Springboard Capstone 3 Final Report: Predicting RNA Secondary Structure from Sequence

Objective and Data Source

The goal is to predict secondary structure information from RNA sequence alone. The dataset in question is a slew of sequences for which nucleotide-level reactivity profiles are available with two different probes: DMS, which reacts with the base-pairing surface, and 2A3, which reacts with the backbone. These are used to experimentally determine RNA secondary structure; we would like to develop a model that can provide a computational alternative. The work below uses the QUICK START dataset from the Stanford Ribonanza RNA Folding Kaggle Competition (https://www.kaggle.com/competitions/stanford-ribonanza-rna-folding/data).

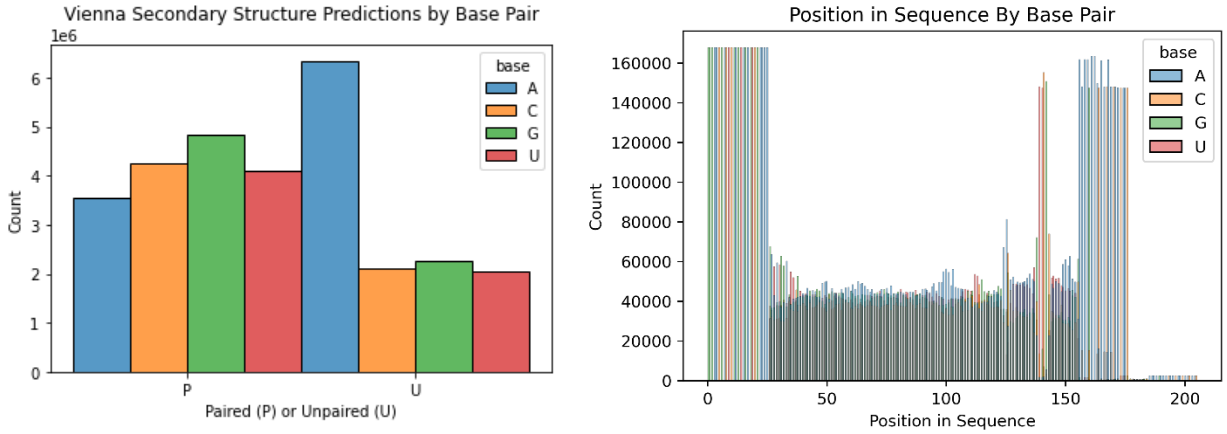
Data Preparation

The QUICK START dataset contains one RNA sequence per row; these are duplicated such that each sequence has one row with the DMS-derived data and another with the 2A3-derived. The reactivity values are normalized so that the 90th percentile value within each dataset is 1.0. Only sequences that contain data for both probes are included, and no duplicate sequences are present. The dataset presented is a subset of a larger one assembled by Stanford scientists; however, competition details were not specific as to whether these were assembled from multiple datasets (implying different labs and potentially different methodologies) or a single data source.

Upon downloading, the data were immediately split into a training and test set. The training set was then converted from one row per sequence to one row per nucleotide. The features detailed were the base identity, the base identities of two neighbors on either side, the position in the sequence (numerical from the 5' end, 1-indexed), and whether it is predicted to be paired (P) or unpaired (U) by the Vienna RNA algorithm (http://rna.tbi.univie.ac.at/). For each row, we are predicting a vector with two positions, corresponding to the DMS and 2A3 reactivities.

These features were chosen because during initial data exploration, the Vienna algorithm prediction and position in the sequence seemed to show differences by base (Figure 1). The neighbor base identities were included to account for the structural impact of base stacking and otherwise account for sequence context. While a simple LSTM-based translation model was not promising on its own, we know that the particular sequence does impact structure and could not be entirely discounted. This information is therefore encoded instead in the neighbor columns and Vienna predictions.

Figure 1. Selected properties by base pair.



Model Development

A Sequential Keras deep learning model with two hidden dense layers with 50 nodes each seemed to immediately produce predictions with strong MAE scores for both training and validation. To identify areas for potential improvement, I took an OFAT approach to determine where attention would be best spent. I evaluated the number of nodes per dense layer, the number of dense layers, and the batch size during fitting. None of these seemed to impact the MAE substantially; furthermore, the MAE was very comparable to the mean experimental error observed in the dataset. The original model was thus chosen as the one to move forward; when evaluated on the test set split off in the initial step, the MAE was 0.22.

Model Findings/Feature Importance

Figure 2. 2A3 prediction SHAP beeswarm plot.

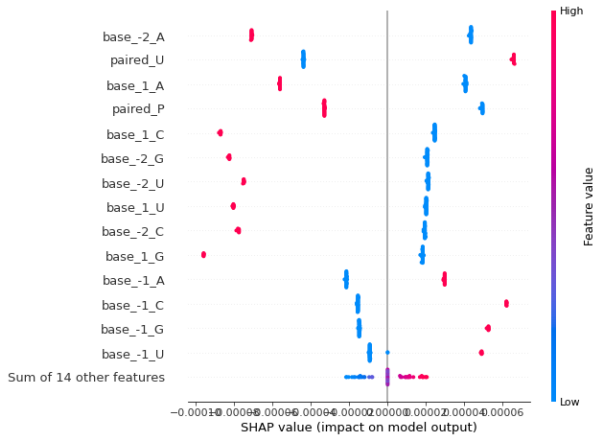
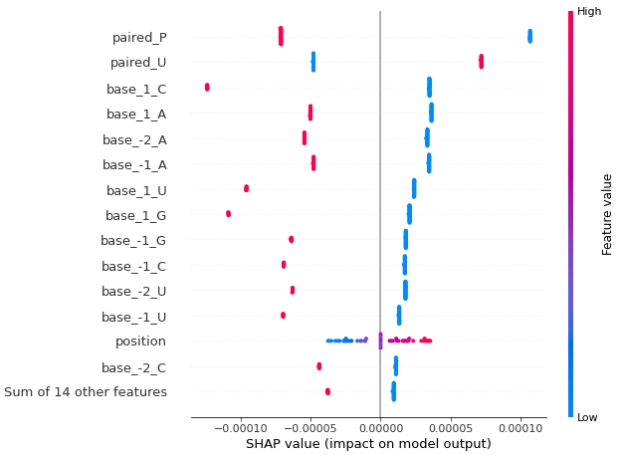


Figure 3. DMS prediction SHAP beeswarm plot.



The beeswarm plots of the SHAP values for the model are shown in Figure 2 for 2A3 probe predictions and Figure 3 for DMS probe predictions. The Vienna predictions are high up on the list for both, although they are the two most important features for DMS. Since the readout is expressly as to whether the nucleotide

is paired or unpaired, it is not particularly surprising that the impact of the readout would be stronger for their ability to react with a probe on the base-pairing surface. For both probes, neighbor nucleotide identity makes a difference, although being towards either end of the sequence doesn't seem to play much of a role. It should also be stated that all the SHAP values are very small.

Next Steps

There are a lot of interesting follow-ups that can be done, depending on where one's particular interests lie. Some potential areas for investigation include:

1. How does the number of neighbor nucleotide identities impact the model predictions?
2. The overall length of each sequence was not included as a feature. Would that have impacted the predictions?
3. Is the prediction accuracy agnostic to probe type? RNA functional and/or structural families?
4. RNA structure is inherently highly dynamic, something that is not captured in a dataset/model like this. Additionally, the model is not able to comment on how changes in conditions (e.g. pH, temperature) would impact the structure. Being able to incorporate these types of nuances into the model would be hugely beneficial.