Building Aging Clocks Using Transcriptomics and EEG Data

Davyd Sadovskyy - Krishnan Lab

Background and Motivation

I had an initial interest in human aging research so Dr. Arjun Krishnan introduced me to his preprint paper "Human pan-body age- and sex-specific molecular phenomena inferred from public transcriptome data using machine". This work curated a transcriptomics dataset to do age and sex modeling on.

Modeling biological age is important for two main reasons. First, a good aging clock model can serve as a tool to quantify whether a particular aging intervention is working to slow the speed of aging. Second, knowing what someone's model predicted biological age is lets us obtain an "aging gap" score, by subtracting the biological age from their real chronological age. These residuals have been shown to be informative of an individual's health - if someone is "below the curve", they tend to be healthier on average.

Aging clocks can be built using a wide variety of molecular data, such as proteomics, metabolomics, transcriptomics, and even brain wave recordings. During my rotation, I was introduced to Dr. Brice McConnell, who had been modeling age and other biological variables using EEG recordings from sleep. My goal was to understand the process used to generate features from EEG data and to learn about and implement novel methods to reduce the feature set.

Accomplishments and Learnings

One of the things I accomplished was critically reading, rereading, and discussing 7 papers about aging clocks that Dr. Krishnan sent to me. Each paper introduced me to a new perspective of building aging clocks. A couple memorable examples include a paper that built aging models using longitudinal instead of cross-sectional data and showed that the omics features had nonlinear relationships over time, and another paper that showed how accuracy can be improved by modeling the age of individual organs instead of the entire organism.

I also learned about and implemented a novel neural network method to predict responses that are ordinal in nature, such as age bins. In the process I gained an understanding of what microarray and RNAseq data is, what its units are, and the necessary normalizations required for each. I showed that performance could be drastically improved by treating age bins as ordinal variables.

When working with the EEG data, I learned new concepts in neural time series analysis by watching Mike Cohen YouTube lectures, which were recommended by Dr. McConnell. I gained a general understanding of how feature engineering of raw EEG data was done by following MATLAB code. I didn't have enough time to create new features based on EEG data, so my goal

was simply to try different feature selection techniques with the hopes of understanding which categories of predictors were the most important. I learned about the traditional feature selection techniques such as recursive feature elimination and SHAP values. I also learned about how the traditional out of bag feature importance scores for random forests, which are still used by many researchers, are flawed. This introduced me to the concepts of knockout variable importance scores and Boruta, both methods that create permutations of the feature matrix in order to obtain more reliable feature importance scores. Additionally, I was pointed to a paper that developed a novel method for selecting features called minipatch variable selection. I had to critically read this paper and implement the package the authors developed for my aging clock use case.

I also read a paper about biologically informed LASSO regression, which is a novel method of selecting features that have known biological importance. Discussing this paper with Arjun and Parker cleared up misconceptions I originally had and helped me understand the mathematical formulas.

Conclusion and Future Directions

I gained a good understanding of the landscape of human aging research, from a computational biology perspective. I read about a dozen research papers which was great practice of critical reading and working through difficult mathematical sections or complicated figures. I learned the novel statistical learning methods of ordinal regression neural networks and minpatch feature selection. This helped me brush up on my machine learning knowledge. It also gave me confidence that I can understand state of the art computational methods in research papers, which was a nice feeling.

An obvious future direction would be to tie up the loose ends with the EEG data analysis and reimplement minipatch learning so that it gives me real predictors, and not just ones from the end of the data set. I suspect there is a bug somewhere in the code. To resolve this, I would implement minipatch using the data used in the original research paper and then try to isolate the issue on my EEG data.

I am continuing my interest in studying human aging in my next rotation with Dr. Joanne Cole. We discussed creating an “aging gap” phenotype and testing the association between it and the genome. I am excited to shift my focus to learning about the genetic aspect of aging.