Development of Semi-Supervised Machine Learning Models to Predict Enhancer Regions in Polygenic Developmental Diseases

Srinivasan, Savitha (School: Interlake High School)

Polygenic developmental diseases affecting humans like autism spectrum disorder often have unknown causes and are difficult to cure. Gene expression and regulation is among the most important mechanistic elements to study. Enhancers, short regulatory stretches of non-coding DNA along chromosomes, signi?cantly impact development through controlling how genes are spatially and temporally expressed. Thus, identifying enhancer regions precisely is key to better elucidating the mechanisms of complex developmental disorders but only a few enhancers have been found with in-vivo techniques. Hence computational approaches to find enhancer regions are appealing. In this study, multiple genomic datasets from the VISTA Enhancer and USCS Genome Browsers were integrated to yield 2200 mus musculus (mouse) enhancers, 10% of which are limb enhancers. Features were extracted by calculating RPKM values for gene expression signatures associated with the enhancers. Supervised machine learning models were developed to baseline the performance of multiple classifiers on limb enhancer prediction. The efficacy of five approaches to address dataset imbalance was systematically investigated. The neural network ensemble developed in this study surpasses prevailing precision/recall rates and was further improved with a newly proposed technique to architect ensemble models based on input zones. New candidate limb enhancers were identified using an algorithm developed with the model's predictions. Finally, semi-supervised learning techniques were investigated to gauge their effectiveness in improving model performance with unlabeled data. The self-training approach enables models to be improved as unlabeled enhancer regions are discovered, thereby supplementing in-vivo techniques effectively.

Awards Won:

Second Award of $2,000

Association for the Advancement of Artificial Intelligence: Honorable Mention