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

1.1 RNA secondary structure

1.2 High throughput probing of RNS secondary structure

1.3 Deep neural network

Chapter 2

Yeast Model

To model in vivo RNA secondary structure, we compiled training data from [1]. 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.

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.

We used the poly-A selected yeast data to compile training dataset consists of mRNAs.

5-fold CV, chromosomes

soft label cross entropy

missing value, loss/gradient masking

TODO RT stop / total coverage

TODO 4 reps

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] 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.

[2] Robert C Spitale, Ryan A Flynn, Qiangfeng Cliff Zhang, Pete Crisalli, Byron Lee, Jong-Wha Jung, Hannes Y Kuchelmeister, Pedro J Batista, Eduardo A Torre, Eric T Kool, et al. Structural imprints in vivo decode rna regulatory mechanisms. *Nature*, 519(7544):486, 2015.

[3] Lei Sun, Furqan M Fazal, Pan Li, James P Broughton, Byron Lee, Lei Tang, Wenze Huang, Eric T Kool, Howard Y Chang, and Qiangfeng Cliff Zhang. Rna structure maps across mammalian cellular compartments. *Nature structural & molecular biology*, 26(4):322, 2019.

[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.