Using Machine Learning to Predict Base Stacking and Solvent Accessible Surface Areas (SASA) for RNA

Carl Klein[1]

[1]University of Michigan, Department of Biophysics, Ann Arbor, Michigan

ABSTRACT

Base stacking interactions in RNA molecules play a key role in maintaining and defining the structural integrity of the macromolecule. Recent advances in using machine learning techniques for solving problems in biophysics has led to the creation of various predictive models that can predict these interactions. Predicting these interactions is very useful for predicting RNA structures from experimental data since the normal structural methods can often be very difficult for some RNA molecules. Likewise, machine learning can also be used to create a regression model to predict the Solvent Accessible Surface Areas (SASA) of a given RNA molecule. The resultant models for both of these biophysical problems show the ability to consistently and accurately predict these interactions on the given test set.

Introduction

Due to the difficulty in characterizing its structure using normal structural methods like X-ray crystallography or cryo-EM, RNA is often the target for using computational methods to predict various structural and molecular interactions. In particular, interactions such as base paring and base stacking play an important role in defining and stabilizing the secondary structure of RNA molecules. Ordinarily, these interactions would be classified by examining the structure of the molecule, but in this case NMR chemical shift data was used instead to create a machine learning model capable of predicting base stacking interactions without the use of any other structural data. Chemical shifts, an experimental measurement resulting from NMR, gives us useful information about the local interactions between neighboring atoms that allows us to infer the secondary structure of the macromolecule.

The Biophysical Challenge

The biophysical challenge presented by RNA is that predicting its structure can often be difficult due to the large number of intra-molecular interactions between its components as well as the large number of degrees of freedom for possible structural confirmations. Predicting the important interactions of RNA and their effect on structure is of great importance in order to better understand their functions aside from their standard role in transcription. Despite the challenges, predictive models and algorithms have continued to advance the field of RNA structure prediction.

State of the Art Techniques in Biophysicsal Machine Learning

Machine learning and deep neural network models have recently begun to find their way into biophysical research in an attempt to predict macromolecular features and interactions that would otherwise be difficult or expensive to determine using traditional experimental methods. The models outlined in this paper make use of the latest version of Scikit Learn's Multi-layer Perceptron classifier neural network algorithm, Random Forest Classifier algorithm, and Stochastic Gradient Descent classifier algorithm. In addition, predicting SASA values is considered a regression problem, so the models used to predict SASA values used a Random Forest Regression model. For all models, the baseline model trained with default hyperparameters was further optimized by using Scikit Learn's Randomized Search algorithm to select the hyperparameters that gave the best results.

Training and Evaluation of the Models

The dataset being used for predicting base stacking contained a total of 3068 samples, each containing the information of whether it participated in a stacking interaction as well as the 19 chemical shifts that were used as the features in the model. Information such as base pairing, orientation, and sugar puckering was excluded from the model. Due to Scikit Learn's algorithms only accepting numerical values, before training the model the string information representing the base stacking status was one-hot encoded as 0 for non-stacking and 1 for stacking interactions. Once the data was properly encoded, the

features (chemical shifts) were scaled using StandardScaler and then the dataset was split into training and testing sets to complete the preprocessing step of the model.

Training the Models

The general approach for training on all model types was to use 70 percent of the dataset for training a baseline model with default hyperparameters before optimizing the model using a randomized search for the best performing hyperparameters. This was done several times for each model in order to validate the performance of the model as well as getting a distribution of results to gauge the reliability of the model.

Base Stacking Models: Evaluation and Comparison

The base stacking dataset was trained and evaluated on three different model types, a random forest model, a stochastic gradient descent (SGD) classifier, and a Multilayer Perceptron (MLP) classifier. Out of all the unoptimized baseline models, the random forest model had the highest f1 score for predicting base stacking, at a score of 0.93. However, the baseline MLP model had comparable results for predicting base stacking (f1 score=0.92), and actually performed better on predicting the non-stacked residues (f1 score=0.28). As for the SGD model, results tended to vary more than the other two models, and generally performed worse overall, especially for predicting non-stacked bases. Optimizing the models lead to marginal improvements to all the model metrics, with noticeable gains in f1 score for the non-stacked bases. In general, the trends seen in the unoptimized models carried over, with the random forest model performing the best overall (f1 score=0.94), with the MLP model giving comparable results (f1 score=0.93).

SASA Models: Evaluation and Comparison

The SASA dataset was only trained and evaluated using a regression type random forest model. Five separate models were created for predicting SASA for all atoms, main chain SASA, sidechain SASA, non-polar SASA, and polar SASA respectively. For each model, the chemical shifts used for the base stacking model were used as the features for the target category of SASA values being predicted. When evaluating the performance of the model, r2 scores were used instead to due to predicting SASA values being a regression problem. The resultant r2 scores for the five models are as follows: 0.31 for all atoms, 0.36 for sidechains, 0.22 for the main chain, 0.38 for the non-polar residues, and 0.28 for the polar residues. The results indicate that the model performs poorly for all types of SASA, with the lowest value being 0.22 for main chain residues. This makes sense, since regression problems tend to have trouble reaching a high degree of accuracy compared to classification problems that are less complex. In every instance of training and evaluating the models, the r2 score was always less than 0.4, indicating that random forest models might not be well suited for studying solvent related interactions.

Discussion: Results and Potential Improvements

The results for predicting base stacking were very promising for all model types (f1>0.9), but there is clear room for improvement when it comes to predicting non-stacked bases (f1<0.3). This result is presumably due to the number of stacking interactions found in the dataset greatly outnumbering the number of non-stack interactions, leading to lower performance. As with all machine learning models, both the performance on stacking and non-stack can be improved by obtaining more data for the model to train. Another potential way to improve the results of the base stacking models might be to try some more complicated deep neural network models aside from an MLP, such as TensorFlow or Keras. Overall, the results for the base stacking model are satisfactory, with the next step being to test the model on more chemical shift datasets. As for the SASA models, performance was poor across the board (r2<0.4) even with optimization of the hyperparameters. In terms of improvement, simply testing out other types of regression models would likely be a good way to improve the results of the current random forest model. In conclusion, the results of these models further emphasize the merits and challenges of using machine learning to tackle biophysical problems such as RNA structure and interactions that were difficult to characterize using past computational methods.

Methods

All models were coded in Colab using a python 3.5 runtime with the latest version of Scikit Learn. The code and datasets can be found at https://github.com/clklein16/Biophysics-435-base-stacking-project