DNAMethyl: Human Age Predictor based on DNA Methylation using an Attention-based Feature Selection Mechanism

Sparsh Gupta* and Vaneet Aggarwal†

*Franklin W. Olin College of Engineering
Needham, MA
sgupta1@olin.edu

†School of Electrical Engineering
Purdue University, USA
vaneet@purdue.edu

Abstract

Epigenetics focuses on the variation of gene functions which are affected by lifestyle and environmental factors. However, epigenetic changes do not alter the DNA sequence of an organism and vary throughout the natural course of an individual's life. One such epigenetic change is DNA methylation, which has various applications in cancer, atherosclerosis and aging. Previous studies have indicated that DNA methylation and aging have a high correlation which can be used to predict the age of individuals. We used the data obtained from the Illumina HumanMethylation BeadChip platform (27K or 450K) in this work which consisted of individuals' DNA methylation frequencies of CpG islands obtained from their blood samples. We analyzed 16 healthy datasets and 10 datasets consisting of individuals with a disease with the ages of individuals ranging from 0 - 103 years. Principal Component Analysis (PCA) was applied to remove outliers from the data and the final dataset consisted of 2363 healthy samples and 1488 disease samples. Pearson correlation between the CpG sites and the DNA methylation frequencies was established and the most highly correlated 100 common CpG sites in the datasets were chosen for the study. The data was standardized and normalized using MATLAB prior to feeding the data to Attention-based feature selection combined with LSTM model performed in Python programming. The 100 features (CpG sites) were passed through the Attention mechanism to sort and select the best and appropriate amount of features to pass on to the LSTM model and achieve the lowest Mean Absolute Error (MAE) in the predicted age of the samples. The model was also applied to 275 saliva samples obtained from 2 datasets. The final achieved MAE is still in progress but baseline results using a simple LSTM model on the healthy dataset obtain an MAE of ~ 6 years on the training set and an MAE ~ 11 years on the test set which indicates that applying Attention-based feature selection would greatly impact the model's results positively and would be state-of-the-art in human age prediction through DNA methylation.