A Novel Visualization Tool: Application to RNA-Sequence Classification

Ayyorgun, Ebru (School: Battlefield High School)

With the rapid growth of digital scientific datasets, data visualization is becoming increasingly important, and even the most popular tools require additional features. The goal of this study is to contribute a novel bivariate colormap feature to matplotlib, an open-source visualization library used by over 5 million scientists. This new feature is crucial in extracting more meaning from datasets that previously couldn't be visualized since they could only be classified by one variable. A novel function was added to the source code that maps two lists to one colormap, creates a customizable legend image, and offers many color options and data normalization. This tool was used to analyze real RNA-sequencing data for atherosclerosis classification with roughly 14,000 attributes. Dimension reduction was performed using UMAP to 3 dimensions for a 3D scatter plot. Using this novel tool, Keratin 8 and HLA-C genes were identified as potential genes to possibly play a role in atherosclerosis and were studied in a bivariate manner. In addition, different stages of atherosclerosis (fatty streak, fibrous plaque and cap, and normal tissue percentages) were studied in a similar fashion. Disease states of atherosclerosis was also mapped with age, gender, and ethnicity on the UMAP projection, which can have implications on the impact of atherosclerosis on different demographic groups. This knowledge can help gain more insight for ongoing research efforts in atherosclerosis. Furthermore, the novel visualization tool is available to the public and allows for more knowledge to be generated from high dimensional datasets at advanced research institutions.

Awards Won:
Fourth Award of $500
Fourth Award of $500