Machine Learning in Predicting RNA Base Stacking and Solvent Accessible Surface Area

Yujin Wu[1,*], Thomas Stewart[1,+], and Luis Fernando Cervantes Vasquez[1,+]

[1]University of Michigan, Department of Chemistry, Ann Arbor, 48109-1055, USA
[*]wyujin@umich.edu
[+]These authors contributed equally to this work

ABSTRACT

RNA molecules play diverse functional and structural roles in cells. To better understand the molecular mechanism of any interested RNA, it is critical and essential to have a clear picture of the RNA structure. In-silico tools and computational techniques have become an integral part of biophysics research. Thus, we explore several different machine learning algorithms to predict two key features, base stacking and solvent accessible surface area (SASA), of the RNA structure.

INTRODUCTION

The main focus of structure prediction and engineer of biological molecules has been on proteins in recent decades due to the relative ease with which their structures can be crystallized, selectively engineered, and tied to function. However, as the field progresses, there has been an increased focus on understanding the complex role that RNA structures play in gene expression and overall control of biological systems. These miRNAs have been credited with controlling key cellular processes including cell growth and differentiation, even implicating these molecules in some forms of cancer. The key regulating attributes of these RNA molecules come from their ability to base pair with other DNA and RNA molecules, making them highly specific and directly related to the function of DNA transcription and translation. However, there are additional properties to consider that determine how these RNA molecules behave in biological systems.

Base Stacking

One of these properties is the ability of the residues to base stack and form more stable tertiary structures. The bases of RNA residues are planar in nature and staking these planes of bases at a distance of 3.4A allows for the exclusion of water molecules and the maximum contribution of Van der Waal interactions to stability of the overall molecule. This stabilization can even play an important role in binding and function of protein/RNA complexes like in the case of RNA-chaperone Hfq molecules. These Hfq molecules catalyses the annealing of bacterial small RNAs (sRNAs) with target mRNAs to regulate expression of certain genes[1]. This particular example shows several different instances of RNA base stacking modulating some functionality to ultimate control larger biological systems.

Given the potential power over regulation that they offer, small interfering RNA (siRNA)-based drugs are being explored as potential therapeutics for gene expression related diseases such as cancer[2]. These drugs could offer increased specificity of treatment given their function requiring specific base pairing. This would lead to highly specific drugs that have very low off target effects. However, the development of such drugs is still limited by the fundamental understanding of RNA structure and function. Being able to engineer and predict such attributes would represent a substantial leap forward in the development of these therapeutics.

These potential applications have encouraged a number of groups to attempt to find ways of predicting RNA structure and interactions. There are currently many software packages available for predicting secondary structure and base pairing of RNA molecule. Some of these packages simply attempt to base pair different portions of an RNA molecule to predict the most likely secondary structure, which is a relatively simple problem to solve. One such package is the RNA structure package which includes predictions for: secondary structure prediction, base pairing probabilities, biomolecular structure prediction, and prediction of a structure common to two sequences[3]. These kind of software packages can provide some insight into RNA functionality but leave out a lot of potentially important information and make no effort to predict tertiary structure. Other groups have attempted to solve this problem by identifying tertiary structure motifs in crystal structures and using these motifs for structure prediction[4]. The inherent difficulties with this type of approach stem from the fact that RNA molecules can be difficult to crystalize and that RNA structures tend to be less defined and globular than proteins.

Solvent Accessible Surface Area

Solvent accessible surface areas (SASAs) are often used as an analysis tool by structural biologists. The intramolecular hydrogen bond is an important component of the driving force for RNA secondary structure[5]. SASA could reflect the amount of the hydrogen bond interactions in the RNA which could be further used in predicting the RNA structure.

To avoid generating the crystallization for a target RNA and then analyzing the structure information, in the following, we explore several different machine learning algorithms to predict both base stacking and SASA based on chemical shifts of a target RNA, which is more easily to be acquired. The final models could be acquired from the github. The solvent accesible surface area (SASA) provides a quantifiable measure of how much contact a specific region in the surface of a molecule (usually a macromolecule such as RNA) has with the solvent. This can serve as a marker in order to understand what types of amino acids (in the case of proteins) or bases (in the case of RNA and DNA) play a key role in the formation of certain interactions that mediate key biological functions and responses. Therefore, in this study, we have developed regression models that use solely chemical shifts as features in order to predict 5 types of SASA, namely polar, non-polar, side-chain, main-chain and all atom SASA for a specific RNA base within a sequence. These models could allow for the accurate prediction of SASA that can then be used in order to predict the SASA of new RNA structures of interest.

MATERIALS AND METHODS

Benchmark Dataset

The chemical shift dataset and the corresponding base stacking and SASA values are provided by the Prof. Aaron's lab (a total of 104 RNAs).[6]

Methods

All models are built based with python script. Sklearn[7] packages are used.

SASA Models

A random forest regressor from sklearn was implemented as a baseline and validated using the leave-one-out (LOO) method by leaving out a specific RNA's chemical shifts as testing sets on which to calculate metrics.

An extended trees regression model from sklearn was also implemented in order to compare to the random forest baseline regression model. Hyperparameters that were optimized where the number of estimators, which refers to the number of decision trees used for the model, as well as the maximum number of features, which indicates how many features are used per decision tree before arriving to a regression result. The hyperparameters that led to the highest r2 scores were 100 estimators using the square root of the number of features as the maximum number of features used per decision tree. While results are not shown, a multi-output deep neural network was constructed using the ReLU activation function for 3 deep layers, involving a dropout rate of 0.25 with 1000 neurons each. Optimization of these and other hyperparameters was attempted, but resulted in very low r2 scores, often times even negative and with an absolute value greater than 1. Different multi-output deep neural network architectures were attempted, but none led to an r2 score of greater than 0.16 for the best SASA type prediction.

RESULT

Base Stacking

We first use a simple classification model without any hyper-layer classifier as the baseline model. The feature vector contains the chemical shifts of the target residue as well as its neighbors. The number of neighbors is set from 0 to 5. The leave-one-out validation test is performed against this model to get the overall accuracy as a function of number of neighbors. The results is shown in table1. With the increase of the number of neighbors, the f1-score and the precision slightly drops and the recall value become 100%. This might results from the fact that the features for each residues become similar to each other with the increase of chemical shifts adopted. On the other hand, given the fact that the base stacking has a much higher percentage in the dataset, it is more likely for the model to classify a residue with base stacking so that it will have a higher recall and lower precision. We then use another random forest classification model which gives has a f1-score around 0.9205 ± 0.0860. Both the random forest classification model and the baseline model has a high accuracy ($> 92\%$) in predicting the base stacking. To reduce the computational cost, we select the baseline model as our final classification model.

To further understand which of the 19 chemical shift types are the most important for accurately predicting base stacking status. We perform another leave-one-out experiment with the number of neighbors set as 2 which gives the best f1-score in the previous experiment. Since the precision is much lower than the recall and could better reflect the accuracy of the model. We then plot a heat map to visualize the change of precision against the average precision we acquired in the previous experiment. (Figure 1).

Table 1. Overall Accuracy in Base Stacking Prediction

Num of neighbors	f1-score	Recall	Precision
0	0.9227 ± 0.0859	0.9985 ± 0.0080	0.8677 ± 0.1272
1	0.9229 ± 0.0807	0.9996 ± 0.0034	0.8670 ± 0.1265
2	0.9225 ± 0.0853	1.0 ± 0.0	0.8661 ± 0.1267
3	0.9220 ± 0.0855	1.0 ± 0.0	0.8653 ± 0.1269
4	0.9205 ± 0.0860	1.0 ± 0.0	0.8648 ± 0.1276
5	0.9205 ± 0.0860	1.0 ± 0.0	0.8628 ± 0.1276

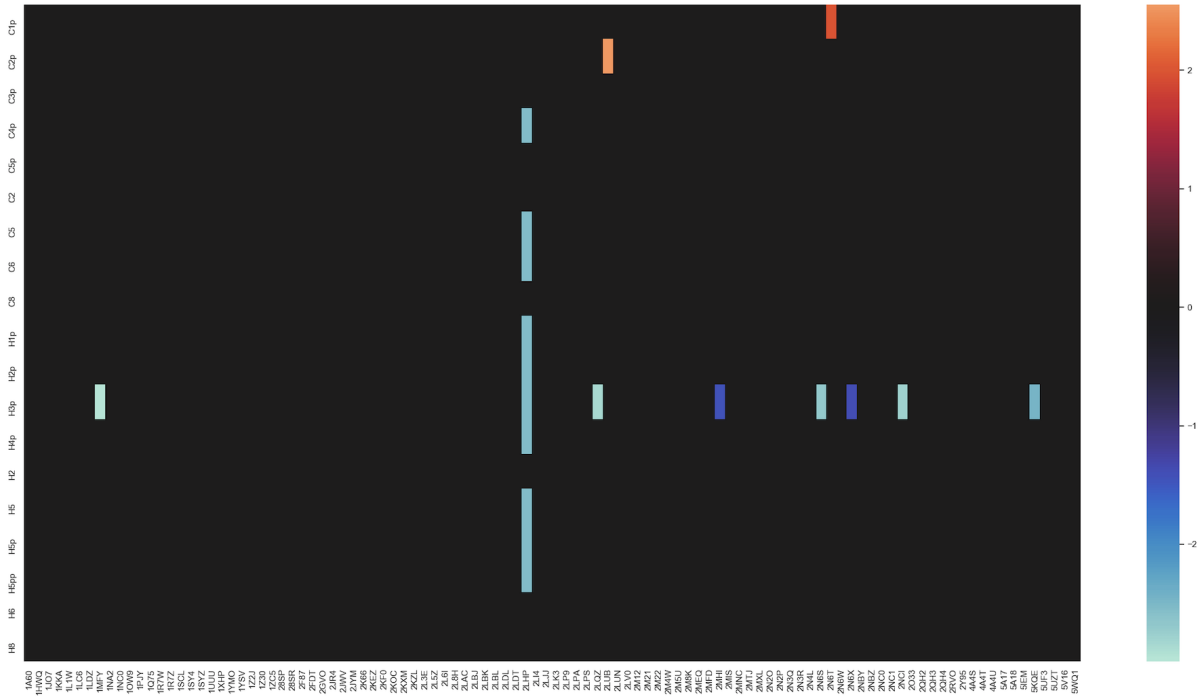

Figure 1. Change of Precision as a Function of Chemical Shift and RNA. The difference in precision is multiplied by 100.

As shown in the figure 1, most of the chemical shifts do not affect the precision except H3p which cause the decrease in precision against 8 RNAs. Another interesting case is 2LHP which has a lowered precision when 10 different chemical shift is dropped. This might results from uncommon RNA secondary structure which is under-represented in the training set. A more robust model will be developed in the future to avoid lower accuracy in predicting base stacking in the uncommon RNA structures.

SASA

Again, we first use a simple multi-linear regression model as the baseline model to predict the SASA. In the simple multi-linear regression model, we only have an r2-score around $10 \sim 20\%$ in the leave-one-out test. This might results from too many numbers of parameters[8]. Thus, we then use ridge regression to avoid this potential issue[8]. This increase the r2-score to $25 \sim 40\%$. However, the r2-score for the in predicting the training set is below 50%. This shows that the baseline model is insufficient in predicting SASA from chemical shifts. Thus, we apple a random forest multi-linear regression model. The accuracy is shown in the table 2.

From table 2, we could find when the number of neighbors is 2, the model gives the best result. In predicting the SASA value from the non-polar atom, we have the highest accuracy. However, in predicting other SASA contributions, our model does not improve too much compared with the baseline model. Another improvement in the random forest model is the r2-score in predicting the training set increased to around 90%.

Another multi-output model that was tested for SASA types prediction is the extended trees model. The results are shown in 3. While not as accurate for the non-polar SASA, it does provide some interesting consistent results with the random forest regression model.

Table 2. Overall Accuracy in SASA Prediction in Leave-One-Out Test Using Random Forest Regression

Num of neighbors	Polar	Non-polar	Side chain	Main chain	All atoms
0	0.3580	0.8677	0.3392	0.2261	0.3322
1	0.5035	0.8670	0.4425	0.4061	0.4803
2	0.4977	0.8661	0.4588	0.4126	0.4927
3	0.4958	0.8653	0.4445	0.4072	0.4836
4	0.4918	0.8648	0.4535	0.4093	0.4891
5	0.4861	0.8628	0.4426	0.4009	0.4801

The accuracy is the average accuracy in the leave-one-out test.
The standard deviation is around 25%.

Table 3. Overall Accuracy in SASA Prediction in Leave-One-Out Test Using Extended Trees Regression

Num of neighbors	Polar	Non-polar	Side chain	Main chain	All atoms
0	0.2721	0.3194	0.2350	0.3576	0.3306
1	0.3510	0.4842	0.4216	0.5277	0.4442
2	0.3432	0.4922	0.4358	0.5202	0.4521
3	0.3475	0.4894	0.4302	0.5190	0.4466
4	0.3437	0.4965	0.4302	0.5148	0.4461
5	0.3413	0.4869	0.4364	0.5092	0.4441

The accuracy is the average accuracy in the leave-one-out test.
The standard deviation is around 25%.

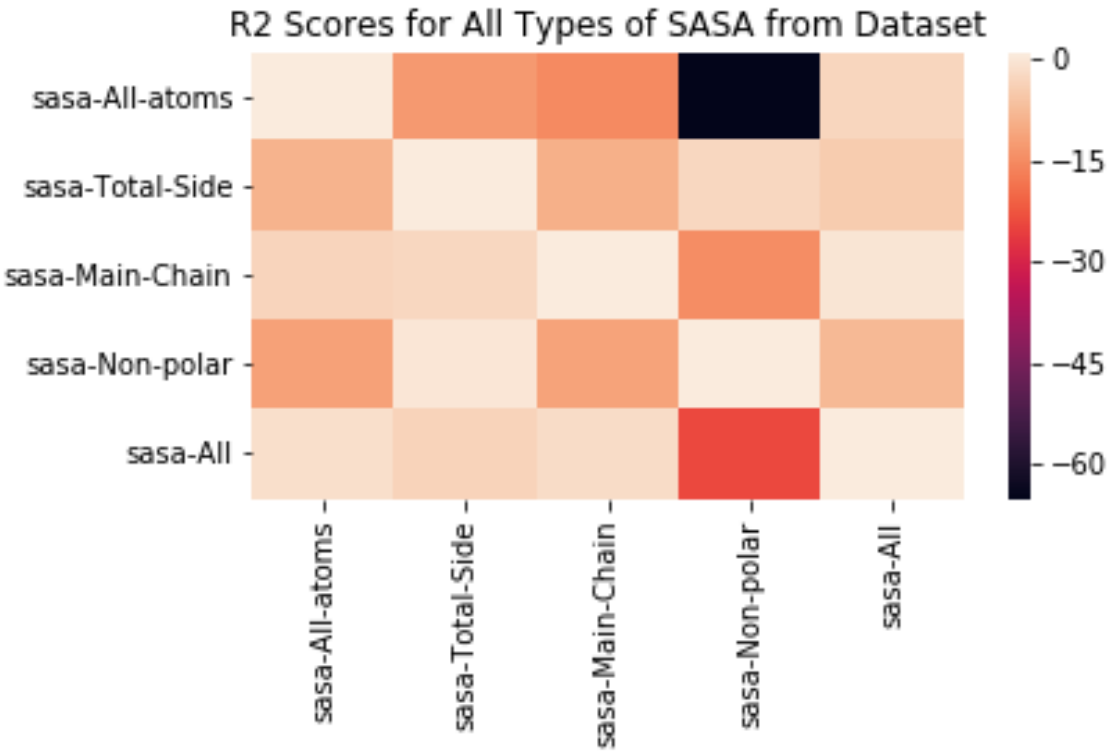

Figure 2. R2 Scores from the overall dataset used to predict different types of SASA.

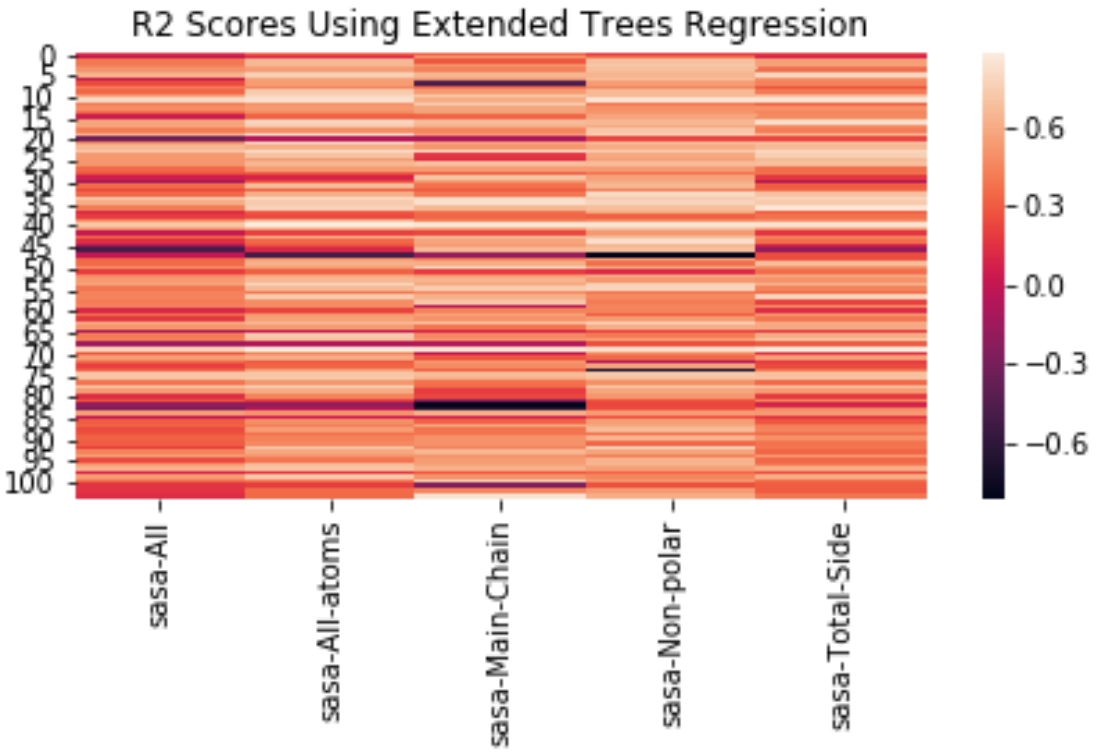

Figure 3. R2 Scores from the overall dataset used to predict different types of SASA.

Discussion

Our results in predicting the base stacking shows that it is not necessary to construct a complex model. Our baseline model actually has a higher f1-score, recall and precision. In the future work, it is more important to construct a more robust model so that it could more accurate in predict uncommon structure such as 2LHP.

Our results in predicting the solvent accessible surface area (SASA) is low. It might be more beneficial using a linear regression towards each target instead of using a multi-linear regression model. Our simple normalization with StandardScaler function in sklearn does not significantly increase the result. Another focus will be developing a new method in normalizing the target values(y) in both training set and testing set.

To analyze multitask regression framework for predicting the five types of SASA, the correlation as measured by the R^2 score needs to be assessed. Looking at the heatmap for the raw R^2 scores from the raw dataset in Figure 3 for the different types of SASA, it can be seen that the correlation between the different tasks are negative. Moreover, it seems that some are anti-correlated. This might explain why predicting these different types of SASA under the same multi-output regressor dataframe yielded such poor correlations, especially for the extended trees regression method, which performed more poorly in general than the baseline random forest regression model. Generally, a multitask regression model works best for tasks that are related in some way, where the true values can be normalized in some way in order to give a linear relationship before learning and the normalization can be undone after prediction. However, in this case, while there is some type of trend among the data from visual inspection, the R^2 scores, as well as some clustering associated with the tasks shows that the tasks are actually not as correlated. It has been shown that if the tasks are not related, this might actually bring prediction shortcomings of this model (Evgeniou and Pontil, 2004). Therefore, it might be possible that a specific model for each SASA task would be of optimal choice. This would involve tuning of the parameters associated with each model and optimizing based solely on one specific task at a time. Not even scaling of the outputs change the results as much.

The random forest regression model as a baseline provided very accurate results (about 0.86 R^2 score) for a specific type of SASA, namely the non-polar SASA, as shown in Table 3. However, compared to the extended trees regression model, the rest of the SASA results are comparable, with an R^2 score close to 0.5 for the most accurate result, which involves a training and testing dataset containing chemical shifts for 2 neighbors on either side of the RNA base whose SASA is predicted.

In the LOO method for the extended trees multi-output regressor, it can be seen that the standard deviation is as high as that

for the random forest regression model. The median R^2 for all of the predictions was higher than the mean, indicating that the distribution of R^2 scores is negatively skewed. In fact, looking at the R^2 heatmaps for each of the neighbors containing PDB ID information in the rows and type of SASA in the columns shows that there are certain RNA PDB ID's that, when left out of the training of the model and used as test data instead, led to increased R^2 across the five different tasks. This leads to the conclusion that there are certain RNA PDB ID's that are either not representative of the entire dataset or perhaps the predictive model is just very accurate for that kind of RNA. Whatever the case may be, the RNA's that were excluded from the training dataset that led to the greatest increase in R^2 scores (up to 0.8 for some tasks and up to 0.67 for others) were identified and were consistent across the number of neighbors used for the LOO validation method. The RNA PDB ID's are 1PJY, 1Z30, 2L5Z, 2MIS, 2N3Q, and 2N3R. The first three correspond to RNA's with a stem-loop structure and the rest correspond to the Neurospora VS Rybozyme active site RNA, which contains a Mg^{2+} coordinating site (Bonneau, et a. 2015). While not a stem-loop itself, the VS rybozyme active site consists of an interaction between two internal loops, one which is a stem-loop. While the rest of the RNA PDB ID's were not analyzed, it is interesting that the types of RNA's that led to stronger predictions could be grouped into a single category of RNA. Therefore, the distribution of RNA PDBID's should be analyzed for the type of secondary RNA structure that is represented before making any judgements as to how one might want to group the datasets for accurate prediction. One could assume that the best approach to predicting SASA in this case might be to tune a predictive model per secondary structure, as well as per type of SASA.

References

1. Schulz, E. C. *et al.* Intermolecular base stacking mediates rna-rna interaction in a crystal structure of the rna chaperone hfq. *Sci. reports* **7**, 9903 (2017).
2. Bora, R. S., Gupta, D., Mukkur, T. K. S. & Saini, K. S. Rna interference therapeutics for cancer: challenges and opportunities. *Mol. medicine reports* **6**, 9–15 (2012).
3. Reuter, J. S. & Mathews, D. H. Rnastructure: software for rna secondary structure prediction and analysis. *BMC bioinformatics* **11**, 129 (2010).
4. Batey, R. T., Rambo, R. P. & Doudna, J. A. Tertiary motifs in rna structure and folding. *Angewandte Chemie Int. Ed.* **38**, 2326–2343 (1999).
5. Cavallo, L., Kleinjung, J. & Fraternali, F. Pops: a fast algorithm for solvent accessible surface areas at atomic and residue level. *Nucleic acids research* **31**, 3364–3366 (2003).
6. Zhang, K. & Frank, A. T. Conditional prediction of rna secondary structure using nmr chemical shifts. *bioRxiv* 554931 (2019).
7. Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *J. machine learning research* **12**, 2825–2830 (2011).
8. Kennedy, P. *A guide to econometrics* (MIT press, 2003).

Acknowledgements (not compulsory)

Acknowledgements should be brief, and should not include thanks to anonymous referees and editors, or effusive comments. Grant or contribution numbers may be acknowledged.

Author contributions statement

Y.W. built the baseline model for bothe classification and regression task and analysed the result. Y.W. built the random forest regression model and analysed the result. T.S. built the random forest classification model and analysed the results. L.C. built the neural network classification model and analysed the result. All authors reviewed the manuscript.

Additional information

All scripts and result could be accessed from the github.