Project Proposal: A single cell RNA-Seq based aging clock for human neurons

Qiao Su
Department of Computational Biology
Carnegie Mellon University
Pittsburgh, PA 15213

Mirudhula Mukundan
Department of Computational Biology
Carnegie Mellon University
Pittsburgh, PA 15213

Jiayin Zhi
School of Design
Carnegie Mellon University
Pittsburgh, PA 15213

1 Introduction

Aging is a biological process which affects every single cell in the body, but the exact molecular mechanisms remain unknown. In order to decipher this mechanism, we can study which gene features are predictive of chronological age. Phenotypes such as aging are usually studied at the level of bulk data, which comes from aggregated expression data from multiple cells.

Past studies have shown some success predicting chronological age from bulk human and mouse gene expression data using elastic net models (1). However, these bulk tissue measurements do not explain how heterogeneous populations of cells differ in terms of their gene expression as aging progresses. Fortunately, single-cell sequencing technology allows us to examine disease biology in unprecedented detail.

However, there are also new computational challenges involved in processing and interpreting this high dimensional expression data. In addition to biological noise from the stochastic process of mRNA production, there is technical noise due to the low amounts of genetic material that can be captured from a single cell. Therefore, deep learning classifiers may be especially well suited to this challenge.

We hypothesize that chronological age can be predicted from features of single-cell expression data using deep learning methods. Towards this goal, we aim to compare deep learning models for predicting age from a real biological dataset, which in this paper is taken from 189432 human neurons from 69 control patients of a study of schizophrenia (7).

2 Background

A comparative study between chronological age and biological age is useful in understanding aging as a phenomenon explained by a set of biomarkers. Recently, studies have shown to indicate that the internal aging clock can be explained by using gene expression data (2; 3). Some studies have utilized this to build models for prediction of age using Genotype-Tissue Expression (GTEx) profile from multiple human tissues using an elastic net algorithm (12; 9) to account for sparsity in data. Another study conducted age prediction analysis in zebrafish with the help of a simple Multilayer Perceptron Model (MLP) with one linear layer, two ReLU activation functions, and a softmax layer to predict the biological age of the zebrafish into three discrete bins indicating stage of age in life (8). In our project, we will try to incorporate single-cell gene expression data from human neurons to predict biological age of the subject using various deep-learning techniques.

10617 Class Project (2022).

3 Methods

3.1 Preprocessing

The entire dataset consists of 189432 human neurons from 69 controls, containing a distribution of ages from 25 to 94 years (Fig 2). To facilitate the fit of the deep learning model, we applied feature scaling of both the target variable as well as z-scoring of the feature matrix using sklearn StandardScaler.

3.2 Architecture and data split

We implemented two baseline models based on Multi-layer Perceptron (MLP) with varied architecture and one linear regression model, in order to perceive how the data would be projected across such different models:

1. An elastic net model with parameters tuned by 5-fold cross validation, for the purposes of comparison with previously published elastic net models. The train/validation and test set consisted of 55 and 14 individuals respectively.
2. Baseline 1: An MLP regressor (skorch NeuralNetRegressor) with 3 linear layer of size 5000, relu activation, and then a linear output layer. For ease of direct comparison, the train/validation and test set was the same as for the elastic net model
3. Baseline 2: A Pytorch implementation of regression using MLP with 2 linear layers, one of size 100 and another of size of 50, with ReLU activation layer, 1-d Batch Normalization and Dropout probability of 0.4. The optimizer used was Adam, with weight decay of 0.9.

On all of these architectures, we report the mean absolute error (MAE) of the classifier on the validation set, when averaged across all the cells for each patient.

4 Baseline Results

The distribution of the ages of the different patients is shown in Figure 2 under the Supplemental section. The baseline results across all three models are shown in Figure 1. The goal of the first baseline NN model was to observe the predictions when we have a large network (5000 neurons), and the second baseline model was implemented to observe the results under the effect of dropout and batch normalization layers, with a relatively smaller network. The Mean Absolute Error (MAE) for the linear model was found to be the the least followed by Baseline NN 1, and then Baseline NN 2. Similarly, the Pearson Correlation value (R) was found to be the highest for the linear model. This shows that the neural network baseline models do not seem to perform as well as the elastic net(glmnet) model.

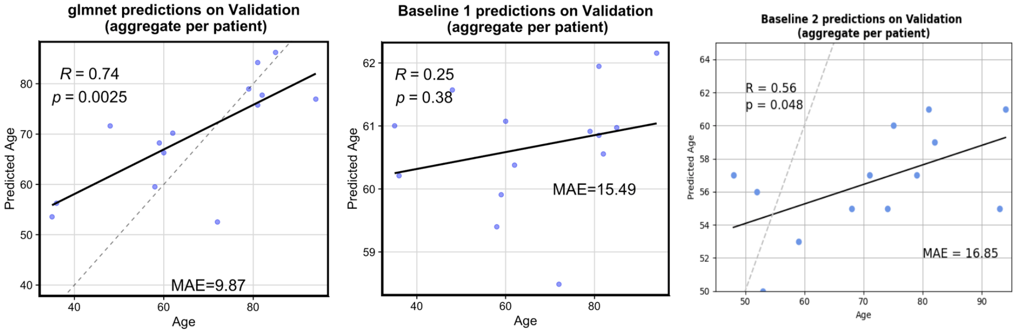

Figure 1: Linear Model vs Baseline result 1 vs Baseline result 2. In each graph, the Pearson Correlation (R) and associated p-value is shown, as well as the MAE. The glmnet elastic net model outperforms the both baseline neural network models.

5 Future Work

Since we have found that architectures from our baseline models are unable to perform as well as a linear elastic net model, which is a sparse model, we next plan to implement dimensional reduction techniques in order to extract features of interest within this expression data. In the section, we discuss three techniques to do this: NMF, VAE and DAE.

5.1 Introduction to NMF data decomposition

The interpretation of scRNA-seq data requires methodological innovations. To find the important biological signals, matrix factorization approaches can find a simplified and thus more interpretable representation of an expression matrix which can also be less costly for a deep learning model to train. For example, exploratory analysis is usually assisted by unsupervised matrix factorization approaches such as principal components analysis (PCA) or t-distributed stochastic neighbour embedding (t-SNE). Non negative matrix factorization (NMF) is another popular matrix factorization approach with the advantage that it has been shown to be able to discover biologically meaningful gene expression programs as latent factors compared to PCA and t-SNE (6). This approach has been successfully used for cell type markers in the past, but it has not yet been applied towards deciphering gene expression programs such as human aging. We will attempt to infer the top 100 latent variables using NMF and check their correlation with Age.

5.2 Introduction to Variational Autoencoder

Studying patterns in the level of transcripts in the scRNA-seq data can lead to identifying key molecular signatures responsible for the aging process. But, this data is highly complex in nature and poses great challenges in both regression analysis and feature selection. Following this, a very high features-to-examples ratio, which is a widely prevalent characteristic of biological and medical data, makes the model close to uninterpretable as well. Scientists have come up with ways to compress large datasets in order to identify and consolidate across multiple latent dimensions to capture key gene expression representations in a quick and efficient manner (11). Previous studies have worked with novel techniques that involve variational autoencoders (VAE) to compress RNA-seq data for extracting latent dimensions (10) or for performing dimensional reduction in scRNA-seq data related to cancer (4). But, no study has yet been conducted on the application of VAEs on ageing related transcriptomic data. Therefore, this project will explore how regularization between a series of linear layers of an autoencoder will help in identifying good latent dimensions for efficient regression.

5.3 Introduction to Denoising Autoencoder

As there are no limits in architecture for latent variable modeling with deep neural networks, such as the depth of the network and the types of layers in-between the network's input layer and output layer, or regularization used, current practice in biological fields has been widely conducted on the application. Also, it can typically make data visualization easier for non-numerical data types as well as finding similarities in a complex dataset. However, the current practice in biological fields on aging data has not been fully conducted on the application of DAE (denoising autoencoder). Therefore, in addition to the practice of deep neural networks models, this project will use a multi-layer DAE (denoising autoencoder) to extract the features which are most representative and informative in constructing a deep neural network (5). In this way, the denoising autoencoder will learn the features which will be tuned using a softmax classifier and fully connected layers. scDAE may be able to outperform other methods on this noisy single cell dataset.

6 Teammates and work division

Q.S. was responsible for implementing the elastic net and Baseline 1, and will implement NMF decomposition for the final report. J.Z. was working on implementing the MLP model, and will try to implement DAE for the final. M.M. implemented Baseline 2 and will be working on VAE for the final report.

7 Supplemental Figures

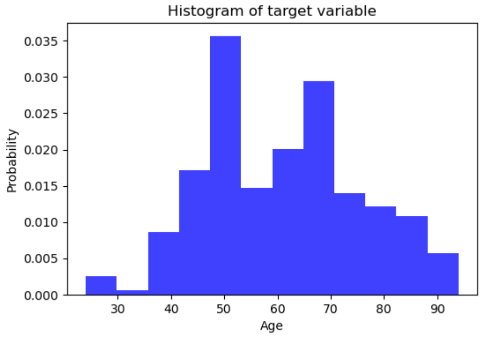

Figure 2: Distribution of target variable (Age) across 189432 individual cells measured.

References

[1] Bergsma, T., and Rogaeva, E. DNA Methylation Clocks and Their Predictive Capacity for Aging Phenotypes and Healthspan. *Neuroscience Insights 15* (2020).

[2] De Magalhães, J. P., Curado, J., and Church, G. M. Meta-analysis of age-related gene expression profiles identifies common signatures of aging. *Bioinformatics 25*, 7 (2009), 875–881.

[3] Frenk, S., and Houseley, J. Gene expression hallmarks of cellular ageing. *Biogerontology 19*, 6 (2018), 547–566.

[4] Grønbech, C. H., Vording, M. F., Timshel, P. N., Sønderby, C. K., Pers, T. H., and Winther, O. scvae: variational auto-encoders for single-cell gene expression data. *Bioinformatics 36*, 16 (2020), 4415–4422.

[5] JOUNGMIN CHOI, J.-K. R., and CHAE, H. scvae: Cell subtype classification via representation learning based on a denoising autoencoder for single-cell rna sequencing. *IEEE ENGINEERING IN MEDICINE AND BIOLOGY SOCIETY SECTION 9* (2021), 14541–14550.

[6] Kotliar, D., Veres, A., Nagy, M. A., Tabrizi, S., Hodis, E., Melton, D. A., and Sabeti, P. C. Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *eLife 8* (jul 2019).

[7] Ruzicka, B., Mohammadi, S., Davila-Velderrain, J., Subburaju, S., Tso, R., Hourihan, M., and Kellis, M. Single-Cell Dissection of Schizophrenia Reveals Neurodevelopmental-Synaptic Link and Transcriptional Resilience Associated Cellular State. *Biological Psychiatry 89*, 9 (2021), S106.

[8] Singh, S. P., Janjuha, S., Chaudhuri, S., Reinhardt, S., Kränkel, A., Dietz, S., Eugster, A., Bilgin, H., Korkmaz, S., Zararsiz, G., et al. Machine learning based classification of cells into chronological stages using single-cell transcriptomics. *Scientific reports 8*, 1 (2018), 1–12.

[9] Wang, F., Yang, J., Lin, H., Li, Q., Ye, Z., Lu, Q., Chen, L., Tu, Z., and Tian, G. Improved human age prediction by using gene expression profiles from multiple tissues. *Frontiers in Genetics 11* (2020), 1025.

[10] Way, G. P., and Greene, C. S. Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2018: Proceedings of the Pacific Symposium* (2018), World Scientific, pp. 80–91.

[11] Way, G. P., Zietz, M., Rubinetti, V., Himmelstein, D. S., and Greene, C. S. Compressing gene expression data using multiple latent space dimensionalities learns complementary biological representations. *Genome biology 21*, 1 (2020), 1–27.

[12] ZOU, H., HASTIE, T., ET AL. Addendum: regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B 67*, 5 (2005), 768–768.