Pre-Journal Entry for Yu Zhang Paper

Summary: DNA methylation is the addition of a methyl or CH3 group to a base of DNA, in the case of aging, a cytosine in a structure called a CpG island. Once methylated, these structures cause the DNA to wind further around a protein hiding the start site of a gene from being accessible. Through extended research DNA methylation has been shown to be an accurate predictor for the age and functional capability of an individual or an individual's organs. For these reasons research has centered on finding the best model to obtain the most accurate age predictions based on the levels of methylation found in DNA.

This paper specifically compares DNA methylation age predictions between various neural network models including those with and without regularization along with linear statistical methods, previously found to have the most accurate age predictions compared to basic neural networks. Neural networks were selected to participate as the focus of the study for their ability to model a nonlinear relationship between the input and the output and locate patterns within a more extensive set of data. However, neural networks can suffer from the "curse of dimensionality," wherein too much input or data can cause the model to become too complex and uses duplicated data points to try to find a data pattern. This results in a model with a lot of variance leading to the incorrect age prediction of methylation levels inserted into the already constructed model.

Knowing that the use of regularization and other approaches with the neural networks can help reduce the complexity of the model, Yu Zhang attempts to correct the issue of neural networks while continuing to use them in the prediction of age with methylation levels. Techniques added to reduce complexity are LASSO (least absolute shrinkage and selection operator), elastic net, drop out, and pre-filtering neural network. All these techniques coupled with neural networks, basic neural networks, and two statistical approaches were evaluated on their accuracy by an assessment of their average number of errors between methylation levels and age.

The results concluded that a neural network approach with the same data (Horvath, 353 sites; Hannum, 17 sites) used for the two statistical regression models resulted in a more accurate prediction of age. Compared to the entire dataset (473034 CpGs) analyzed by a neural network without one of the techniques to the neural networks including one of the techniques, all but one model including one of the techniques was found to reduce error in the prediction of age. LASSO was found to increase the number of errors likely due to the sparsity of the model. LASSO involves reducing the amount of data used to construct the model. Since LASSO removes data randomly that produce similar outcomes in the model, it can randomly select a CpG site affecting age in the way many might. This then may make it difficult for a relationship to be found resulting in the increased number of errors. The elastic approach counters this problem by reducing the complexity and making the model sparser, but still incorporating highly associated CpG sites with age into the model, making the relationship between the two variables more prominent. Dropout, however, beats this method by reducing the overall complexity of the model through the thinning of connections between CpG sites and age resulting in more of the model's resources being used to compute other outputs, and a lack of increased significance attached to duplicate outputs only found in sets used to make the model and not in the actual dataset. Overall, the correlation pre-filtered neural network approach (CPFNN) introduces a statistically significant decreased average number of errors. This approach reduces the number of inputs resulting in decreased complexity of the model, a faster computational speed, and more importance given to CpGs with a higher correlation to age.

To address applications of this more accurate model in age related diseases, CPFNN along with the two statistical models were used to calculate the age acceleration of patients with schizophrenia and then with patients with down syndrome. Age acceleration is a measure of the difference between an individual's chronological age, and their biological age ascertained from their DNA methylation levels coupled with the models. Compared to the control, the age acceleration of the patient was found to be statistically significant in down syndrome for all three models. However, with schizophrenia the two statistical approaches found no statistical significance while the CPFNN model found a statistically significant association. With the use of external studies backing the association found with CPFNN, the model was further proven to be a more accurate prediction of age with regards to DNA methylation levels even in the detection of weak patterns.

Importance: Based solely on the results and application of the findings of this paper, the research proves itself to be important and a vital piece to the overall discovery of age and its influence on age related diseases. Not only does the paper provide new and more accurate models for the prediction of age with methylation levels, but it also demonstrates how one of these models, CPFNN, can be used to identify weaker patterns between age and age-related diseases. These results are important since the implications of these could lead to an earlier diagnosis or scanning for potential diseases and a more knowledgeable account of diseases associated with increased methylation levels and thus aging. If diseases are known to present statistically significant differences between an individual's chronological age and their biological age, the testing of an individual's DNA with these models can lead professionals to examine if they may be at increased risk for diseases that present these differences. Furthermore, the study suggests how to utilize all the features of DNA methylation by using an automatic grad search algorithm to find the exact number of data needed to be included to produce an accurate prediction of age. However, these results would not be applicable to further research if the research conducted was not controlled for confounding variables. While the reason for datasets being chosen for certain aspects of the experiment needs to be further explained, the researchers went to great lengths to prevent batch effects, to prevent bias in the splitting of data sets between training and test sets and unequal age distributions, and to prevent the differences of intensity of methylation across data sets. Due to the increased control of confounding variables expressed within the study, the results appear to be usable for further research. Since the results are valid and the results present new information to the field, I deem the research to be important.

Core Concepts: Methylation of DNA usually occurs through epigenetic factors meaning environmental factors result in DNA methyltransferases mutating the DNA leading to the molecular change of the expression of genes and their products. Increased levels of DNA methylation are associated with aging and age-related diseases leading to the decrease in an individual's fitness to survive and thus continue to reproduce. The aging referred to in the article assumes one understands RNA is transcribed from a gene within the DNA and then translated into a functional protein, and thus by restricting certain genes from being accessible through methylation this leads to a lack of informational flow. A lack of typically produced proteins can lead to structural changes of the cell, which can ultimately lead to a decreased functional ability of a decent amount of cellular processes. Assuming the structure of the cell undergoes substantial damage, the electrochemical gradient of charged molecules could be disrupted by the decreased impermeability of the lipid bilayer leading to altered movement of molecules across the membrane. This impairment of cellular function weakens the cell making it more susceptible to lysis and ultimately as more cells within an individual age they can be more susceptible to disease and infection especially if this aging affects their immune cells. If the immune cells of the body are negatively impacted, then the influence of disease on other systems of the body will become more severe.