Biophys 435 Final Project

Cameron Craig[1,+], Isaac Payne[2,+], and Itai Palmon[1,*,+]

[1]Program in Biophysics
[2]Program in Biochemistry
[*]ipalmon@umich.edu
[+]these authors contributed equally to this work

What is the biophysical challenge being added and why is it important?

The Solvent Accessible Surface Area (SASA) of a given molecule was first defined by Lee and Richards to describe protein accessibility with respect to an implicit or explicit solvent, or a given binding molecule (Richards 1971). In particular, the binding between RNA and proteins is important for its role in the regulation of gene expression. SASA analysis of RNA and RNA-binding proteins is demonstrably useful for predicting RNA-protein binding, by the ability to determine free-energy of binding through the SASA of both macromolecules in co-crystallized structures (Bahadur 2018).

What is the state of the art?

Our project seeks to create a model for predicting the SASA of RNA molecules via deep learning. It is state of the art for its use of a multi-layer perceptron regressor to correlate SASA with the sub-features of RNA molecules.

How did you train your baseline and deep neural model?

We trained our baseline model via a leave-one-out (LOO) cross-validation algorithm. The deep neural model was constructed using a Multi-Layer Perceptron Regressor, using “adam,” a stochastic gradient descent algorithm to optimize the predictor. We found the best results with the sigmoid or tanh activation functions for the hidden layer, which respectively use the functions shown below, 1 and 2.

$$f(x) = \frac{1}{1+e^{-x}} \tag{1}$$

$$f(x) = \tanh(x) \tag{2}$$

How does your model compare to baseline models?

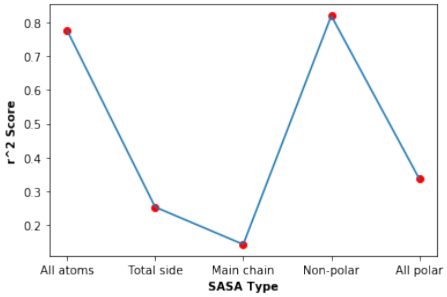

Figure 1. There’s a wide range of R2 values that the model predicted for the SASA chemical shifts, ranging from around 0.1 to 0.8. These results mean the chemical shifts allow the model to predict the SASA value in some cases, but not in others.

How can your model be improved?

Obviously, a majority of the LOO models had low accuracies in terms of R2, ranging from about .5 to .8. Since R2 implies how much of the variation in data is explained by the independent variable changing, in this case, the chemical shifts. Therefore, it is likely that there is some other factor in determining if a surface area is solvent accessible. With this data, we might be able to construct a better regressor. Additionally, some other featurizations of the RNA structure may explain poor accuracy of the model. Including the shape of the RNA molecules may allow for the model to determine if a surface area is solvent accessible, as the bending of an RNA molecule may affect the surface area.

References

1. Lee, B. & Richards, F. The interpretation of protein structures: Estimation of static accessibility. *J. Mol. Biol.* **55**, 379–400, DOI: https://doi.org/10.1016/0022-2836(71)90324-X (1971).

2. Mukherjee, S. & Bahadur, R. An account of solvent accessibility in protein-rna recognition. *Sci. Reports* **8**, 10546, DOI: https://doi.org/10.1038/s41598-018-28373-2 (2018).

12