Using Deep Learning to Predict Long-Range Regulatory Networks Based on Protein-Protein Interactions

Albert Xue[1] Binbin Huang[2] and Jianrong Wang, PhD[2]

[1]Duke University, Durham, NC 27708 [2]Department of Computational Mathematics, Science & Engineering, Michigan State University, Lansing, MI 48824

Introduction

The vast majority of disease-associated genetic variants are located in non-coding regions, which represent a major portion of the human genome [1, 2]. A systematic delineation of the mechanism by which such non-coding variants induce diseases requires accurate identification of downstream target genes whose expression levels are regulated by these variants in diverse tissues [1, 2]. Since these target genes are highly tissue-specific and are usually located far away in the 1D genome, prediction has thus far proven to be difficult. Based on our preliminary analysis, we propose to leverage protein-protein interactions (PPI) as features to predict long-range chromatin interactions. We will integrate PPI, transcription factor (TF) binding, chromatin and epigenetic signals in order to train a convolutional neural network to achieve better accuracy on long-range regulation predictions.

Background

In the one-dimensional representation of the genome, an enhancer (regulatory element) and its linked promoter can often be millions of base pairs away (Fig. 1). Such distance makes it difficult for current biological methods, which rely heavily on proximity, to predict their linkages.

Figure 1: An enhancer (blue) and its linked promoter (orange) and gene (green) are often millions of base pairs away in the human genome.

However, in the three dimensional representation of the genome, the paired enhancer and promoter become spatially proximal in order for mediating transcription factor complexes to bind to both sites, bending the chromatin in the process (Fig. 2).

Figure 2: The presence of a two-protein complex (pink, red) causes the genome to bend in 3D. As a result, a previously distal enhancer and promoter pair becomes spatially proximal.

We hope to build a deep learning model which uses the protein-protein interactions between the transcription factors bound to each enhancer and promoter to indirectly predict whether or not a given enhancer-promoter pair is linked (Fig. 3).

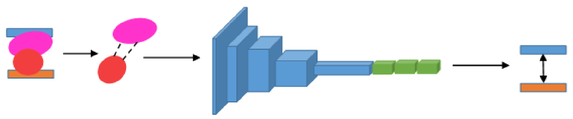

Figure 3: We use protein-protein interactions between transcription factors on a given enhancer and promoter as features for a convolutional neural network, which attempts to determine whether or not the pair are linked

In summary, we perform binary classification on a set of protein-protein interactions for linked and unlinked enhancer-promoter pairs.

Methods and Materials

Encoding protein-protein interactions in the form of an image allows their use as features in a convolutional neural network.

We create a set of negative (unlinked) enhancer-promoter pairs controlled by the distance distribution of an original linked set predicted from the Roadmap Epigenomics Consortium [3]. Furthermore, we integrate known TF interactions from [4] as well as T-cell specific TF expression data from [3] to build a one-hot encoded protein interaction matrix (Fig. 4) which serves as a template for the features of the network.

For each enhancer-promoter pair in the total set, we use binding data from [5] to filter a copy of the original template matrix (Fig. 5). As a final product, for a position *i, j* represented by cluster *i* and cluster *j*, we give

$$[i, j] = i \text{ binds to enhancer} \wedge i \text{ expressed} \wedge j \text{ expressed} \wedge j \text{ binds to promoter}$$

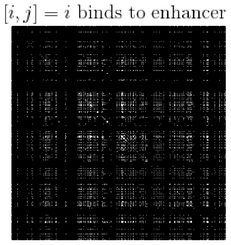

Figure 4: The (399, 399) template matrix, which encodes interactions between transcription factors.

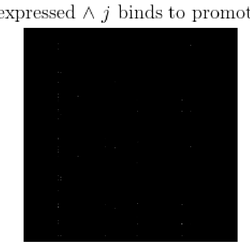

Figure 5: A sample filtered matrix, which is noticeably sparser than the original.

Model

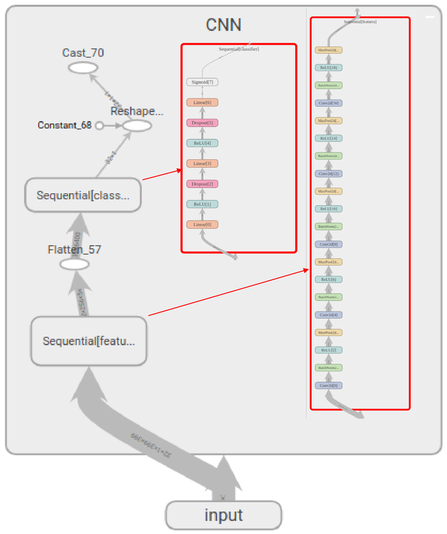

Figure 6: Model implemented in PyTorch [6] and visualized through a PyTorch implementation of TensorBoard [6, 7, 8]

Results and Conclusion

The model was trained for almost 350 epochs using the SWATS optimizer detailed by [9], 60% dropout probability in the classifier, and a slow learning rate of 10^{-4}. Over the training set, the network achieved a loss of 0.1767 and a prediction accuracy of 91.80% (Fig. 7, 8).

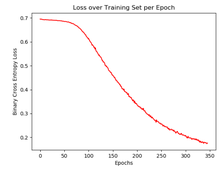

Figure 7: The network's loss decreased steadily down to 0.1767

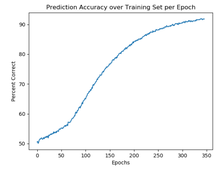

Figure 68 The network's prediction accuracy increased steadily up to 91.80%

However, the model performed poorly over the validation set, achieving a prediction accuracy of only approximately 54%. This suggests a high degree of model overfitting, which may be due to the sparsity of our dataset and the limited number of data points available.

In the future, implementing regularization methods such as even higher dropout rates, weight decay/training schedulers, or other methods may improve network performance. In addition, using transcription factor clustering information from [5] will drastically reduce data sparsity.

Future Work

Future work consists mainly of feature extraction. Determining the most important protein-protein interactions for enhancer-promoter linkage classification can lead to further studies on their mechanistic effects on chromatin folding.

One proposed solution is image occlusion, first detailed by Zeiler and Fergus in [10], where an 'empty' gray filter is used to obscure a certain area of the image at a time. By sliding the filter across the image and observing the resultant effects on classification accuracy, the most important protein-protein interactions for linkage classification can be determined.

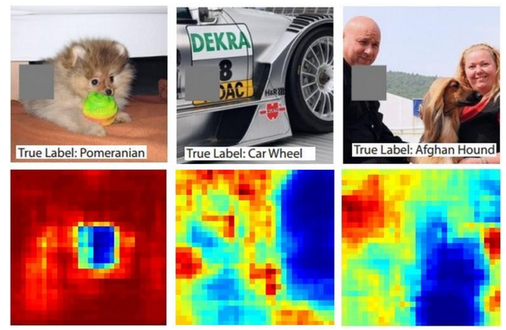

Figure 9: By constructing a heatmap of classification accuracy when a certain part of the image is occluded, we can identify the most important parts of the image for classification (blue)

Acknowledgements

We acknowledge support from the MSU ACRES REU program, which is supported by the National Science Foundation through grant ACI-1560168.

References

1. Whalen S, Truty RM, Pollard KS. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nature genetics*. 2016;48(5):488-496. doi:10.1038/ng.3539.
2. Zeng W, Wu M, Jiang R. Prediction of enhancer-promoter interactions via natural language processing. *BMC Genomics*. 2018;19(Suppl 2):84. doi:10.1186/s12864-018-4459-6.
3. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, et al. Integrative analysis of 111 reference human epigenomes. *Nature*. 2015;518(7539):317-330. doi:10.1038/nature14248.
4. Li T, Wernersson R, Hansen RB, et al. A scored human protein–protein interaction network to catalyze genomic interpretation. *Nature methods*. 2017;14(1):61-64. doi:10.1038/nmeth.4083.
5. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74. doi:10.1038/nature15393.
6. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in pytorch. 2017.
7. https://github.com/lanpa/tensorboardX
8. Martín Abadi, Paul Barham, Jianmin Chen, et al. 2016. TensorFlow: A system for large-scale machine learning. In OSDI
9. Shirish Keskar, Nitish & Socher, Richard. (2017). Improving Generalization Performance by Switching from Adam to SGD.
10. Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks. CoRR, abs/1311.2901, 2013. Published in Proc. ECCV, 2014.