# CS282R PROJECT PROPOSAL

YAKIR RESHEF AND FULTON WANG

## 1. The problem

Deep learning has led to major advances in our ability to predict the biological function of DNA sequences. One specific task that has seen a great deal of progress has been the task of predicting, given a short (say 1000 base pairs) DNA sequence, whether a given *transcription factor* (TF) will bind to that sequence. A transcription factor is a protein that binds to DNA in specific places and turns genes in those places on/off, typically in a coordinated fashion to achieve some goal such as cell division, response to an immune stimulus etc. Because transcription factors play an important role in regulating biological processes within the cell, there is considerable interest in understanding what determines their behavior.

One specific aspect of TF binding that is important to understand is the effect of *genetic variants* on the binding of TFs. Specifically, if a patient has, say, at A instead of a T some specific location, and we're trying to understand whether that variant is disease-causing, we'd like to be able to say whether it increases or decreases the binding of disease-relevant TFs. One approach to this problem is to query a black-box predictor about the binding of a TF of interest by feeding in two different sequence inputs, once with an A at the location of the variant and one with a T at the location of the variant, and to see which gets a higher prediction and by how much. However, the results of this method come with many caveats due to the black-box nature of the predictor. For example, the "alternate" version of the sequence might not be represented in the distribution the predictor is trained on, such that we don't have guarantees about the generalization capability of the predictor.

## 2. The proposal

We propose to leverage a more mechanistic understanding of TF binding to "constrain" our predictor in order to improve the interpretability of its predictions on sequences that are not well represented in the training set. We aim to do this using the fact that there are databases that contain short DNA sequences called *motifs* that have been observed to be bound by different transcription factors. (Each motif corresponds to one transcription factor.) Rather than training a generic convolutional neural network to predict binding, we plan to train a neural network whose first layer of convolutions is fixed to match these motifs. This way, when we look at the difference in the predictions for two possible values of a genetic variant, we know that the variant must act by changing the strength of one of these biologically relevant motifs.

## 3. Goals for 1 month

The goals fall into the categories of 1. data processing, 2. model building, and 3. model evaluation.

(1) Data Processing: Assemble the data used in [2] by obtaining human reference genome sequences for sites where at least one of 919 chromatin marks (labels) is present (positive class). Assemble a set of motifs from [1], making sure the number of motifs is not too big, and that they are all of uniform size (for reference, [2] uses 320 filters that are 8 base pairs long, whereas [1] contains 402 human transcription factor motifs).
(2) Implement in Tensorflow the neural network classifier. Should be able to specify the hardcoded list of motifs when constructing the model. Do not worry about tuning hyperparameters yet - assume they are all specified in a configuration file.
(3) Evaluate prediction accuracy on sequences from chromosome 8/9.

## References

[1] Ivan V Kulakovskiy, Ilya E Vorontsov, Ivan S Yevshin, Ruslan N Sharipov, Alla D Fedorova, Eugene I Rumynskiy, Yulia A Medvedeva, Arturo Magana-Mora, Vladimir B Bajic, Dmitry A Papatsenko, Fedor A Kolpakov, and Vsevolod J Makeev. Hocomoco: towards a complete collection of transcription factor binding models for human and mouse via large-scale chip-seq analysis. *Nucleic Acids Research*, 46(D1):D252–D259, 2018.

[2] Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods*, 12(10):931, 2015.