Coding Task: Transcription Factor Binding Predictor

Mahmoud Hossam

August 11, 2021

1 Introduction

In this report I share my findings on building and training a deep learning classifier to predict the binding probability of a certain Transcription Factor (TF) to DNA sequences. All the code and detailed information of the model and the experiments are in this Google Drive folder. In the following sections I discuss the results, the motif interpretation findings, and future work.

2 Results and Discussion

2.1 Effect of creating the negative samples dataset

Using CNN architecure to build the TF prediction classifier, several experiments were conducted under different architectures and hyperparameters. The main finding is that the model architecture (convolutional or recurrent) along with the hyperparameters (regularization, optimizer, number of neurons, etc..) only improve the model marginally. However, the technique used for generating the negative samples (that do not bind to the TF) has a significant effect. Table 1 shows the effect of the used negative sample generation technique on the accuracy and area under the ROC curve (AUC).

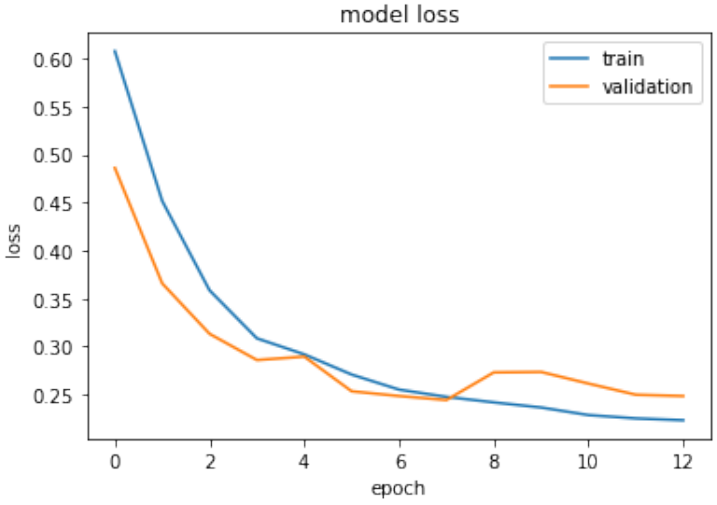

Figure 1: Training and validation losses.

It can be easily noticed that the more the generated negative samples are random or lose their original structure due to heavy shuffling, the better the classifier performs. However, it is important to choose techniques that allow some structure in the created samples in order to generalize better to other DNA sequences in other domains. Therefore, a mixture of the three methods is found useful in this regard and performs reasonably well. Furthermore, K-folds model selection was used to select the best model. The validation loss follows the training loss in decreasing manner (Figure 1), and the training is stopped early before validation loss diverges from the training loss. These design choices might further help to generalize of the model on held-out and different DNA test data.

Table 1: Test Accuracy and AUC for the model under different negative samples generation techniques

	Random bases flipping	N-Grams Synthesis	Positions Shuffling	Test Data Accuracy	AUC
Classifier	Ratio of seq. length	(N)		%	
CNN	0.1	–	–	<70.0	< 0.70
	> 0.3	–	–	>99.0	>0.99
	–	3, 5, 10	–	>99.0	>0.99
	–	–	yes	>99.0	>0.99
Logistic Regression	0.25	yes	yes	68.5	0.745
CNN	0.25	yes	yes	**91.4**	**0.954**

2.2 Inferences on binding preferences of TF

By using a simple method of computing importance scores for DNA nucleobases through deleting them consecutively, we could uncover insights regarding what motifs are related to the binding behavior of the TF. The results in Figure 2 show high importance scores for the motif CCGCA appearing in the positively classified DNA sequence. There are some additional motifs that also appear to have some effect on the TF binding, like GTTT, GTTG and GTTA, and they usually precede the main CCGCA motif by few nucleobases.

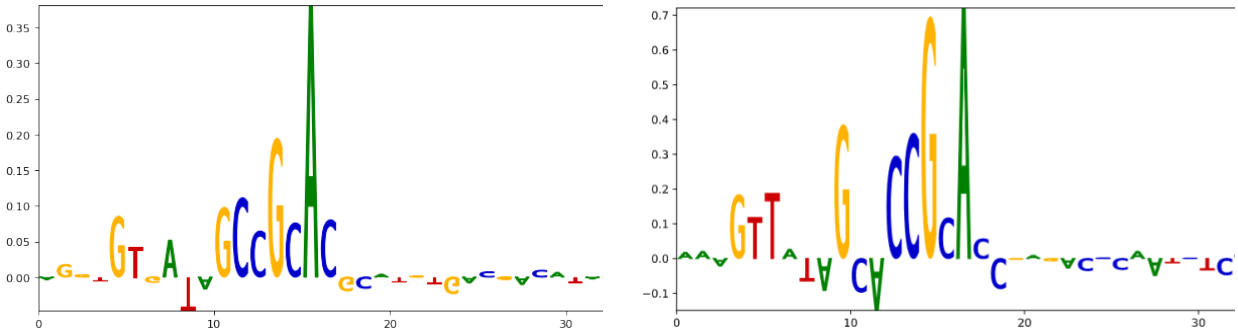

Figure 2: Visualization of importance scores towards prediction probability, showing the discovered *CCGCA* motif.

3 Conclusions and Future Work

Through a set of experiments on CNN TF classifier, we show that the choice of negative samples generation is key to better model performance and generalization. Additionally, There is plenty of room for future improvements:

1. **Negative samples generation**: There are more methods that can be considered that might allow better generalization, including: *i)* Training a model to generate DNA-like sequences in an unsupervised way, and use it to create negative samples, and *ii)* use DNA sequences from other sources that are independent of the transcription factor TF to use as negative samples.

2. **Nucleobase importance scores**: More advanced methods could be considered: *i)* Generating saliency maps using the gradient of the model's prediction with respect to each individual nucleobase. *ii)* Employ an interpretable model like "Learning to explain (L2X)", where importance scores can be learned during the training, and later be used to identify important motifs and TF binding preferences.