Illuminating Gene Dysregulation in Cancer: Deep Learning Identification of Disrupted Transcription Factor Binding Sites

Chiang, Bryan (School: Lynbrook High School)

Over 90% of cancer-associated mutations are non-coding, driving tumor development by impeding transcription factor binding – the “on and off switches” regulating key cell life and death genes. However, few of the disrupted binding sites underlying cancer progression have been identified and studied due to high experimental costs. In this study, a novel computational framework was developed to pinpoint and characterize sites of atypical transcription factor binding in cancer. High-capacity convolutional neural networks integrating epigenomic data were first constructed to detect genome-wide regions of binding across cell lines for 8 transcription factors linked to oncogenesis, while current methods can only recognize binding in a single cell line. The networks were evaluated against five accuracy metrics, significantly outperforming current methods with an average auROC score of 98.7%. Statistical methods were leveraged to rigorously screen for over 300 new sites of abnormal binding in previously uncharacterized breast cancer cell lines ($p<0.05$). The discovery process was validated with literature-mined sites of irregular binding and breast cancer mutation data. Furthermore, gene set enrichment and Ingenuity Pathway analyses highlighted 240 top-scoring downstream genes, protein-protein interactions, and canonical pathways, suggesting new mechanisms of biological disruption in breast cancer. This study provides the most thorough computational characterization of gene dysregulation in cancer to date and proposes the first-ever systematic approach for identifying regions of disrupted transcription factor binding. This research sheds new light on the molecular underpinnings of tumor biology and paves the way for future cancer diagnostics, prognostics, and clinical therapies.

Awards Won:
Third Award of $1,000
American Statistical Association: Certificate of Honorable Mention
Air Force Research Laboratory on behalf of the United States Air Force: First Award of $750 in each Intel ISEF Category