The Genetics of Human Aging: Predicting Age and Age-Related Diseases by Deep Mining High Dimensional Biomarker Data

Guan, Hannah (School: BASIS San Antonio Shavano Campus)

Aging is traditionally thought to be caused by complex and interacting factors such as DNA methylation. The traditional formula of DNA methylation aging is based on linear models and few works have explored the effectiveness of neural networks, which can learn nonlinear relationships. DNA methylation data typically consist of hundreds of thousands of feature space and a much smaller number of biological samples. This leads to overfitting and poor generalization of neural networks. In this research, I developed the Correlation Pre-Filtered Neural Network (CPFNN) model that uses Spearman Correlation to pre-filter the input features before feeding them into neural networks. I compared CPFNN with the statistical regressions (i.e. Horvath's and Hannum's formulas), the neural networks with LASSO regularization and elastic net regularization, and the dropout neural networks. CPFNN outperformed these models by at least 1 year in terms of Mean Absolute Error (MAE), with a MAE of 2.7 years. I also tested for association between the epigenetic age with Schizophrenia and Down Syndrome ($p = 0:024$ and $p < 0:001$, respectively). I discovered that for a large number of candidate features, such as genome-wide DNA methylation data, a key factor in improving prediction accuracy is to appropriately weight features that are highly correlated with the outcome of interest. My research is one of the first to adapt neural network algorithm in biological aging predication. The CPFNN model can be applied to a wide range of high-dimensional biomarker data, and ultimately improve understanding of aging process and benefit public health.

Awards Won:
American Statistical Association: Second Award of $1,000