Machine Learning Models to Predict Base-Stacking and Solvent Accessible Surface Area in RNA

Joseph Mekhael[*,+,2,4], **Andrew Looka**[+,1,4], **and Kyrillos Abdallah**[+,3,4]

[*]jmekhael@umich.edu
[1]University of Michigan Department of Astronomy
[2]University of Michigan Ross School of Business
[3]University of Michigan Department of Biophysics
[3]University of Michigan Department of Chemistry
[+]these authors contributed equally to this work

ABSTRACT

Over the last decade, there has been a greater emphasis placed on understanding the role of RNA in cellular processes. The 2006 discovery of RNA interference marked a new wave of RNA discovery and placed RNA research as a central topic of focus. In recent years, machine learning has begun to play a larger role in both basic scientific discovery as well as therapeutic drug development. The collaboration of humans and machines has been hailed by many as the next wave of cutting edge research. The onset of machine learning in the life sciences and performance of research in silico has allowed for both an increase in the research able to be investigated as well as a decrease in cost for research. Our team chose to investigate RNA base-stacking, a stabilizing interaction occurring when 2 bases a certain distance apart interact with each other leading a more stable RNA structure, as well as the solvent accessible surface area (SASA), also known as the Lee-Richards molecular surface, a way to visualize a macromolecule by displaying the area of the molecule that is accessible to the solvent. Base-stacking and SASA have been observed as major determinants of higher order RNA tertiary structure. Specifically, our team built and tested different types of machine learning models using chemical shift signals in specific RNA as an input to predict base stacking status of the different bases and five different types of SASA found in RNA. This research is part of the larger biophysical problem of predicting RNA folding patterns and ultimately 3D structure and configurations.

Please note: Abbreviations should be introduced at the first mention in the main text – no abbreviations lists. Suggested structure of main text (not enforced) is provided below.

1 Introduction

Base Stacking

With regards to the tertiary structure of nucleic acids, there are two primary components that contribute to structural stability: hydrogen bonding between complementary bases and base-stacking interactions. Base-stacking interactions are most prevalent with classic Watson-Crick base pairing, which is why stacking contributes significantly to the stability of the DNA double helix[1]. However, because of non-canonical base pairing interactions along with other H-bonding interactions, RNA stability is noticeably less influenced by base-stacking interactions; however, base-stacking in RNA remains a relevant stabilization interaction, especially when considering supramolecular complexes involving multiple RNA helices[1]. The physical basis of base-stacking interactions is a combination of multiple different types of interactions. The hydrophobic effect plays a large part, sequestering the hydrophobic bases in the interior of the RNA duplex while forming a sugar-phosphate backbone. Additionally, van der Waals interactions between the bases within one strand and between two stands also contributes to the overall stacking interaction. Lastly, some research has suggested that the aromaticity of the bases plays an important role in stacking, but this is currently being debated[2]. Regardless, base-stacking interactions are influential on the structure of an RNA molecule, thus affecting its function.[1,2]

Solvent Accessible Surface Area

The solvent accessible surface area (SASA), also known as the Lee-Richards molecular surface, is one way to visualize a macromolecule by displaying the area of the molecule that is accessible to the solvent[3]. The SASA relies primarily upon two factors: the van der Waals radius of the atoms of the surface and the choice of solvent. To determine the SASA, a "rolling ball" algorithim is used. In these simulations, a probe molecule (the solvent) is "rolled" along the surface of the molecule, determined by the van der Waals radius of the surface atoms. By calculating the SASA of a molecule, relative solubilities can

be determined between molecules and between different conformations of the same molecule[3]. Additionally, the SASA allows for analysis of the contact surface between protein domains.

Challenge and Importance

The biophysical challenge being tackled in this paper is using chemical shift data from specific RNA residues to predict both the base-stacking status of that specific residue as well as several different types of SASAs for a given residue. Base stacking is extremely important in RNA as it is a determinant of RNA structure post-folding. Since structure is so closely linked with function, being able to predict the base-stacking status of an RNA moves us one step closer to being able to predict an RNA's structure based on a given set of input variables. Similarly, SASA can be used to help predict different conformations as well as RNA interactions with different proteins and small molecules.

2 Methods

There were two biophysical problems the team set out to build models to characterize: base-stacking status in RNA molecules and different types of solvent accessible surface area (SASA). For each of these problems we will stereotype the problem, characterize the data available to us, and describe the model used to predict the output variable.

Base-Stacking Model

The goal of the base-stacking predictor was to use chemical shift data for specific residues in different types of RNA as input variables to predict if the specific residue was stacked with another residue in the RNA. The output of the model would be binary based on if the residue was base paired to another residue. It could not predict which specific residue the input residue would be paired to. Part of this was due to the data available to us to be fed into the model. The data, which can be found in the github repository under file name "Base-Stacking Data" is a 3069 x 28 matrix. It contains 104 different RNAs and for each RNA has a specific residue, its chemical shift data, and whether or not the given residue is base-stacked. The chemical shift data was given in the form of a float with two decimals, base-stacking status was a boolean integer with 1 indicating a base-stacked residue and a 0 indicating the residue was not base-stacked. There were other data points built into the data such as base-paring, sugar puckering, orientation, and pseudoknot status. These columns of the data set were taken out during the initial data processing and were not fed into the model.

In order to predict the base-stacking status of a given residue, 2 models were built. The first was a multi-layer perceptron (MLP). The MLP had 4 hidden layers of size 50, 50, 50, 50. The activation function was tanh. The solver was adam and we used a constant learning rate with an alpha of .0001. The second model that was built was a Random Forest Classifier. This Random Forest Classifier had 10 trees in the forest. The models were trained by using the Leave-One-Out (LOO) approach by splitting the original data set. This means that when the initial data was being processed, two different data sets were made, a train set and a test set. The train set was used to fit the original models for both the MLP and the Random Forest Classifier. After the two models were fit, the test set was used to test the accuracy of the model. Logistically, the model was used to predict the output of the test set using only the input values. The predicted output was then compared to the actual value as found in the data set.

Solvent Accessible Surface Area Model

The goal of the SASA predictor was to predict solvent accessible surface areas from chemical shifts. Specifically, there were 5 SASA values to predict: total SASA, the SASA of the main-chain, the SASA of the nucleobase, the non-polar SASA , and the polar SASA. The data for the SASA predictor was a 3069 x 28 Matrix. The reason this data set was smaller was because it used the same input chemical shift information from the Base-stacking predictor. We were able to match the id, resname, and resid columns in this data set against the "Base-Stacking Data" to pull in the chemical shift information to make the regression model. The data can be found in the github repository under file name "SASA Data." Stereotyping the data, the chemical shift data was exactly the same as those used in the base-stacking predictor. The new outputs for SASA were floats with two decimals.

To build the SASA predictor we used three different types of models: a MLP, Random Forest Regressor, and a DeepChem Multitask Regressor. The MLP model had 1 hidden layer of 25 neurons, the activation function was relu, solver was sgd, alpha was .0082, and an adaptive learning rate. The Random had 10 trees in it. The DeepChem model had a 8 hidden layers of size 1000, 1000, 500, 400, 300, 200, 100, 10. It used a dropout of .5 and ran for 10 epochs. Similarly to the Base-Stacking Predictor models, the SASA predictive models were trained and fit in a similar fashion. First, the data was split into train and test data using the LOO approach. After the data was split, each model was fit against the train data. Once the models were fit, each was tested against the train data set to compare the predicted output against the actual output in order to establish accuracy of the model.

3 Results

Base-Stacking

To evaluate the Base-Stacking model, the F1 score of the model was calculated and used as the evaluation metric. F1 score is a measure of a model's accuracy that considers both precision (calculated as true positives divided by the sum of true positives and false positives) and recall (calculated as true positives divided by the sum of the true positives and false negatives). The equation for the F1 is $(2 * recall * precision)/(precision + recall)$. The two models for predicting base-stacking gave very similar F1 scores. The MLP gave an F1 score of .82. The Random Forest Classifier gave an F1 score of .88. on the testing set.

Solvent Accessible Surface Area

To evaluate the SASA models, the pearson R^2 was used as the evaluative metric of choice. The R^2 provides a measure of how much of the variance in outcomes is predicted by the model and thus measures how well the observed outcomes are replicated by the model. R^2 was used as the evaluative metric for SASA vs. F1 score because of the difference in the type of output. The F1 score can be used to measure accuracy when the output is categorical (like a boolean value) where the difference between categories is the same. It cannot account for the magnitude of difference in a continuous possibility of outputs. The R^2 value takes into account the magnitude of difference in the prediction. It is thus both the better and only metric between the two to evaluate regression models. The R^2 score for each of the 5 outputs for the 3 models can be seen in Table 1 below:

Model	total SASA	SASA of the nucleobase	SASA of the main-chain	non-polar SASA	polar SASA
MLP	.062	-.544	-.626	-.409	-.267
Random Forest	.128	-.061	.32	.145	-.043
DeepChem	.015	.011	.006	.007	.014

Table 1. R^2 scores for the three predictive models for SASA

4 Discussion

Base-Stacking

When looking at the F1 score for both the MLP and Random Forest Classifier models, we observe roughly similar F1 scores of .83 and .88 respectively. An F1 score of 1 indicates a perfectly accurate model on the test set and a score of .5 indicates a model no better than random guessing. This means that both models, although better than random guessing, still have room for improvement. There are several reasons why are model may not be perfectly accurate and a few ways to possibly improve. First, more data points in the train set would help make each of the current models more accurate as it gives a larger set to learn and fit from and help raise the model's accuracy. Second, the underlying data set from which we are training may be skewed so that there is an imbalance between base-stacked and non base-stacked RNAs. The way to find out definitively would be to compare these two numbers in the original training data set that is made on the model. If it was found that there were imbalances in the underlying chemical data, then a weight matrix could be made to help balance the side that is lacking in numbers, whether it be the base-stacked or not, so that the model could more accurate in its predictions. Third, the underlying chemical shift may not be a perfect predictor of base-stacking status. If this is the case, then additional inputs would be needed to make up for the information that chemical shifts lack. This also means that the chemical shift data alone could not produce a perfect or near perfect predictive model. One final way we could have improved the models would be to further optimize the hyper parameters in each of the models. In our case, we used the "best params" function to find the best hyper-parameters from a random search and manually changed the number of trees in Random Forest Classifier. If we had access to greater computational power, we could search for the best hyper parameters across a wider array of parameter combinations to further optimize the model.

Solvent Accessible Surface Area

When comparing the R^2 scores of the three models, there are several high level insights we can derive initially. First, none of the three models are great at predicting the different SASA values given chemical shift information alone. Second, although different models were better at predicting specific types of SASA (i.e. Random Forest for SASA of the main-chain), overall, the DeepChem was the best predictive regression model. The negative R^2 scores were of particular interest as the researchers were under the impression that R^2. Further research indicated that a negative R^2 score was possible as the function evaluated against a null hypothesis of a horizontal line, no change. This means that a negative R^2 score indicated that the chosen regression followed an opposite trend and was a particularly bad predictor.

In terms of reasons for the poor predictions, there are three possible explanations that could explain the poor results across

all three models. First, the underlying output values for the SASA data might be skewed in either direction giving a misrepresentation of what is found in nature. If the data were heavily skewed, it would make it difficult to predict the output from the chemical shift data. To confirm this suspicion, each set of SASA data would need to compared against a large, random population's data and compared to see if the differences were statistically significant. If a significant difference was found, then transformers or a weight matrix could be used to transform the data before feeding it into the model. Second, further optimizing the models through changing the hyper-parameters could potentially improve the R^2 score. The authors don't believe that hyper-parameter optimization would have yielded much success for two reasons. First, hyper-parameters were already optimized based on different random searches and best parameter functions and additional optimization would yield a diminishing marginal return. Second, all three models were nearly identically inaccurate showing that this lack of relationship between chemical shift data and SASA outputs may be model agnostic. The final reason for the poor model performance, and the one the authors believe to play the biggest role in performance, is a lack of relationship between the chemical shift data and SASA. The chemical shift may have no relationship with the SASA that the model can exploit to predict a SASA that is close to what is observed in nature based on chemical shift data alone. In this case, better inputs would need to be found that map closer to SASA outputs to build a more accurate predictive regression model.

5 Conclusion

In this study, we sought to build deep learning models that could predict a residues different SASA values as well as base-stacking status based on chemical shift data. The different models built showed that chemical shift could be used a predictor for base-stacking status, but was a not a strong predictor for any of the five SASA values. Further analysis would need to be done to optimize the models, add additional inputs to better predict, and see if there was truly an underlying relationship between chemical shifts and our outputs.

References

1. Yakovchuk, P. e. a. Base-stacking and base-pairing contributions into thermal stability of the dna double helix. *Nucleic acids research* **34,2**, 564–74, DOI: 10.1093/nar/gkj454 (2006).
2. Šponer J., e. a. Nature and magnitude of aromatic base stacking in dna and rna: Quantum chemistry, molecular mechanics, and experiment. *Biopolymers* **99**, 978–88 (2013).
3. ML, C. Solvent-accessible surfaces of proteins and nucleic acids. *Science* **221**, 709–13, DOI: 10.1126/science.6879170 (1983).