Project Title

Semi-supervised learning-based prediction of RNA-protein sequence and structure binding preference

Team Members (including Andrew IDs)

Yuxuan Wu (yuxuanwu)   Yifan Wu (yifanwu3)

Project Idea

Our project would design a multi-instance learning based model with gated attention mechanism to make a binary prediction of the RNA protein binding preference. We would use iDeepS paper as our baseline model, which used CNN+LSTM model to include both sequence and structure information of the model.

Data Resources

We would obtain our data from the iDeepS [1] repository in github (https://github.com/xypan1232/iDeepS/tree/master/datasets/clip). We might use one typical CLIP-seq dataset with representative RBPs Ago2. The dataset is pre-cleaned and processed in fasta file format, with labels indicated as either 1 (positive) 0 (negative) and each sequence length is 101 bp. They also separated the training and testing sample, which is safe to use directly. This Github repository has over 30 different dataset, and each dataset contains 30,000 training samples and 10000 test samples. And we believe this is sufficient data for our model.

Feature Engineering

Given the sequence data (101bp) and binary label achieved from the data, we are planning to use a sliding window with fixed size (tunable) and stride (tunable) to obtain the instance level feature. After that, we would use one hot encoding to convert each single nucleotide to a 1*4 vector.

Models Design

We would incorporate the idea of multi-instance learning and gated attention mechanisms together to extract significant information from limited data. Multi-instance learning, an algorithm originally applied in image classification, where the labels are available at image level rather than object level, is opted to work on the current RNA binding protein preference prediction [2]. As the specific binding site/motif is unknown. It might be helpful to use attention mechanisms to identify the patterns with higher weights in the classification.

Evaluation Methods

We would like to use five parameters to evaluate the performance of our model, which are model accuracy, recall score, precision score, F1 score, the Matthews correlation coefficient (MCC), area under the ROC Curve (AUC), area under the precision-recall curve (auPRC). We have two benchmarks; one is from the DeepBind [3] (CNN) and the other is from iDeepS (CNN+LSTM).

Potential Improvements

In iDeepS, they utilized the existing package to generate the RNA structure data in model training, however, our primary focus of this project is to test whether semi-supervised learning could increase the overall performance, if time allows, we would evaluate the performance with structure information included.

Reference

[1] X. Pan, P. Rijnbeek, J. Yan, and H. Bin Shen, “Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks,” *BMC Genomics*, vol. 19, no. 1, pp. 1–11, 2018, doi: 10.1186/s12864-018-4889-1.

[2] D. Huang, B. Song, J. Wei, J. Su, F. Coenen, and J. Meng, “Weakly supervised learning of RNA modifications from low-resolution epitranscriptome data,” *Bioinformatics*, vol. 37, pp. I222–I230, 2021, doi: 10.1093/bioinformatics/btab278.

[3] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, “Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning,” *Nat. Biotechnol.*, vol. 33, no. 8, pp. 831–838, Aug. 2015, doi: 10.1038/nbt.3300.