Predicting Cancer Stem Cell Biomarkers with Machine Learning

Liu, Julia (School: Nikola Tesla STEM High School)

Cancer stem cells (CSCs) are subpopulations of cancerous cells that are theorized to be the cause of tumor initiation and relapse. CSCs are especially dangerous because they are resistant to current cancer treatments. This project investigated the question of whether CSC biomarkers can be identified with machine learning by combining machine learning models with publicly available transcriptome data. Five machine learning models were trained, tested, and validated with human stem cell lines, somatic cell lines, and normal tissue samples. The model with the highest accuracy was determined and then used to classify cancer samples from 13 different cancer types into stem-like cancer samples and non-stem-like cancer samples. Differential expression analysis then identified overexpressed and under-expressed genes between cell and tissue samples as potential CSC biomarkers. The random forest model had the highest accuracies with 96.4% testing accuracy and 100% validation accuracy. The random forest model and differential expression analysis identified ANKRD35 as a CSC biomarker for 11 cancer types and IGFBP6 and SEPTIN14P12 for 10 cancer types. Highly differentially expressed CSC biomarkers for all 11 cancers with stem-like samples were also determined. All biomarkers are statistically significant with a p-value < 0.05. A study has already conducted research on IGFBP6 in cancer cells, but further research should also be done on ANKRD35, SEPTIN14P12, and other identified biomarkers. The identified CSC biomarkers are potential therapeutic biomarkers. Future developments in drugs that can target the shared CSC biomarkers may reveal new treatments for multiple cancers.

Awards Won:
Fourth Award of $500