RNA SHAPE Prediction

James Swomley [1] Leon Kloker [1]

[1]Stanford University, Institute for Computational and Mathematical Engineering

Overview

The ability to easily compute the local reactivity of each nucleotide in an RNA sequence is an important intermediate step allowing researchers to then determine the secondary and tertiary structure of a given RNA molecule. In this project, we built an ML model that accurately computes these reactivity values using only the RNA sequence as input. Our best model consists of a BERT-style foundation model with an appended GRU prediction head.

Background

The secondary and tertiary structure of an RNA molecule are crucial to its transcription of proteins, since they determine the reactivity of its individual nucleotides.

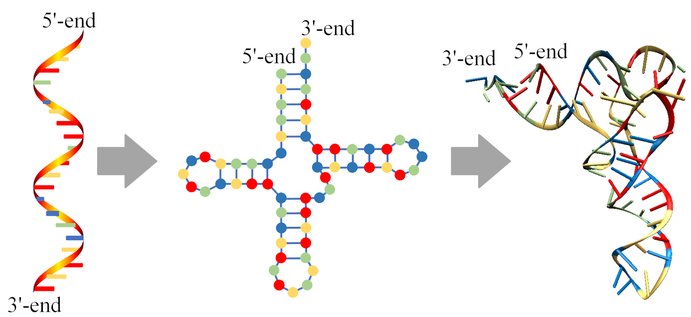

Figure 1. An RNA sequence and its secondary and tertiary structure

Directly measuring the structure of individual RNA molecules remains difficult even with modern methods such as cryo-EM. To bridge the gap between RNA sequence and structure, we introduce selective 2'-hydroxyl acylation analyzed by primer extension, or **SHAPE** data [1], which consists of a scalar value for each nucleotide that measures its reactivity. SHAPE data is expected to completely encode tertiary structure.

⇒ A model that computes SHAPE data from RNA sequence could very rapidly create a massive dataset enabling the development of new model that predicts tertiary structure.

Methods

Three main types of models were deployed.

1. Basic sequence model: A sequence model (GRU/LSTM/Transformer/CNN) is trained from scratch directly on the RNA sequence to predict SHAPE.
2. Sequence model trained on foundation model embeddings: RNA sequences are passed through RNA-FM [2] to generate embeddings, then a sequence model (the prediction head) is trained on those embeddings. RNA-FM is a BERT-style self-supervised foundation model trained on a massive corpus of unannotated RNA sequences.
3. Fine-tuned foundation model: Some of RNA-FM's transformer blocks are left unfrozen to train along with an appended prediction head.

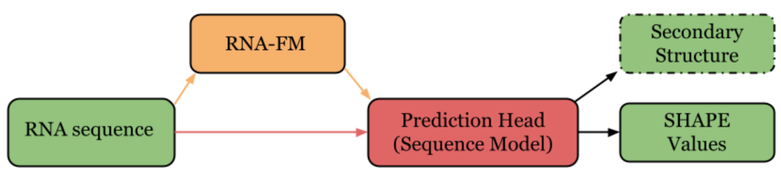

Figure 2. Model Outline

Experimental data about the RNA's secondary structure can also improve our model without changing its objective. By adding another prediction head that predicts whether each nucleotide is bound or unbound and incorporating this in our loss function, we can force the model to learn structural information that is closely related to SHAPE profile, improving SHAPE prediction accuracy.

Evaluation Metrics

The loss function used in all experiments is L1 loss. Moreover, the performance of the model is also evaluated by the median of the Pearson correlation coefficients:

$$r_{seq} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(y_i^* - \bar{y}^*)}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(y_i^* - \bar{y}^*)^2}}$$

where y_i and y_i^* are the experimental and predicted SHAPE values at nucleotide i of the sampled sequence, respectively. $\bar{y}$ and $\bar{y}^*$ denote their mean values.

Experiments

All 4 sequence models were trained in setting (1) and (2) with similar sizes of ∼1.5M parameters for comparability.

method	metric	CNN	Transformer	LSTM	GRU
(1)	MAE	.204	.391	.295	**.190**
	Pearson	.594	.165	.278	**.637**
(2)	MAE	.177	.195	.169	**.162**
	Pearson	.707	.623	.735	**.751**

Table 1. Model Performance Metrics

The results show that utilizing the foundation model as in (2) improves over the baseline method (1) in all cases. We plan on evaluating method (3) for the GRU.

References

[1] Matthew J Smola and et al. Selective 2′-hydroxyl acylation analyzed by primer extension and mutational profiling (shape-map) for direct, versatile and accurate rna structure analysis. *Nature protocols*, 10(11):1643–1669, 2015.

[2] Jiayang Chen and et al. Interpretable rna foundation model from unannotated data for highly accurate rna structure and function predictions. *bioRxiv*, pages 2022–08, 2022.