Predicting Human Age via DNA Methylation Profile of a Blood Sample

Cole Killian
McGill Statistics and Computer Science

SCIENCE& SYNERGY
SUS ACADEMIA WEEK 2020

Introduction

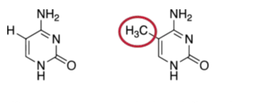
Why investigate human aging?

Recently, increasing efforts have been dedicated towards understanding the biological aging process. Questions relating to slowing and possibly even reversing aging have been making their way to the spotlight.

This paper suggests an improved technique for predicting a human's age based on their DNA methylation profile, especially when sequenced from a blood sample. This could be a step towards a better understanding of what it really means to age. While DNA methylation has been identified as the most promising molecular biomarker for the prediction of age, the upper limit for accuracy and precision is still unclear (Daunay, 2019).

What is DNA Methylation?

DNA methylation is a biological process by which methyl groups are added to certain positions in the DNA molecule; i.e. changing a cytosine to a 5-methylcytosine. While an organism's DNA sequence is mostly static throughout its lifetime, an organism's DNA methylation is subject to dynamic change (methylating and demethylating a certain position); this enables changing the activity of a DNA segment without changing the sequence.



Cytosine

methylated Cytosine

Overarching Question:

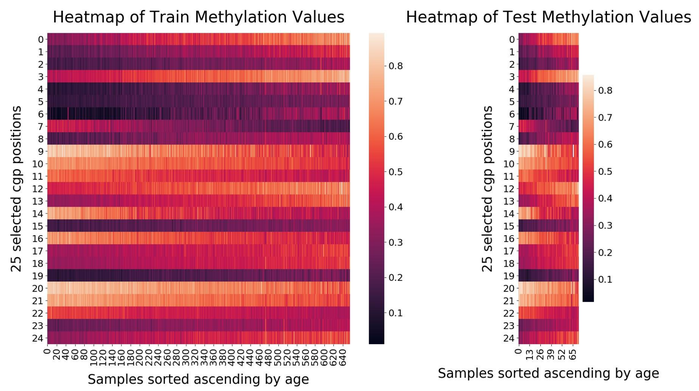
How accurately can we predict human chronological age from the DNA methylation of a blood sample?

Methodology

Data comes from the NCBI Gene Expression Omnibus GSE 87571. The data includes DNA methylation at 476,366 sites throughout the genome of white blood cells from a population cohort (N=421) ranging in age from 14 to 94 years old. There were 732 samples. The values of these features range from 0 to 1, where 0 corresponds to 0% of the white blood cells being methylated at that position, and 1 correpsonds to 100% of the white blood cells being methylated at that position.

Note that performing machine learning on 732 samples with 476,366 features is a sure way to overfit the model. Therefore the first step after preprocessing was to identify the features most absolutely correlated with age. However, doing so before splitting the data could lead to some biases being inserted into the model. Therefore I first split the data into training and testing sets at a ratio of 9:1, and selected the 25 most correlated features in the training set. Each of these features had a correlation with age between 0.83 and 0.94.

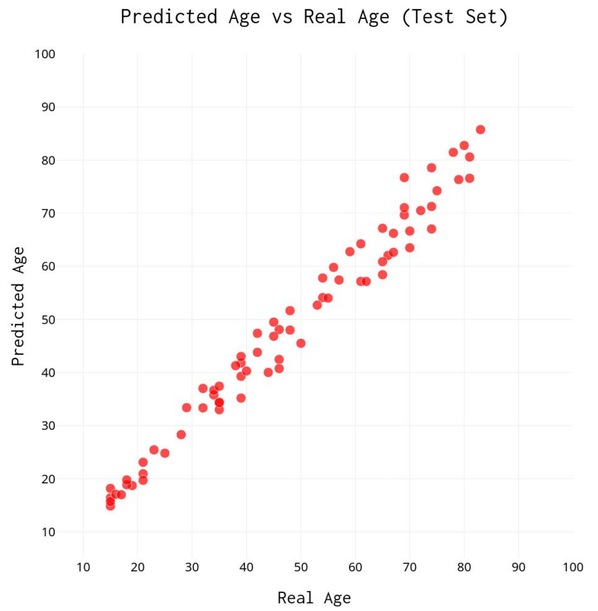
The training data was then fed into a sequential neural network. The model includes three dense layers each with 1024 ReLU nodes, three dropout layers each with a rate of 0.2, and a final dense layer with a single node corresponding to the predicted age. The model was trained for 50 epochs.



The heatmap above allows for a visualization of the correlation of each of the 25 features with age for both the train and test sets. The y-axis represents the selected methylation positions, while the x-axis represents the samples sorted in ascending order by age. It is clear that the values for each of the features tend to trend consistently up or down as age increases. While the trends in the test set are not quite as clear as those in the training set, they are still present.

Results

After training the model on the training set, it was applied to the test set. When comparing the model predictions with the real ages, there was a mean absolute deviation of 3.19 and standard error of the estimate of 0.46. We can compare these to the best age-prediction performing models found with the Bekaert and Thong models (MAD of 4.5–5.2, SEE of 6.8–7.2), followed by the Zbiec-Piekarska 1 model (MAD of 6.8 and SEE of 9.2) (Daunay, 2019). While my model outperformed previous attempts, note that these statistics are not based on the same test sets. Applying each of these models to the same test set would be the next step in more rigorously comparing them.



The scatterplot above displays the predicted age plotted against the real age for the testing set. All predicted ages are within ±10 years of real age; over 70% are within ±5 years of real. Check out https://epigenosys.com for a live demonstration!

References

Daunay, A., Baudrin, L. G., Deleuze, J.-F., & How-Kit, A. (2019, June 20). Evaluation of six blood-based age prediction models using DNA methylation analysis by pyrosequencing.

Li, X., Li, W., & Xu, Y. (2018, August 21). Human Age Prediction Based on DNA Methylation Using a Gradient Boosting Regressor.

Horvath, S. (2013). DNA methylation age of human tissues and cell types.

Johansson A, Enroth S, Gyllensten U. Continuous Aging of the Human DNA Methylome Throughout the Human Lifespan. PLoS One 2013;8(6):e67378. PMID: 23826282

Acknowledgements

McGill Artificial Intelligence Society
Office of Science Education, the Science Undergraduate Society, Teaching and Learning Services and University Advancement at McGill University.