Utilizing Chemical Shift Data to Predict Base-Stacking Interactions in RNA

"Biophysics is the science of taking features of nature and translating them into computer models" - OpenAI's new machine learning model

Pujan Ajmera[1,+], Anna Argento[1,+], Kyle Finos[1,+], and Myra Visser[1,+]

[1]University of Michigan Department of Biophysics, Ann Arbor, 48104
[+]these authors contributed equally to this work

ABSTRACT

Base stacking, also known as π-π interactions, are attractive non-covalent interactions found between aromatic rings, particularly those in DNA and RNA. These interactions are prevalent in nucleotide structure, protein structure, materials science, and molecular recognition. Previous studies have displayed indirect relationships between Nuclear Magnetic Resonance Spectroscopy (NMR) and base stacking status of nucleic acids[1], but a complete theoretical model of what causes base-stacking interactions is not available. We used NMR chemical shift information in an attempt to predict base-stacking interactions using neural network algorithms. Using Google's Colab, we implemented algorithms such as multi-layer perceptrons (MLP) and a 1-dimensional graph convolution models to predict base-stacking status of various RNA nucleobases. Furthermore, we used the chemical shift data to predict five domains of the solvent-accessible surface areas (SASA) of these RNAs. The connections between the success of predictive models and the types of models provide direct insight into the molecular mechanisms behind base stacking in RNAs.

Introduction

It is well known that ribonucleic acid, or RNA, molecules are very complex and fine-tuned structures, influenced by their sequence as well as external parameters, such as pH, temperature, and protein environment. This makes predicting RNA structure solely from sequence a challenge without having significant intricate knowledge about the molecule. Many methods have been created to improve detection of RNA structure including, but not limited to, ligand kinetics, X-ray crystallography, and energy minimization evaluations. However, many of these approaches are very hands-on and require a massive amount of human intervention, and do not actually consider the sequence itself. Algorithms that predict RNA structure by computing free energy minimums based on sequence are among the most popular methods used. However, many RNAs don't fold to their minimum energy structure in the biotic environment; their functional roles come from dynamic structures that maintain a biological potential energy rather than optimal structures that meet minimum energy requirements. The longer the RNA sequence, the more complex it's secondary and tertiary structure, and the more its biological potential energy status deviates from the minimum energy status, making predictive models based on free energy minimization less accurate[2]. Furthermore, RNA structure goes through many transient intermediate states from the first few nucleotides exiting RNA polymerase to the final mature RNA.

With the uprising of NMR in the last four decades as a non-invasive imaging tool, more RNA chemical shift data is being uploaded into the PDB. Since chemical shifts depend on the electronic environment of the given residue, chemical shift data can reveal much about the local structure of an RNA. Therefore, we believe chemical shift data is an appropriate parameter to be introduced to artificial intelligence methods. Recent work[3] used machine learning methods to predict the base-pairing status of RNA residues based on their NMR chemical shifts. These base-pairing statuses were then used as restraints in RNA folding algorithms which accurately predicted the two known structures of microRNA-20b.

Similarly, we believe NMR chemical shift data can be used to predict the nucleobase stacking status of RNAs. Stacking interactions between adjacent bases is another major determinant and stabilizer of RNA tertiary structure. Base stacking interactions, a type of π-π interactions, occur when the electron orbitals of two aromatic rings overlap, forming a non-covalent, stabilizing interaction[4]. These interactions have been observed in many RNA structures, including ribozymes, tRNAs, and the genomic RNA of retroviruses[5]. Base stacking includes stacking between adjacent bases and coaxial stacking between two separate helical regions of the RNA which can form kissing loop interactions and pseudoknots.

Getting a complete picture of the structure of an RNA is very important in order to understand translation, gene regulation, RNA modifications, and RNA-mediated catalysis. In fact, two of the most important reactions in the cell, the condensation of amino acids in the peptidyl transferase ribosome center and eukaryotic mRNA splicing, are catalyzed by ribozymes[6]. These reactions depend on the structures of RNA, including base pairing, stacking and, most importantly, solvent accessible surface area, or SASA. SASA is defined as the surface area of a biomolecule that is accessible to a solvent[7]. As with base stacking, we will use NMR chemical shifts in order to predict five SASA values: the total SASA, the SASA of the backbone, the SASA of the sidechain, the non-polar SASA, and polar SASA.

Results

Multi-Layer Perceptron (MLP) Classification

The accuracy of our results for base stacking and non-base stacking classification were primarily represented in the form of f1 scores. An f1 score is a weighted average of the precision and the recall of the model, with 1 being the best value and 0 the worst. When training and testing the algorithms, the Leave-One-Out (LOO) approach was taken, as well as the Random Data Segmentation (RDS) approach. The former allows us to take a single-entity approach to obtaining total averages, whereas the latter allows us to obtain ensemble averages of the RNAs that are being tested.

Leave-One-Out (LOO)

The first MLP was constructed using sklearn MLPCLassifier, and the MLP was ran over all the RNAs, leaving out one RNA for testing and the other 103 for training. This data was used to obtain an optimum number of neighbors for further analysis. The average f1 score was obtained for number of neighbors (N) = 0, 1, 2, 3, 4, and 5 in three trials.

N	f1 score
0	0.8314 ± 0.0072
1	0.8290 ± 0.0006
2	0.8296 ± 0.0077
3	0.8300 ± 0.0019
4	0.8316 ± 0.0051
5	0.8285 ± 0.0021

Table 1. This table includes an average f1 score with standard deviation for 3 trials over varying neighbor numbers. F1 scores were obtained with MLP Classification Report

As shown in Table 1, the number of neighbors clearly does not affect the f1-score in any significant manner. Any changes in f1-score are accompanied by a larger standard deviation. Therefore, to obtain consistent and comparable results for further tests, we chose the f1-score that produced the least standard deviation, which was N=1. Note that this MLP does not include drop-out layers, as the MLPClassifier function does not permit it and the main goal of this network is to act as a baseline model for the other models created. Furthermore, for computational efficiency, N=0 was run for some of the experimental trials.

Random Data Segmentation (RDS)

A second MLP was constructed using Keras Deep Learning in order to create an algorithm that optimally predicts the base stacking of the RNAs. Keras Deep Learning tools were ideal for our application due to its layer-by-layer approach to building models. Keras permitted the use of drop-out layers and custom activation functions. Custom class weights were implemented due to the imbalanced (stacked: 87%, unstacked: 13%) dataset.

The results are reported as an evaluation of a testing set consisting of 30% of the data. The model had an f1 score of 0.93. More detailed insight into the training regimen can be viewed in the tensorboard graph in Figure 1, which compares accuracy as a function of epoch.

1-Dimensional Convolutions

Utilizing Keras' Conv1D to construct an elementary 1-dimensional convolutional network, the data was fit using the LOO and RDS approach described in Methods. The data was calculated for N=1, and the chemical shift input vector is as shown in Figure 2. The number of epochs ran was 3, as the accuracy was noticed to converge very quickly and we aimed to prevent overfitting.

The graph convolution ran kernels over this vector, and due to the ordering being fairly arbitrary, with the exception of being ordered numerically and grouped by atom, it is important to note that we can not expect this graph convolution to provide formal direct insight into the important chemical shifts/atoms that determine base-stacking status. Using the LOO method, an f1 score of 0.92018 was obtained.

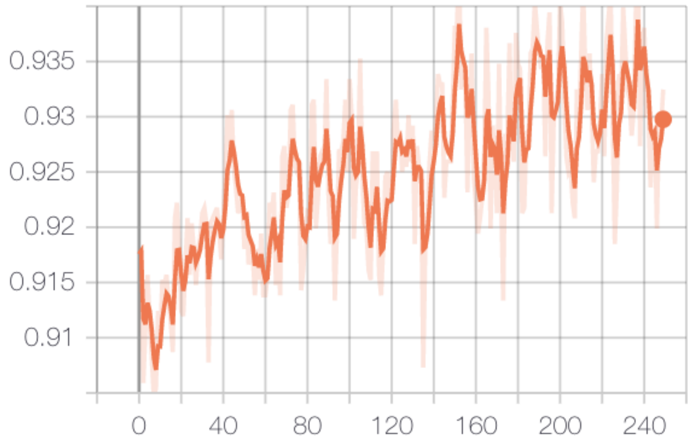
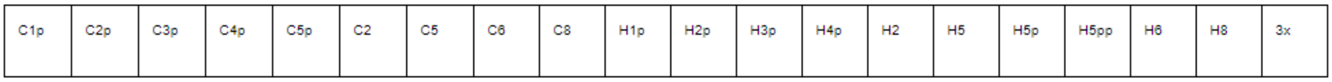


Figure 1. Tensorboard image shows accuracy increasing along further epochs, indicating a successful training regime



C1p	C2p	C3p	C4p	C5p	C2	C5	C6	C8	H1p	H2p	H3p	H4p	H2	H5	H5p	H5pp	H6	H8	3x

Figure 2. Vector showing the ordering of chemical shifts of the RNAs. '3x' indicates that this vector is triple of what is shown due to including 1 neighbor on either side of the nucleobase

Solvent-Accessible Surface Areas (SASA)

The solvent-accessible surface areas of RNA residues are reported in five domains: side-chain, main-chain, non-polar, all-polar, and all atoms. The same chemical shift data used for base stacking was used for SASA prediction. A multiclass regression neural network model was designed with all 5 outputs. Originally, the large variation in the five SASA parameters presented a challenge for the network; there was up to a five-fold difference between variables for residues. With this setup, an R2 score of only 0.15 was achieved, indicating low correspondence between predicted and actual SASA data. A significant improvement was realized by normalizing each SASA label independently. Under this format, a mean percent error of only 6.1% with an R2 score of 0.30 was achieved on a testing (30% of all data) dataset.

Discussion

Through varying deep learning approaches, we were able to successfully model physical aspects of RNA by experimenting with chemical shift data and residue type. Our models were able to predict base stacking with a robust F1 score (0.93) and solvent accessibility (6.1% error, 0.3 R2 correlation). For practical considerations, we believe that the base stacking performance of our model qualifies it as a legitimate predictive model. Yet it may still be improved with minimal modification by adding more labeled NMR data should it become available.

A 1-Dimensional graph convolution showed to be useful, with an f1 score of 0.92. This shows that there are certain patterns within the stated vector that are being exploited for the estimation of base-stacking. Further tests of leaving out specific chemical shifts could help narrow down this vector into one that only has pertinent chemical shift information to determine base-stacking status of nucleobases. Another method would be to directly look into the weights of the model, as it could give insight into which sections of the vector in Figure 2 are weighted higher and are therefore more important to determining base-stacking.

Our results show that our project benefited from the increased level of customization provided by Keras. By comparing our baseline model- a simple MLP classifier designed to choose the number of optimal neighbors- to our custom Keras model, we see that the Keras model provides a significant increase in performance considering its f1 score of 0.93 compared to 0.83. The obvious advantage of using a neural network for analyzing NMR is a significant decrease in computation power and analyzation time compared to traditional methods. Traditionally, base stacking was determined by arduously creating a 3-dimensional structure via multiple NMR experiments, which can take days to weeks and intense computational power. Next, structure would be manually analyzed for base stacking interactions by finding bases within 4 angstroms of each other. In contrast, our model shows a robust, albeit imperfect, prediction of base stacking from a far more simple chemical shift experiment. Our model would be able to predict base stacking for a given residue in a matter of seconds. We hope that in the future our model will be useful to experimenters who value swiftness in their analysis of RNA.

Our model for solvent-accessible surface area was similarly quick and requires low computation to run relative to traditional experimental methods. We believe that our model is an interesting proof-of-concept in showing relations between chemical shift and solvent accessibility. However, it likely has more limited research applications beyond base-stacking prediction because of lower performance. For example, an R2 score of 0.30 indicates that our model was only partially predictive of the variation in SASA data. We believe that model optimization would likely be fruitless, as our predictive abilities are subject to our limited chemical shift features. SASA is a complex structural property of RNA bases, and it is possible that chemical shifts simply can't fully describe the parameters that determine solvent-accessibility. In future research, we would like to improve this model by featurizing data from other NMR experiments in conjunction with chemical shift values.

Further analysis on the LOO models can be done as well to determine types of RNAs that have different deterministic chemical shifts. This is analogous to the single molecule approach to biophysics, in which subpopulations of an ensemble can be analyzed. Understanding and determining these subpopulations of the RNAs can also improve the robustness of the LOO models, as they have been shown to underperform the RDS model.

Many RNA researchers are currently probing different ways to model RNA structure with computational approaches. While energy minimization techniques used for secondary structure prediction remain the standard, some researchers have developed alternative methods for investigating tertiary structure, much like we have started to do by modeling base stacking. For example, Yesselman et al. used the conformational variation present in RNA helices from X-ray crystallographic data to build a thermodynamic model of tertiary assembly. Their ensemble model showed how sequence-dependent conformational effects of helical elements can influence the thermodynamic stability of RNA tertiary structure with high quantitative accuracy[8].

Another method includes using RNA homologs to improve tertiary structure prediction. This is based on the observation that RNA sequences from the same RNA family will fold into a conserved structure, and parallel modeling of these homologs provide insights that other models have yet to exploit[9]. As far as machine learning, others have devised a pattern-detecting algorithm, PATTERNA, to mine RNA motifs in structure profiling experimental data sets. This is a model inspired by speech recognition patterns, and it does not rely on energy minimization calculations or known references, but instead trains itself to learn the parameters of the model directly from the data. Ultimately, PATTERNA can be used to accurately detect structural motifs in a diverse range of datasets, which can then be more carefully analyzed by a combination of traditional modeling methods. This can facilitate the discovery of structural elements of functional importance[10].

A model that also uses NMR data is iFoldNMR, a discrete molecular dynamics model that integrates atomic topological constraints by interpreting imino chemical shift data in order to efficiently and accurately predict complex tertiary RNA motifs such as pseudoknots and non-canonical base pairing. With imino-based NMR data alone, the developers were able to predict complex tertiary structures of RNAs of up to 56 nucleotides efficiently and accurately. The group anticipates extending the capabilities of iFoldNMR to longer RNAs in the future by working on experimental design and data analysis[11] Overall, our research indicates correlations between chemical shift data and base stacking. Through further optimization of our method and combinations with other techniques such as iFoldNMR, the complete tertiary structure of RNAs and other macromolecules can be computationally elucidated.

Methods

Structure and Chemical Shift Dataset

For 104 RNAs, NMR chemical shifts, stacking classification, and other structure metric data were downloaded from Github. We modified the data by deleting the base pairing status, orientation, sugar puckering, and pseudoknots because these structures introduce information about stacking and would thus release bias into the model. Using one-hot encoding, numbers were assigned to each residue based on the residue type. Chemical shift data from neighboring residues, up to five neighbors on each side of the residue in consideration, was also factored into the model, resulting in different accuracies. The canonical base-stacking status of each residue in the RNA was determined from the LOO MLP model using the optimal number of neighbors, which we determined to be 1.

Multi-Layer Perceptron Classifiers

To predict the base-stacking status of individual residues in an RNA based on the chemical shifts of atoms in those residues, we trained a set of multi-layer perceptron (MLP) classifiers. The input features, hidden neuron layers, and output labels for the optimum keras model can be visualized in Figure 3.

The baseline model that was used to understand how the number of neighbors affected the f1-score of predicting base stacking had three dense layers of 50, 100, and 50 neurons respectively, each with the default activation function for MLPClassifier. The default MLPClassifier loss and optimization functions were used. The code to run this was compiled from Github.

The keras model that achieved highest results consisted of three deep layers and one output layer. Dropout layers of 30% were used between the first three layers to prevent overfitting. Activation functions of both rectified linear unit (ReLU) and sigmoid were used. Adaptive Movement Estimation (ADAM) and binary cross-entropy were utilized for optimization and loss respectively. The model was iterated over 250 epochs. Data was randomly split and trained on 70% and tested on 30%. F1 score was used to evaluate the model. The keras model can be found at Github

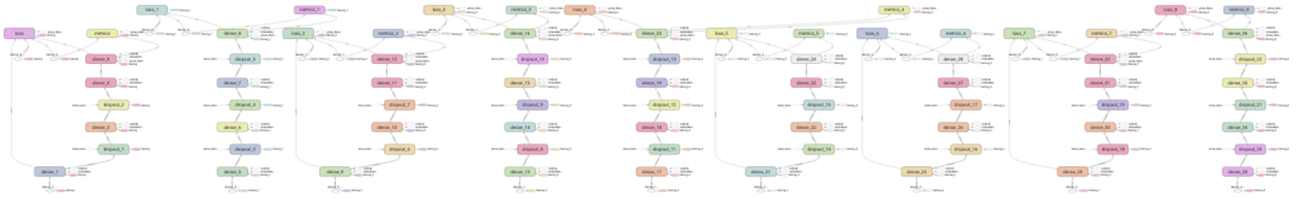


Figure 3. Tensorboard visualization of MLP Classifier input features, hidden layers, and output labels

RDS

RDS, or Random Data Segmentation, is an alternate method to train and validate our machine learning models. Instead of removing just one RNA for testing and the rest for training, this approach entails taking a random subset of the RNAs and setting them aside for training, and setting the rest aside for testing. In general, we kept the training set larger than the test set in order to create a robust model. RDS has the benefit of preventing overfitting to the dataset, which helps improve validation scores of models. The code to integrate RDS into the neural networks is specified in the RDS MLP file mentioned previously.

SASA Model

The solvent-accessible surface areas data for the same 104 RNA residues is reported in Github. Due to discrepancies between the number of data points in the SASA data and the chemical shift data for five of the RNAs, we removed these five RNAs. Our model for predicting solvent accessibility for RNA residues was distinctly different from base stacking in the sense that it was a regression model, which must predict a continuous distribution rather than discrete classification. Furthermore, there were five domains which had to be predicted; side-chain, main-chain, non-polar, all-polar, all atoms. A multiclass regressor was designed with 5 outputs to predict all variables at once.

This model was architected with three layers of 100 neurons each. Between these layers, a random dropout of 40% was implemented to prevent overfitting. RELU activation function was used in intermediate 'deep' layers. ADAM optimizer was chosen because of its efficiency and tendency to converge. Mean squared error was chosen as the loss function, however our preliminary experiments showed little variation in performance with other loss functions. The SASA model was compiled in the same code alongside the RDS MLP algorithm.

1-Dimensional Graph Convolution

The 1-Dimensional graph convolution was created using a keras sequential model. In it, there was, in order, an input layer, one Conv1D layer, followed by two dense layers and a final output layer. The number of Conv1D layers was limited to one because the vector was only 57 units long, and having more than 1 convolution layer could cause overfitting and limit useful pattern detection by the algorithm. All the layers of the network had a ReLU activation, except the output which used softmax. The model was compiled using a binary cross entropy loss function and an ADAM optimizer. The code for running this model can be found on Github.

References

1. Condon, D. E. *et al.* Stacking in RNA : NMR of four tetramers benchmark molecular dynamics. *J. Chem. Theory Comput.* **11**, 2729–2742, DOI: https://doi.org/10.1021/ct501025q (2015).

2. Zhang, H. *et al.* A new method of rna secondary structure prediction based on convolutional neural network and dynamic programming. *Front. Genet.* **10**, DOI: https://doi.org/10.3389/fgene.2019.00467 (2019).

3. Zhang, K. & Frank, A. T. Conditional prediction of RNA secondary structure using NMR chemical shifts. *BioRxiv preprint* DOI: https://doi.org/10.1101/554931 (2019).

4. Ramsundar, B., Eastman, P., Walters, P. & Pande, V. *Deep Learning for The Life Sciences* (O'Reilly Media, 2019).

5. Paillart, J.-C., Westhof, E., Ehresmann, C., Ehresmann, B. & Marquet, R. Non-canonical interactions in a kissing loop complex: the dimerization initiation site of HIV-1 genomic RNA. *J. Mol. Biol.* **270 (1)**, 36–49, DOI: https://doi.org/10.1006/jmbi.1997.1096 (1997).
6. Wilson, T. J. & Lilley, D. M. RNA catalysis–is that it? *RNA* **21(4)**, 534–537, DOI: 10.1261/rna.049874.115 (2015).
7. Lee, B. & Richards, F. The interpretation of protein structures: Estimation of static accessibility. *J. Mol. Biol.* **55(3)**, 379–400, DOI: https://doi.org/10.1016/0022-2836(71)90324-X (1971).
8. Yesselman, J. D. *et al.* Rna tertiary structure energetics predicted by an ensemble model of the rna double helix. *BioRxiv preprint* DOI: https://doi.org/10.1101/341107 (2018).
9. Magnus, M., Kappel, K., Das, R. & Bujnicki, J. M. Rna 3d structure prediction guided by independent folding of homologous sequences. *BMC Bioinforma.* **20(512)**, DOI: https://doi.org/10.1186/s12859-019-3120-y (2019).
10. Ledda, M. & Aviran, S. Patterna: transcriptome-wide search for functional rna elements via structural data signatures. *Genome Biol.* **19(28)**, DOI: https://doi.org/10.1186/s13059-018-1399-z (2018).
11. Benfeard, W. I. *et al.* Structure modeling of rna using sparse nmr constraints. *Rucleic Acids Res.* **45(22)**, DOI: 10.1093/nar/gkx1058 (2017).

Acknowledgements

We thank Dr. Frank for his useful suggestions to the project.

Author contributions statement

All authors prepared and ran Python scripts. All authors reviewed the manuscript.