Discovery of Hidden Gene Regulators: A Novel Machine Learning Approach to Transcriptional Pause Site Determination

Golla, Anudeep (School: Fairview High School)

Transcriptional pausing, the inhibition of RNA polymerase (RNAP) elongation during transcription, is a vital regulator of DNA transcription that offers increased control over gene expression. Erroneous or absent pausing can lead to genetic diseases such as Down Syndrome, Cystic Fibrosis, Huntington's Disease, and Sickle Cell Anemia. Currently, the only way to determine pause sites in a genome is to spend multiple days and thousands of dollars on genome sequencing, which still produces extremely inaccurate results. The objective of this study was to develop a more effective and less expensive way to determine pause sites. Using previously supported pause sites in the E. coli genome and aggregating sequential and parametric DNA data, this study developed and optimized probabilistic and neural machine learning models to predict pause site incidence in DNA segments. A randomized trees algorithm and novel sequential memory model proved to accurately predict pause site locations at an unprecedented 0.995 confidence level across all test environments. In addition, the pause site predictions of these models were then back-tested through in vivo biochemical experiments that analyzed the effect of inserting the predicted pause sequences on the expression of Green Fluorescent Protein (GFP) in a non-pathogenic strain of E. Coli, resulting in a strong correlation between real and predicted locations. The results of this study ultimately provide a highly favorable alternative to pause site prediction. Furthermore, gaining insight into the relationship between transcriptional activities and the human body can potentially help discover the genes that are directly related to genetic diseases.

Awards Won:

Second Award of $1,500