Refined Genome-wide Prediction of Transgenerational Epimutations and an Ensemble-based Solution to the Imbalanced Class Problem in Epigenetics

Apostol, Jason (School: Palos Verdes Peninsula High School)

Paternal exposure to environmental toxicants, malnutrition, and stress is associated with increased disease incidence in their children, grandchildren, and great-grandchildren. Recent research suggests exposure-specific DNA methylation is the primary driver of this phenomenon. This project sought to design an in silico model to predict a larger and more accurate set of potential differentially methylated regions (DMRs) associated with exposures that have yet to be investigated in vivo. A novel motif-detection schema, custom R scripts, and RepeatMasker were used to annotate 332 features on 6,954 known DMRs, extracted from twelve in vivo studies, and a simple random sample of 15,000 non-DMRs. An autoencoder was trained on this unbalanced dataset of annotated regions and created the encoded dataset. After normalization, sampling techniques were employed on both unencoded/encoded datasets to create six final datasets. 60 combinations of various architectures and datasets were then trained, hyperparameter tuned, and evaluated. The top-performing ensemble, SMOTE+AdaBoost, classified 100% of 6,954 known DMRs correctly and 99% of 15,000 non-DMRs correctly, a 3% improvement over the previous top-method. It identified 65,248 potential DMRs in a reference genome, and subsequent cluster analysis identified 150 DMR clusters. Of these, 19,691 regions and 37 clusters were previously unidentified. Clinical diagnostics targeting the novel set of DMRs predicted in this study could provide meaningful information to fathers-to-be on their risk of passing on epigenetically inherited disease. Future research utilizing this list of clusters can efficiently investigate other epigenetic phenomena that contribute to transgenerational programming.

Awards Won:
Third Award of $1,000