Deep Learning to Evaluate the Combinatorial Impact of Genetic Variants on Gene Expression

Wang, Collin (School: Detroit Country Day School)

DNA-binding proteins play an especially important role in regulating gene expression. These proteins called transcription factors (TFs) recognize and bind to DNA-sequences called DNA motifs, allowing for the control of timing and location of the transcription of DNA. This type of interaction is critical to regulating gene expression. However, current methods for analyzing coding mutations that affect TFs are often computationally inefficient or require large population samples, rendering them unable to detect the impact of rare mutations on gene expression. This research proposes a runtime efficient model that uses recent advances in deep learning like convolutional neural networks (CNNs) and the U-Net, a CNN architecture popular in biomedical image segmentation, to predict the effects of protein-coding mutations on transcriptional regulation. By leveraging publicly available protein-DNA co-crystal structures in the DNAProDB database, the U-Net learns to annotate DNA-binding protein residues directly from protein sequence using few training examples.   Using this approach, a generalizable model that predicts the impact of diverse coding genetic variants on transcriptional regulation was built. The trained U-Net can predict the changes to the DNA-binding profile of TFs with high accuracy and computational efficiency relative to other current models that do so. This research thus provides mechanistic insight into the behavior of how coding mutations alter the genomic regulatory machinery, contributing to our understanding of how personalized coding mutations affect disease predisposition.

Awards Won:

Fourth Award of $500