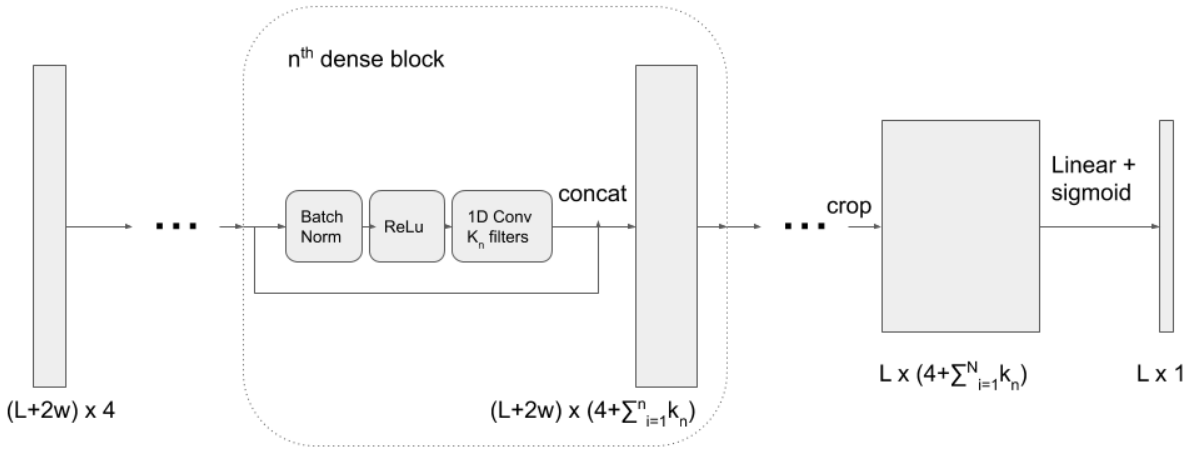

Figure 3.1: Densely connected neural network used for the yeast model

As shown in Fig 3.1, to make inference on a stretch of RNA sequence of length L, we need to pad the sequence with w bases on each side, where w is half of the receptive field size, as determined by all convolutional layers. 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 of four components:

1. Batch Normalization
2. ReLu nonlinear activation
3. 1D dilated 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 3.1: Dense block parameters

We use 5 dense blocks in this work. The parameter of each layer is as shown in Table3.1. 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.

3.3 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 3.2: Chromosomes used for each fold

We use 5-fold cross validation, split by chromosomes, as shown in Table 3.2.

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 1e−3 and momentum 0.9. To prevent the models from overfitting, L1 and L2 regularizers with weight 1e−6 was added to the loss, and training was stopped when 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.

3.4 Result

3.4.1 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 3.2 shows the distribution of Spearman correlation across all transcripts.

3.4.2 Ribosomal RNA

Next we used our model to predict the reactivity for all A/C bases in yeast $18S$ and $25S$ ribosomal RNAs, both were never seen by the model (neither in the training nor validation set).

Raw read count data for $18S$ and $25S$ was downloaded from GSE45803, and was processed similarly to the training dataset, since the experimental protocol was identical.

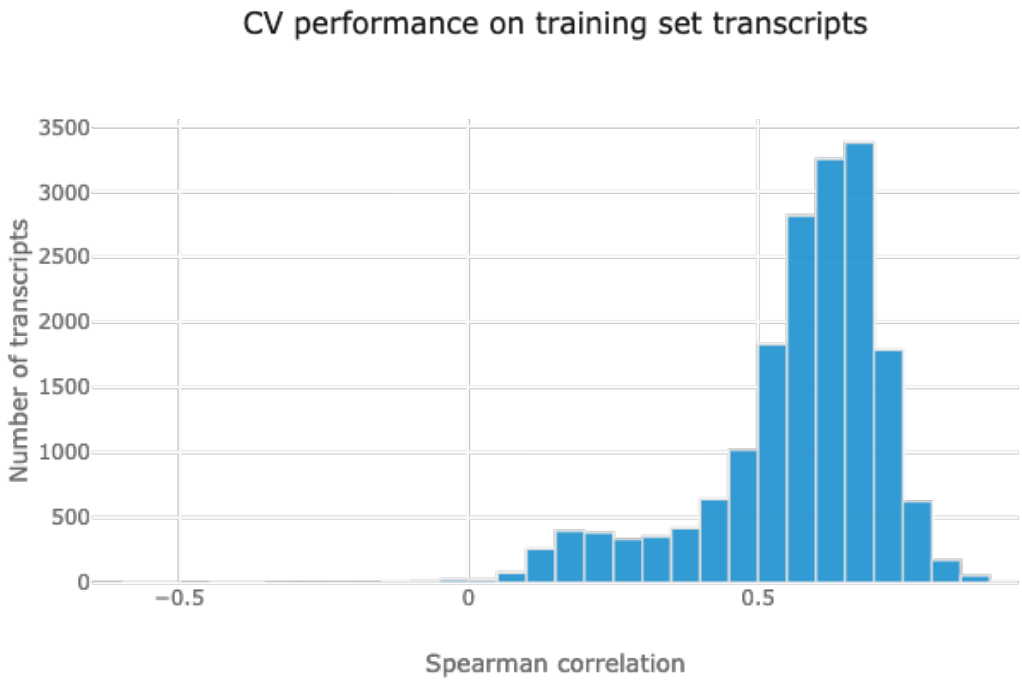

Figure 3.2: Densely connected neural network used for the yeast model

Correlation between prediction and the normalized read count is shown in Fig 3.3 and Fig 3.4, 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 $25S$, respectively.

3.4.3 Noncoding RNAs

To evaluate whether the model generalizes to noncoding transcripts and different experimental protocols, we processed yeast data from the ModSeq paper[7], where yeast was treated with DMS or no-DMS (as control), from which 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 Fig 3.5.

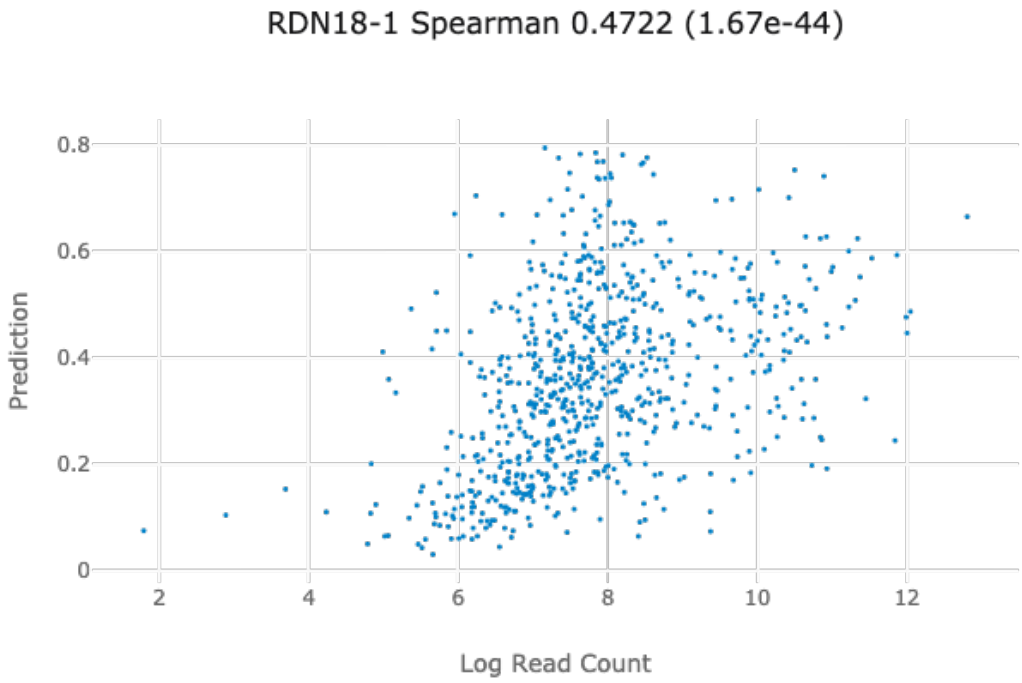

Figure 3.3: Performance on yeast tibosome RNA 18S.

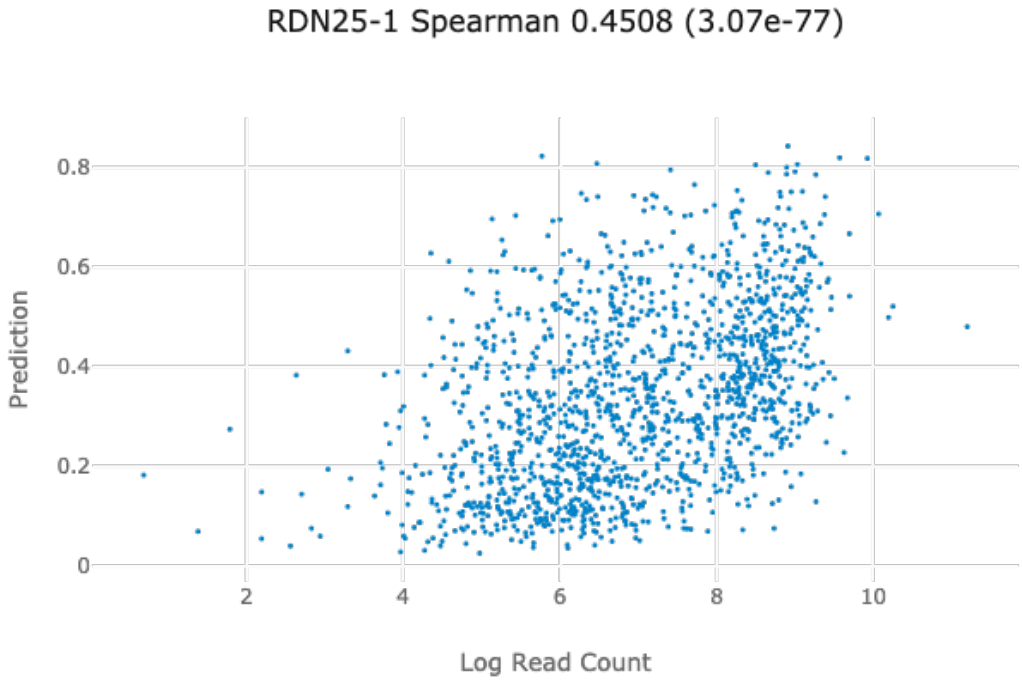

Figure 3.4: Performance on yeast tibosome RNA 25S.

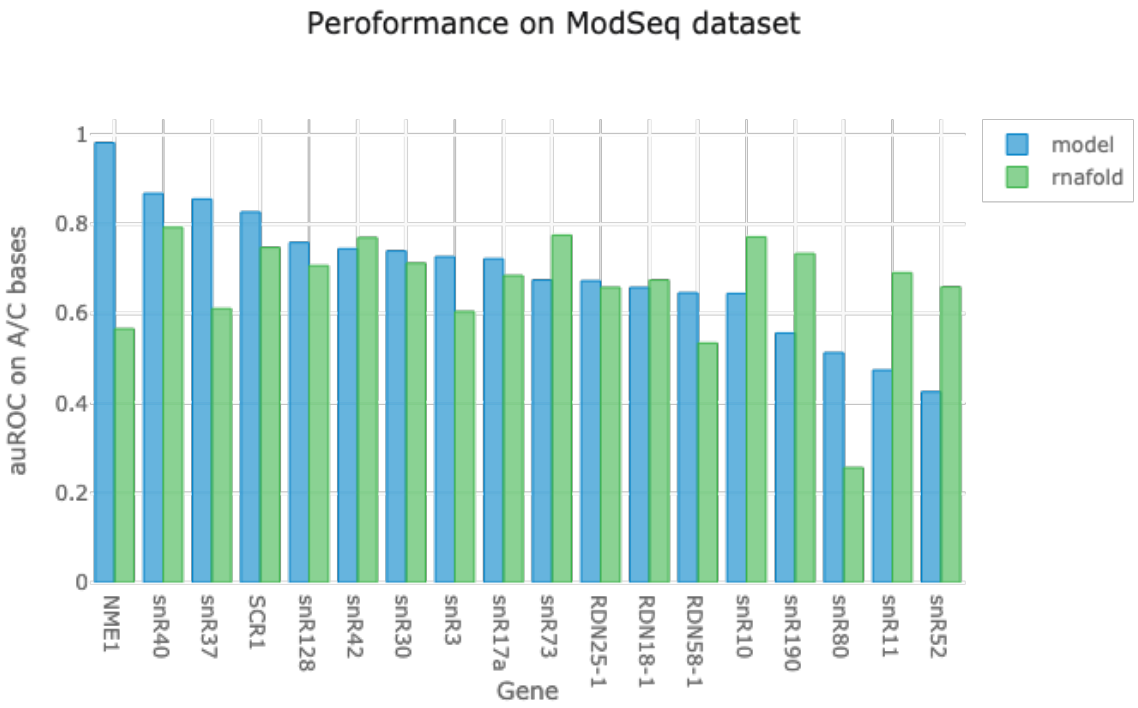

Figure 3.5: Performance on ModSeq dataset.

Bibliography

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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, 2010.

[6] 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.

[7] 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.

[8] Zhipeng Lu, Jing Gong, and Qiangfeng Cliff Zhang. Paris: Psoralen analysis of rna interactions and structures with high throughput and resolution. In *RNA Detection*, pages 59–84. Springer, 2018.

[9] Jong Ghut Ashley Aw, Yang Shen, Andreas Wilm, Miao Sun, Xin Ni Lim, Kum-Loong Boon, Sidika Tapsin, Yun-Shen Chan, Cheng-Peow Tan, Adelene YL Sim, et al. In vivo mapping of eukaryotic rna interactomes reveals principles of higher-order organization and regulation. *Molecular cell*, 62(4):603–617, 2016.

[10] Eesha Sharma, Tim Sterne-Weiler, Dave O'Hanlon, and Benjamin J Blencowe. Global mapping of human rna-rna interactions. *Molecular cell*, 62(4):618–626, 2016.

[11] 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.

[12] Stefan R Stefanov and Irmtraud M Meyer. Deciphering the universe of rna structures and trans rna–rna interactions of transcriptomes in vivo: From experimental protocols to computational analyses. In *Systems Biology*, pages 173–216. Springer, 2018.

[13] 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.

[14] 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.

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

[16] Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature biotechnology*, 33(8):831, 2015.

[17] Shreshth Gandhi, Leo J Lee, Andrew Delong, David Duvenaud, and Brendan Frey. cdeepbind: A context sensitive deep learning model of rna-protein binding. *bioRxiv*, page 345140, 2018.

[18] Hannes Bretschneider, Shreshth Gandhi, Amit G Deshwar, Khalid Zuberi, and Brendan J Frey. Cossmo: predicting competitive alternative splice site selection using deep learning. *Bioinformatics*, 34(13):i429–i437, 2018.

[19] Michael KK Leung, Andrew Delong, and Brendan J Frey. Inference of the human polyadenylation code. *Bioinformatics*, 34(17):2889–2898, 2018.

[20] David R Kelley, Jasper Snoek, and John L Rinn. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome research*, 26(7):990–999, 2016.

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

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

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

[28] 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.

[29] 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.

[30] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.