Deep Motif: Visualizing Genomic Sequence Classifications

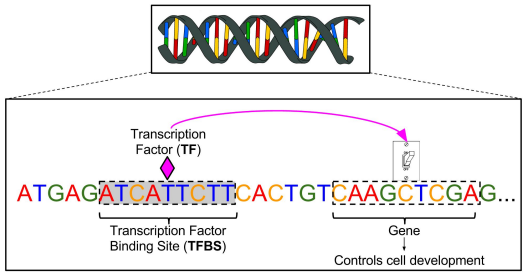
Jack Lanchantin, Ritambhara Singh, Zeming Lin, & Yanjun Qi

University of Virginia, Department of Computer Science

ICLR 2016

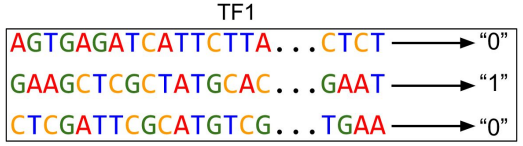
1. Overview

Transcription factors (TFs) are proteins which bind to DNA to regulate cell machinery. Being able to accurately predict and understand the binding sites of transcription factors (TFBSs) will help lead to a better understanding genomics in general. We apply a CNN framework to classify genomic sequences on the TFBS task. To make the model interpretable, we propose an optimization driven strategy to extract "motifs", or symbolic patterns which visualize the positive class learned by the network.

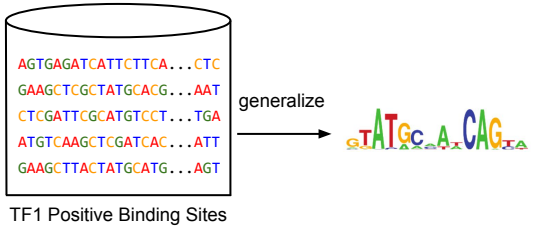


2. Objectives

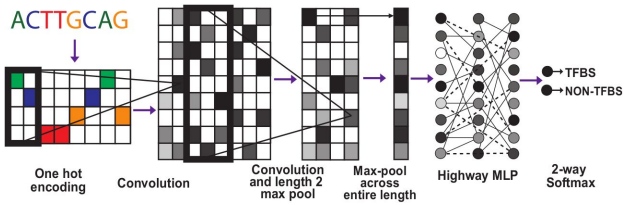
1. Accurately predict where TFs will bind on a DNA sequence (binary classification task) using a convolutional neural network.



2. Understand, or visualize, the locations of the positive binding sites by optimizing the positive binding site class of the trained network.



3. ConvNet for TFBS Classification



The final model hyperparameters were selected based on training set accuracy for each individual TF. Each model has 3 to 4 convolutional layers with 128 hidden units, and 5 to 7 highway MLP layers with 32 hidden units.

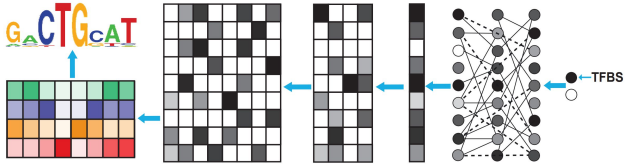
4. Class Optimization/Visualization

After obtaining high accuracy classification results, we seek to find the model's interpretation of the positive TFBS class.

We seek to optimize the following equation, where $P_+(S)$ is the probability of the input sequence S (matrix of *input length* × 4, where 4 is our alphabet size) being a positive TFBS computed by the softmax output of our trained model for a specific TF:

$$\arg\max_S P_+(S) + \lambda \|S\|_2^2$$

We find a locally optimal S through backpropagation, where the optimization is with respect to the input sequence. We clip the optimized values to the interval [0, 1] and convert S into a PWM, using Laplace smoothing.



References

[1] Alipanahi et. al.. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. Nature biotechnology, 2015

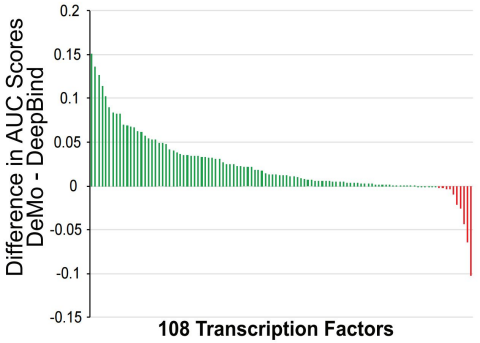
[2] Mathelier et. al. Jaspar 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic acids research, 2015

[3] Simonyan et. al.. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034, 2013

5. Experiments and Results

We ran DeMo on the same 108 leukemia cell TF datasets used in [1]. Each TF dataset has an average of 30,819 training sequence. Due to the separate train/test data for each TF, we train a separate model for each individual TF dataset.

1. DeMo outperforms DeepBind [1] (baseline) on 92 of the 108 datasets. DeMo median AUC: 0.951 DeepBind median AUC: 0.931



2. Of the 57 tested, 36 of our motifs significantly match JASPAR [2] (baseline) motifs, and 29 of our motifs outperform JASPAR motifs using the average motif affinity scoring tool.