Predicting Optimal Combinations of Regulatory Transcription Factor Perturbations to Simulate Complex Cell-State Conversion Paths Using Combinatorial Optimization and Machine Learning

Agrawal, Shaurya (School: Adlai E. Stevenson High School)

Cellular reprogramming holds immense promise for novel therapeutics in fields like disease reversion and cellular regeneration. This field aims to control or perturb gene expression to induce specific downstream impacts of cell states and types. However, current research remains limited to basic use cases as large-scale in-vitro experiments are extremely expensive and impractical and in-silico models can only predict simple conversions with limited clinical applicability. Thus, the goal of this project was to create different, generalizable optimization models that can predict compounded transcription factor perturbation paths for more complex cell-state conversions. For our analysis, perturb-seq data from the genome-wide study done by Replogle et. al was used. After preprocessing, the data was used to prepare benchmarking models based on leading techniques from past literature due to no direct existing models for comparison. These benchmarks were then compared with three models, custom vector-addition-based greedy algorithms, constrained OLS, and non-linear SQP solver that perform the main task of predicting combinations of transcription factor perturbations for converting erythroleukemia cells to retinal pigment epithelial cells. The models were compared using two different similarity metrics, the industry-standard cosine similarity, and the custom projection score. The custom vector addition-based greedy algorithm achieved the highest score for both metrics, almost double the benchmark, and 90% for the cosine similarity score. All models predicted established transcription factors and novel TFs for this drastic conversion task proving generalizability and clinical applicability to further cellular reprogramming as a potential cure for a myriad of diseases.