iNAP: NUCLEIC ACID BINDING CLASSIFICATION FROM A PROTEIN SEQUENCE USING BiLSTM-CNN

Stella Park, Jungmin Kim
Columbia University
New York, NY 10027
COMS4762 Machine Learning for Functional Genomics

Abstract

Proteins that bind to the nucleic acid molecules serve important function in human cells including the control of transcription, translation, splicing, apoptosis, and DNA repair. Proteins that bind to the nucleic acid sequences can be classified depending on what they bind to: RNA binding protein (RBP), DNA binding protein (DBP), or DNA RNA binding protein (DRBP). Changes in nucleic acid sequence of these proteins may have altered affinity for the binding or they may lose the binding ability. Moreover, the question of identifying if newly identified protein sequence produces RBP, DBP, or DRBP is a focus of many ongoing research since these proteins serve key roles in live cells. There are existing machine learning models that takes up the DNA sequence and aim to classify the protein into RBP, DBP, or non-nucleic acid binding protein. However, most models do not consider the fact that some proteins may bind to both DNA and RNA. Here, we developed iNAP, a model that takes up the amino acid sequence and classify the protein into one of the four categories: RBP, DBP, DRBP, or non-nucleic acid binding protein. iNAP utilizes Convolutional Neural Networks (CNN) to extract the local features of the sequence and Long Short-Term Memory (LSTM) network to effectively preserve the contextual information from the long protein sequences that determines the nucleic acid binding characteristics. iNAP uses bidirectional LSTM (BiLSTM) to consider the context from forward and backward direction, since the overall protein characteristics are determined from the complete structure of the whole protein. We compared iNAP to the performance of other models that predict if the protein is DBP or not. iNAP outperformed existing models with a high Area under the ROC Curve (AUC) score of 0.922 and a fair accuracy of 74.65%. Among the compared models, iNAP was the only model that utilized BiLSTM. Overall result shows that BiLSTM is an effective approach in working with the protein sequences since the biological functions of the proteins are scribed innately into the context of the complete sequence.

1 Introduction

1.1 Biological Motivation

Interactions between protein - RNA, protein - DNA, and protein - RNA/DNA serve an important role in control of transcription, translation, DNA repair, splicing, apoptosis, etc. Noticeable number of proteins are known to bind to both DNA and RNA(DRBP), while a large portion of the proteins are either DNA binding protein(DBP) or RNA binding protein(RBP). Even though they serve very different biological functions, it is hard to predict which type of nucleic acid the protein would bind

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

to, given novel protein sequences. Many past methods have evaluated the RNA/DNA affinity of R/DBPs, with minor consideration of a possible affinity to both RNA and DNA. However, they often confuse if the given protein is RBP, DBP, or DRBP. This would be largely due to the fact that DNA binding prediction was only done with DBP, while RNA binding prediction was only done with RBP. Since alternative splicing events can affect the binding affinity of a given protein and alter the characteristics of D/RBP, we developed a multi-class classification model that differentiates the types of protein based on which nucleic acid it binds to, given protein sequences.

1.2 Project Goal

Existing Approaches Yan, et al.[5] is a pre-existing study in this area which used HMM and logistic regression to predict RBP and DBP. The study utilized HMM to assess the features of given protein sequences and extract the feature score for each amino acid residue for binding affinity in each case (RNA or DNA). This information was then fed into logistic regression and the spatial proximity between high score residues were calculated. The method used in Yan, et al. has limitations that it does not consider the innately determined protein sequence structure, since binding properties are largely dependent on the region of protein when folded.

Model Design We developed iNAP (Identification of Nucleic Acid-Protein Interaction), a machine learning model that takes the amino acid sequences as the input and classifies the sequence into one of four categories: RBP, DBP, DRBP, or non-nucleic acid binding protein. Amino acid sequences used to produce proteins in human cells contain highly contextual information; each amino acid seems independent, but the function of the proteins largely depends on the overall structure of the sequence. Thus, we chose to implement the Recurrent Neural Network (RNN) to model the sequence of data. Specifically, we implemented Long Short-Term Memory (LSTM) networks to resolve the vanishing gradient problem. LSTM allowed us to effectively preserve the characteristics of historical information in long sequences. Here, we chose bidirectional LSTM (BiLSTM) to consider the sequence context in both forward and backward directions.

2 Data

We downloaded the protein sequences from Uniref100 [3]. We only filtered for Homo Sapiens proteins those functions are known. Due to the training time issue and data size, we randomly selected a subset of the protein sequences that are composed of 28.7% RBP, 18% DBP, 8.7% DRBP, and 44.6% non-nucleic acid binding proteins. To categorize the proteins, we used the following criteria.

- DNA Binding Proteins: We used UniProt annotations to label DNA Binding Proteins [1]. Among DBPs, a random subset was included in our dataset.
- RNA Binding Proteins: We used the annotations generated by Gerstberger et al. [2]. Among the listed RBPs, a random subset was included in our dataset.
- DRBP: We used the DRBP annotation generated by Leung at al. [4]. Among the listed DRBPs, a random subset was included in our dataset. We filtered out any proteins that were also reported as RBP only or DBP only.
- Non-binding proteins: To strictly categorize the protein as non-nucleic acid binding protein, we only selected the proteins from Uniprot that were validated to be non-binding experimentally. Also we excluded any proteins that had DNA, RNA, nucleic acid, nucleotide, or ribosomal in their names. Function annotation with DNA binding, RNA binding, nucleic acid binding, or nucleotide binding function was also filtered. We followed the guideline by Yan et al. [5]

The data was extracted in FASTA format. A total of 5897 protein sequences were used. We split the complete dataset into 80% training set, 10% validation set, 10% test set. The summary of the dataset statistics is in Figure 1. All of the rare amino acids outside of 20 common amino acids included in the original dataset were substituted to "X," to indicate that the model is not considering it for the prediction.

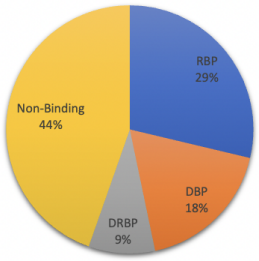

Figure 1: Dataset Composition

3 Methods

3.1 Sequence length

We used one-hot encoding to tokenize the sequence and the labels. Since we needed the input sequences to be of a uniform length, we had to find an appropriate length to truncate the sequence. We used histograms to visualize the sequence length. The length of the longest input sequence was 7388, however as seen in Figure 2, the histograms show that most sequences fall below 1000 in length. We chose to truncate/pad the input sequence to 256, since despite the short length, 55.0% of training data, 42.7% of validation data and 48.8% of testing data had lengths smaller than 256.

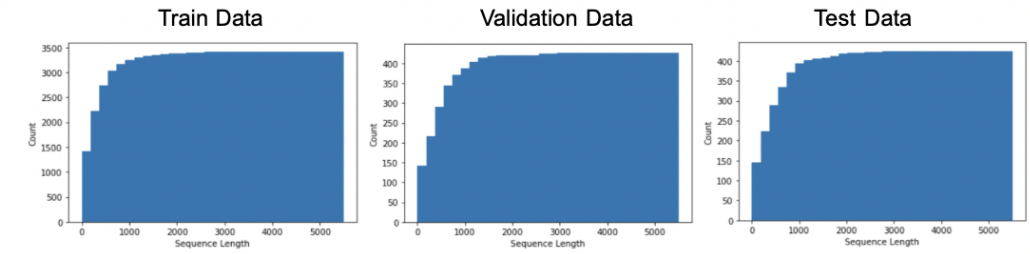

Figure 2: Cumulative count for each sequence length in train, validation, and test data

3.2 Model Structure

We built a BiLSTM - CNN model, which combined both the CNN's pattern recognition abilities and BiLSTM's recurrent memory and long-distance dependency capturing ability. A total of 314,204 parameters were trained. First, the amino acid sequences went through the embedding layer with an embedding vector length of 100, followed by a dropout layer. Next, the sequence went through a series of convolution and pooling layers to extract local features of the sequence. We implemented two convolution layers with Rectified Linear Unit (ReLu) activation function. The first convolutional layer had a kernel size of 10, and the second convolutional layer had a kernel size of 5. The max-pooling layer that follows the convolutional layers had a pool size of 2. After another dropout layer, it went through the Bidirectional LSTM layer with a recurrent dropout rate of 0.5. It was then activated with the softmax function that classifies the data. Dropout layers of 50% rate were implemented to prevent overfitting the model to the training dataset. The model structure is summarized in Figure 3.

4 Results

4.1 Performance Evaluation

We trained the model for 50 epochs before the evaluation. Figure 4 shows the model accuracy on our training and validation data set across the epochs. As seen, we observed that iNAP accuracy on the training set kept improving as epochs increased. However, the marginal benefit we

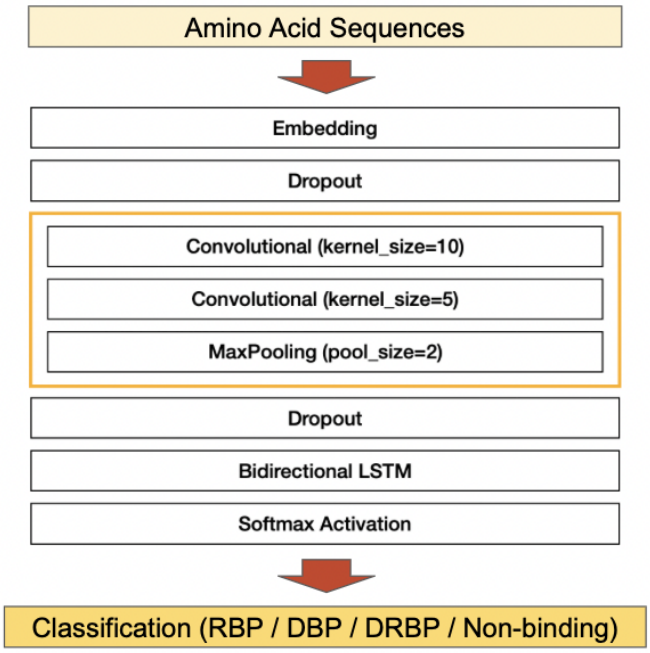

Figure 3: Model Structure of iNAP

can get from increasing the number of epochs from here seemed small since the amount of time consumed significantly increased to train the model and the validation accuracy was already converging. To prevent overfitting too much, we limited the number of epochs to 50. After training the model and evaluating with the test data set, iNAP reached an accuracy of 74.65% and the Area under the ROC Curve (AUC) of 0.9222. iNAP is available for the readers in the following link: "https://drive.google.com/drive/folders/1iAKv4IavkxOjRReyM12hh57qp31FmNxR?usp=sharing"

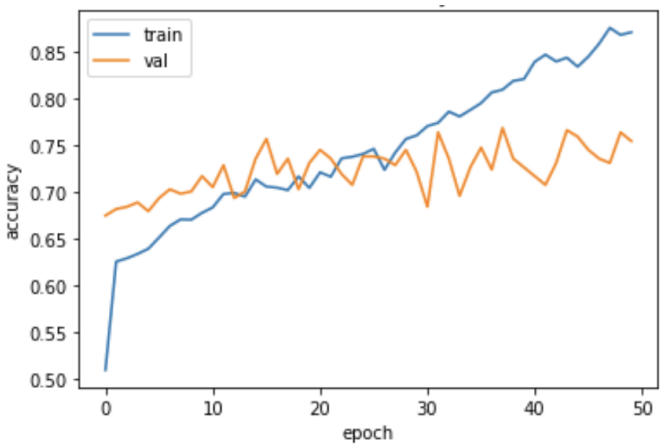

Figure 4: Model accuracy on the training and validation data set across the epochs

4.2 Comparison

There are existing studies that focus on identifying a DNA-binding protein from its sequence. We compared iNAP to existing DNA-binding protein prediction models since we could not find an existing model that also classifies proteins into the four categories we have determined. The summary is shown in Figure 5. Among the existing studies, HMMBinder[6] showed the best performance with

an accuracy of 86.33% and an AUC of 0.9026. The approach of HMMBinder is largely different from iNAp; HMMBinder used the hidden Markov model (HMM) profiles to extract monogram and bigram features of the protein sequences and utilized Support Vector Machines (SVM) to classify the proteins.

Compared to the existing models, iNAP had a high AUC that outperformed all listed models. While iNAP showed an exceptional AUC, its accuracy was fair enough but around the accuracies of other models, not necessarily superior to others. We believe the reason to be one of two. First, this can happen if our model achieved a good performance in predicting the positive class, at the cost of a high false negatives rate. Thus, the prediction of the positive class given by iNAP is likely to be reliable, while its prediction of the negative class may be too harsh. Second, this may have happened if the threshold we used for the AUC evaluation was not suited to our model, even though our model performance was good in reality. Nevertheless, since our accuracy was comparable to other models and our AUC outperformed all of them, after trials of improvements, we decided to conclude with what we currently have.

Method	Accuracy	AUC
iDNA Pro-PseAAC	76.76%	0.8392
DNA Binder (dimension 21)	73.95%	0.8140
DNA Binder (dimension 400)	73.58%	0.8150
DNA-Prot	72.55%	0.7890
iDNA-Prot	75.40%	0.7610
iDNA-Prot|dis	77.30%	0.8310
PseDNA-Pro	76.55%	-
Kmer1 + ACC	75.23%	0.8280
Local-DPP	79.20%	-
HMMBinder	86.33%	0.9026
iNAP	**74.65%**	**0.9222**

Figure 5: Model Performance Comparison of iNAP with other DNA-Binding Protein Prediction models [6]

5 Discussion

The focus of this model development was to classify the proteins by reading the innate language of the protein sequence - to demystify the protein characteristics that are hidden under the sequence of amino acids. Competitive performance of iNAP shows that BiLSTM-CNN model effectively captures the characteristics of the protein sequences and well performs the classification task.

One discussion point is determining the length of the amino acid sequence that is used for the model training. The length of the sequences that were in the dataset we gathered varied from as short as 14 amino acids to as long as 7388 amino acids long. We chose to truncate or pad it to 256 amino acids, and this allowed 53% of the total data to be expressed fully. We have also tried with the length of 1024, and this allowed 93% of the data to be fully expressed. However, contrast to our expectation that choosing longer length would improve the model, while longer length slowed down the training time significantly, the performance was not any better. This was something that we did not expect during training, and we would like to explore the question of finding the optimal sequence length further. Moreover, we envision the future work to include tailoring iNAP so that it can also return the nucleic acid binding affinity score along with the classification or predicting the specific binding domain.

References

[1] The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1):D480–D489, 11 2020.

[2] Stefanie Gerstberger, Markus Hafner, and Thomas Tuschl. A census of human rna-binding proteins. *Nature reviews. Genetics*, 15(12):829—845, December 2014.

[3] Baris Suzek, Hongzhan Huang, Peter Mcgarvey, Raja Mazumder, and Cathy Wu. Uniref: Comprehensive and non-redundant uniprot reference clusters. *Bioinformatics (Oxford, England)*, 23:1282–8, 06 2007.

[4] Ricky Wai Tak Leung, Xiaosen Jiang, Ka Hou Chu, and Jing Qin. ENPD - A Database of Eukaryotic Nucleic Acid Binding Proteins: Linking Gene Regulations to Proteins. *Nucleic Acids Research*, 47(D1):D322–D329, 11 2018.

[5] Jing Yan and Lukasz Kurgan. DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues. *Nucleic Acids Research*, 45(10):e84–e84, 01 2017.

[6] Rianon Zaman, Shahana Yasmin Chowdhury, Mahmood A Rashid, Alok Sharma, Abdollah Dehzangi, and Swakkhar Shatabda. HMMBinder: DNA-binding protein prediction using HMM profile based features. *Biomed Res. Int.*, 2017:4590609, November 2017.