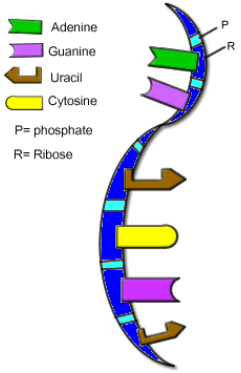
Predicting mRNA Degradation Rates with Feature Extraction

Alex Beeston

https://github.com/alexbeeston/kaggleChallenge
https://www.kaggle.com/c/stanford-covid-vaccine/data

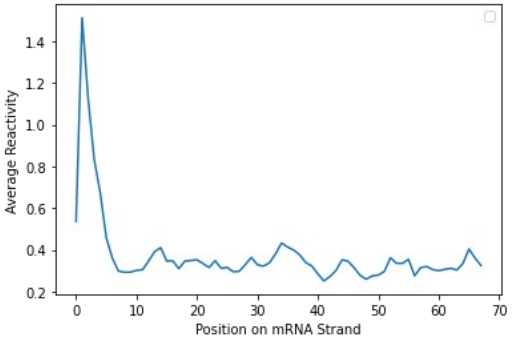
Background Information



- mRNA is a promising component to COVID-19 vaccines
- mRNA is a comprised of a sequence of bases
- Certain mRNA strands spontaneously degrade without extensive refrigeration
- Each base has a probability of reactivity, which is proportional to degradation at the base
- Knowledge of probabilities of mRNA degradation would help evaluate vaccine effectiveness computationally

Data Summary

- Reactivity for 68 bases in 2,400 strands were experimentally determined
- Structure and sequence of each strand provided
- n = 2,400 * 68 = 163,200
- μ = 0.3749 (for first experimental condition only)
- σ = 0.4469 (for first experimental condition only)
- Average reactivity by position on strand plotted below

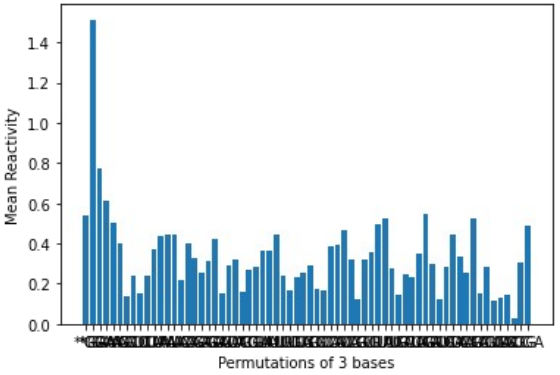


Task

Given the sequence and structure of bases in an mRNA strand, predict the reactivity at each base in the strand for five different experimental conditions.

Approach

Reactivity was grouped by unique reading frames of length 3 among all strands, which, as shown in the bar chart below, moderately discriminated reactivity.



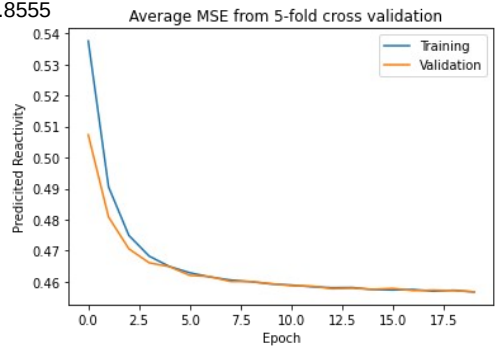
A feature set for each base, comprised of the following characteristics, was extracted from the strands:

- Index of the base in the strand
- Previous 3 bases, one-hot-encoded
- Distance to paired base (0 if base is not paired)
- Number of surrounding pairings
- Number of adjacent paired bases

A neural network with two hidden layers containing 10 and 5 neurons, respectively, was used to predict the reactivity of each base. A separate network was learned for each experimental condition. Because only 2,400 samples were available, 5-fold cross validation was used.

Results

After about 15 training epochs, the training and validation average mean squared errors (MSE) converged to 0.4567 and 0.4568 for the first experimental condition, as shown in the figure below. The average training and validation MSE across all five experimental conditions was 0.8552 and 0.8555



Analysis

Although the model improved its accuracy over the epochs, the final MSE loss is about the same as the standard deviation of the data set. This implies that the model does not do a good job of estimating the reactivities of the bases. The submission ranked 1,385 out of 1,636 on Kaggle with a public score of 0.40471.

Conclusions

The author hoped that extracting a feature set for each base and applying a multi layer neural network to the feature sets would predict the reactivity as well as a recurrent neural network would. However, such was not the case, as recurrent neural networks are known to have predicted reactivities with half the error that the traditional neural network in this approach did. This work convinced the author that, indeed, recurrent neural networks are better suited for sequenced data than traditional neural networks.