Improved Visualization of Dimensionality Reduction Plots With Controlled Downsampling

Hsu, Ashley (School: Commack High School)

Large biomedical datasets routinely consist of more data points than can be displayed faithfully on a computer monitor. Computer screens are incapable of representing high-volume datasets with perfect one-to-one pixel resolution, resulting in arbitrary data loss. This is problematic for researchers since critical details are obfuscated. The general problem of loss of data due to insufficient sampling is called the aliasing effect, when signals are sampled at a rate incommensurate to that of the original signal. Displaying data on a computer screen is a form of downsampling since the screen only samples a portion of the complete data. For single-cell RNA-sequencing data, this implies ignoring entire cells or genes, which is not consistent with biological principles. In this project, I aim to lessen the information loss between the original data and what is visible to the researcher. To accomplish this I have designed and implemented software, called AHggregate, to automatically cluster similar data and then downsample, with a targeted use case of single-cell RNA-sequencing data. I apply my methodology to publicly available single-cell transcriptome data obtained from Tabula Muris, and demonstrate with heatmaps and dimensionality reduction visualizations that the salient features of the data are preserved. While this study focuses on single-cell RNA-sequencing data, I expect that the methods developed here have applications to other large datasets. AHggregate creates better visualizations, advancing computational biology by enabling scientists to better understand the human condition through more accurate diagnoses, better targeted treatments in personalized therapies, and more precise insights into tumor activities.

Awards Won: