Pathway Scores: Feature Engineering for Age Prediction from Gene Expression Data

Yash Pershad
Department of Bioengineering
Stanford University
Stanford, CA
ypershad@stanford.edu

Rishabh Kapoor
Department of Biology
Stanford University
Stanford, CA
rishabhk@stanford.edu

Abstract—With increasing life expectancy, the disease burden posed by aging has become increasingly salient. Individuals may vary in their rates of aging due to genetic and environmental factors, inspiring the concept of a biological age, which is related to, but distinct from, chronological age (years since birth). To contribute to an understanding of the biological basis of aging, we apply regression techniques to predict the age of patients from gene expression data and thereby uncover patterns of gene expression associated with age. Our key innovation is to map individual gene expression values into a lower-dimensional pathway score space, representing the level of activity within expert-curated biological gene sets. Due to an emerging understanding of aging as occurring at the level of pathway dysregulation, we hypothesized that models trained on pathway activity scores may outperform models trained on individual gene expression. With a leave-one-out cross validation evaluation scheme, our pathway-based linear regression models performed better than the gene-based models (R^2 = 0.68 vs R^2 = 0.48) and a shallow neural net (R^2 = 0.573) . These results suggest that the pathway scores are a viable feature for prediction of age from transcriptomic data and provide support for the pathway dysregulation hypothesis of aging.

Keywords—transcriptomics, aging, linear regression, gene set enrichment analysis

I. Introduction

Aging can be defined as "intrinsic, progressive, and irreversible deterioration of virtually every bodily function".[1] Dramatic increases in life expectancy in modern times have exposed age-associated diseases, which contribute a significant burden in the US healthcare system ($135 billion).[2] Understanding the biological basis of aging is a central problem in modern biology and may provide insights into anti-aging treatments to reduce age-related illnesses.

Increasingly, researchers are realizing that chronological age is an inadequate predictor of senescence due to heterogeneity in aging rates across individuals. Instead, a biological notion of aging would allow for quantification of age based on expression of age-associated genes or other biomarkers. One approach to determining the biological determinants of aging is to train machine learning algorithms to predict chronological age from gene expression data. Genes that are highly predictive of age might then be causally linked to aging-related physiological changes.

Therefore, we built a linear regression model that could predict an output of chronological age from gene-expression-based features. We also aimed to analyze feature importance to understand the biological determinants of aging. We trained several models with different inputs. Our baseline model used inputs of all genes for which we had expression values. We obtained improved correlation of predicted and observed ages by feature maps to the raw gene expression data to produce pathway activity scores. This explicit feature engineering reduces dimensions by a factor of approximately 34 and performs better than a regression model with automated feature engineering via a neural network.

II. Related Work

A. Predicting age from gene expression

Recent advances in RNA-sequencing (RNA-Seq) technology have spawned an interest in predicting chronological age from gene expression data. Fleischer et al use gene expression data publicly available in Gene Expression Omnibus (GEO) from healthy individuals of ages ranging from 1 to 94 years old to train an ensemble of linear discriminant analysis classifiers that successfully predicts age with a median absolute error of 4 years using normalized gene expression values as features.[3]

While Fleischer's approach represents a key advancement in the application of machine learning to chronological age prediction, we have identified several key limitations. First, a large body of research suggests that biological aging occurs primarily at the level of pathways (e.g., nutrient sensing, proteostasis, DNA repair) rather than individual genes. As an example, perturbations of any gene in the nutrient sensing pathway upstream or downstream of the protein mTOR1 can have a similar effect on aging rate. Secondly, in the large literature on prediction of complex disease states from gene expression data, machine learning classifiers trained on genetic-pathway-level activity scores have outperformed models trained on expression of individual genes.[4,5,6] This is partly because reducing the number of features increases robustness to individual gene expression differences and because pathophysiology, like aging, occurs at a pathway level. Pathway scores can therefore be interpreted as a biologically-inspired feature engineering technique that reduces the dimensionality of transcriptomic data to prevent overfitting while capturing the structure of age-relevant variation in the original data.

B. Feature mapping

Here, "pathways" consist of expert curated sets of genes involved in a common biological process. Gene Ontology (GO) is a hierarchical structure of pathways with annotated gene functions.[7] The gene sets of interest are sets of genes involved in biological processes. As an example, an "inflammation" pathway would consist of a set of genes encoding products that participate in inflammatory processes.

A number of possible pathway score generation algorithms (feature maps) have been developed, as summarized and evaluated in Zhang et al.[8] We selected two

algorithms based upon their interpretability and computational efficiency.

1) Enrichment scores as features

The first feature map is an adaptation of single-sample Gene Set Enrichment Analysis (ssGSEA), a technique to obtain a metric of pathway activity from a ranked list of genes by expression level. Briefly, the algorithm produces a Kolgomorov-Smirnov-like statistic ("enrichment score") for each pathway by iterating through the ranked list and adding a "reward" every time a gene in the pathway is encountered, and subtracting a "penalty" when the gene is not in the pathway. The enrichment score is determined as the maximum positive or minimum negative statistic obtained while iterating through the list. If a pathway is upregulated (activated gene expression) more than random expectation, then the enrichment score is large and positive, because many rewards are added for genes in the pathway before non-pathway genes are encountered. If a pathway is downregulated (deactivated gene expression), then its enrichment score will conversely be large and negative. Traditionally, the resulting enrichment score is compared to a null distribution to determine a p-value. For our purposes, we propose using the raw enrichment score for each pathway as our feature. The ssGEA algorithm is well-suited for feature mapping because it can be computed from a single input gene expression vector (i.e., does not require a comparison to a training set).

2) Aggregated pathway z-scores as features

The second feature map involves using gene sets and calculating an activity score for each gene set from the z-score of genes in that pathway, as was done in Lee et al.[4] First, we calculate a z-score for each gene expression value relative to the other samples. For each gene, we compute a Pearson correlation statistic between each gene's z-score and patient chronological age. Then, for each gene set, starting from the highest ranked gene, gene z-score vectors are averaged to produce an output activity score vector until the correlation of the activity score vector with age ceases to increase. The average of the z-scores of the genes included in the subset represents the activity score of the pathway. Thereby, only a subset of genes in a gene set are used to compute the activity score. As compared to enrichment scores, this algorithm has the advantage of excluding genes uncorrelated with age, thus reflecting a filter feature selection strategy.[9] However, the z-score generation requires comparison to all other elements in the training set and thereby cannot be computed from a single input gene expression vector.

III. Dataset

We used Fleischer et al's gene expression data, in transcript frequency per kilobase million (FKPM), which is publicly available in GEO (GSE113957).[3] The study included data for 27,142 transcripts per individual from 133 healthy individuals, with ages ranging from 1 to 94 years old. The ages are roughly uniformly distributed over this age (Fig. 1). We analyzed the data for the influence of sex and found no significant differences by this potential confounds on PCA (data not shown). There were also 10 patients with Hutchinson-Gilford progeria syndrome (HGPS), an early-aging disease. Z-score normalization for each gene was only performed for "pathway z-score" feature mapping, as described above.

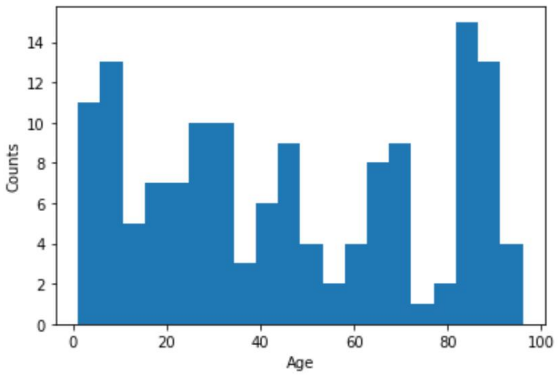

Fig. 1. Histogram of ages in dataset from Fleischer et al.

IV. Methods

A. Feature mapping

In order to obtain pathway scores from our dataset, it was first necessary to convert our feature names from RefSeq transcript IDs to Entrez Gene IDs, which are conventionally used in curated GO pathway sets. This mapping was performed using a standard UniProt dictionary to generate 16,609 genes as features.[10] The number of genes is lower than the number of original transcripts because not all transcripts correspond to protein-encoding genes with Entrez IDs and because multiple transcripts can map to the same gene, in which case we aggregated the expression values into a single column by summing across all transcripts.

We obtained single-sample GSEA (ssGSEA) enrichment scores using the *ssGSEA* function of the *GSEApy* package.[11] Subsequent models were trained with these enrichment scores as features for 2,286 pathways. We implemented the algorithm for obtaining pathway z-scores as features according to the pseudocode provided in Lee et al (as described in the Related Works Section), resulting in 4,154 pathway scores.[4] Pearson correlation values were obtained using the *stats* module from *scipy*. Pathway activity scores were used as features in downstream modeling.

B. Data visualization with Principal Component Analysis (PCA)

PCA provided a readily interpretable means of visualizing important patterns of variation in our high-dimensional data set in a lower-dimensional space. Briefly, PCA works by rescaling all features such that they have a mean of zero and variance of 1. Then, we find unit basis vector u for a new subspace such that the projection of the original data onto the basis vector maximizes the variance of the projected data, by optimizing the following equation subject to $||u||_2 = 1$, as shown in Fig. 2.[12]

$$\frac{1}{n}\sum_{i=1}^{n}(x^{(i)T}u)^2$$

Fig. 2. Equations for finding the basis vector in principal component analysis[11]. u is the unit basis vector and $x^{(i)}$ is a sample in the original data.

The solution to the above constrained optimization problem is the first eigenvector of the covariance matrix of the data. To obtain a two-dimensional subspace, we selected the first and second eigenvectors of this matrix as the

horizontal and vertical axes of our lower-dimensional subspace. Data are then projected onto this subspace by computing their projections onto the basis vectors u_1 and u_2. The fraction of the variance explained by each principal component is computed as the ratio of its corresponding eigenvalue to the sum of all of the eigenvalues in the covariance matrix.

We performed PCA on the gene expression data and transformed pathway features with the *PCA* function from *decomposition* in the *sklearn* package, with 2 principal components. We plotted the dimensionality reduced data with *matplotlib* and visualized patterns of variation with age with the *color map* function.

C. Building ordinary least squares regression models

Using the *sklearn* package,[13] we fitted an ordinary least squares (OLS) model to each of our three sets of features: (1) gene expression for 16,609 genes, (2) GSEA enrichment scores for 2,286 pathways, and (3) aggregated z-scores for 4,154 pathways. OLS works by finding the weight vector w that minimizes the squared deviation of the predictions (Xw) from the labels y according to the equation in Fig. 3. The coefficient of each feature from w is used to compute feature importance for models with pathway scores as features.

$$\min_w ||Xw - y||_2^2$$

Fig. 3. Optimized loss function for linear regression for *sklearn*.

With 133 training examples, training a model on the full set of features would result in an underdetermined regime vulnerable to overfitting. To reduce the model complexity, we applied multiple filters. First, as done in Fleischer et al,[3] we restricted analysis to genes with at least one sample with an expression level greater than 5 FPKM and a fold change of at least 5 FPKM between any two samples. These filters address the high degree of stochastic variation for low level read counts in RNA-Seq and ensure resulting genes have biologically significant variation, respectively. Then, for each respective model, we applied filter feature selection to select the top 500 features by mutual information with age using the *SelectKBest* model in *feature selection* in *sklearn*.[9,13]

D. Regression with neural networks with one-hidden-layer

Our GSEA and aggregate z-score feature maps represent feature engineering based upon domain knowledge. An alternative approach to explicit feature extraction is to train a neural net, which learns salient features automatically. Using the *keras* library of the *tensorflow* package, we trained a neural network with a single hidden layer with 500 neurons; the number of neurons therefore corresponds to the number of engineered pathway features selected from our maps. The single hidden layer of our neural net accepted all 16,609 genes as inputs, then produced a rectified linear unit (ReLu) activation for each of the 500 neurons. The outputs of this hidden layer serve as features for the final layer of the model, which maps $\mathbb{R}^{500} \rightarrow \mathbb{R}^1$ with a linear activation function. We trained over 100 epochs with a batch size of 10, with a loss function of mean absolute error. Model parameters for each layer are learned to minimize mean absolute error using the stochastic gradient descent-based Adam algorithm with a standard backward propagation approach. We visualized the loss function as a function of the number of epochs.

E. Model evaluation

To evaluate our regression models, we performed leave-one-out cross-validation (LOOCV) with the *LeaveOneOut* function of *feature selection* in *sklearn*. LOOCV involves holding out one sample, training on all other samples, and predicting the label of the held-out sample for every sample. This process generates a predicted label for each sample. As compared with k-fold cross-validation, LOOCV is well-suited to our dataset with a relatively small number of samples relative to the number of features. In accordance with Fleischer et al, we use the coefficient of determination (R^2) between predicted and observed ages as our primary metric of validity. R^2 quantifies the fraction of the variance in actual age explained by variance in our predicted ages, and is thus well suited to the objective of understanding how variation in gene expression contributes to aging.

V. Results and Discussion

A. Feature Comparison with Principal Component Analysis

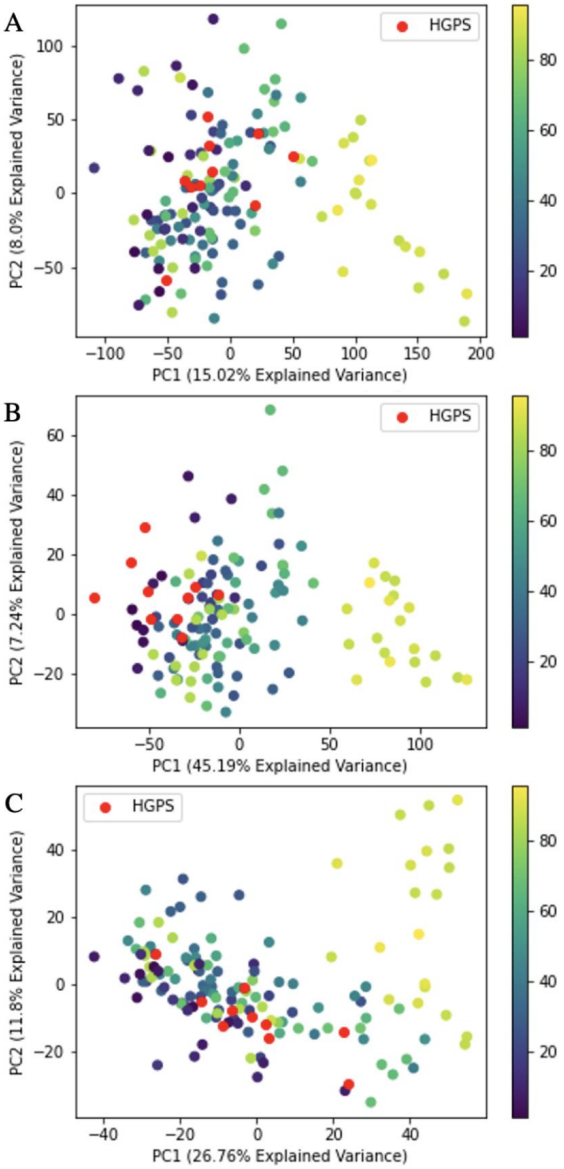

Fig. 4. Principal components 1 and 2 for (A) genes as features , (B) aggregate pathway z-scores as features, and (C) single-sample GSEA (ssGSEA) enrichment scores as features. The scatterplots are colored on a scale by age. Red points represent patients with the early-aging condition Hutchinson-Gilford progeria syndrome (HGPS).

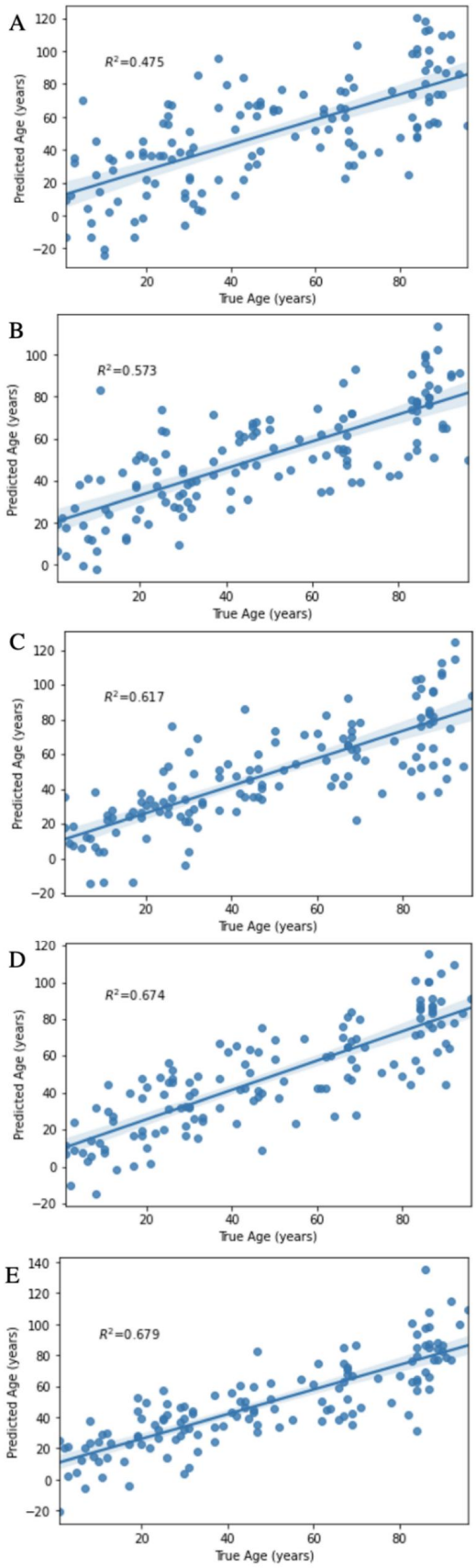

Fig. 5. Scatterplots of linear regression predictions after leave-one-out cross-validation versus true age of samples. (A) Ordinary least squares (OLS) regression on the top 500 gene expression features (R^2 =0.475), (B) Regression with a neural network with a single hidden layer with 500 neurons (R^2=0.573), (C) OLS regression with the genes constituting the top 500 pathways from the aggregate pathway z-score method, (D) OLS regression with top 500 features of the aggregate pathway z-scores, and (E) OLS regression with the top 500 features from the single-sample GSEA enrichment scores as features.

Fig. 4 shows the results of PCA dimensionality reduction when applied to our three types of features. The first principal component (PC1) in all three cases is correlated with age (i.e., R^2 between PC1 and age of about 0.60): older patients have the highest values when projected onto this axis. The second principal component for ssGSEA scores also captures some variation with age. These results show that age is an important contributor to variation in patterns of gene expression in the data, However, for the PCA on the gene expression data (4A), there is poor delineation between young and mid-aged patients, and relatively little (15.02%) variation in the original data is captured by the first principal component. In contrast, the principal components for both types of engineered pathway scores explain a higher fraction of the variance. Most importantly, there is a clearer gradient from younger to older patients along the entire length of PC1 for aggregate pathway Z scores, and along the diagonal between PC1 and PC2 for ssGSEA enrichment scores. The gradient in the PCA plot shows that, by averaging out irrelevant stochastic variation in individual gene expression, pathway scores are better able to capture patterns of variation in gene expression that correlate with age throughout the full range of ages in our dataset.

B. Regression Model Performance

Fig. 5 shows the predicted age versus observed age for all combinations of feature and model choice tested. As can be seen, all models produce a significant ($p < 0.05$) positive correlation between observed and predicted ages, but the model trained on gene expression values alone has the worst performance (Fig. 5A). Our two modes of feature extraction, pathway aggregate z-scores and ssGSEA enrichment scores, have the best performance, with an R^2 of 0.674 and 0.679, respectively. This shows that grouping genes into pathways appears to improve the ability of the OLS model to capture transcriptional variation salient to age prediction. Since the algorithm for selecting pathway aggregate z-scores involves selecting genes within a pathway according to their correlation with age (see "Related Work"), we also tested the performance of a classifier trained on all the non-aggregated z-scores of genes selected to compute aggregate z-scores. The resulting classifier (Fig. 5C) performs worse than the aggregated scores (Fig. 5D), showing the utility of aggregating z scores into pathways. The improvement between Fig 5C and 5D specifically shows the utility of mapping the gene expression values into features rather than just subsetting the genes involved in the pathways as features directly.

Finally, the neural network regression model outperformed training on genes alone, but performed worse than either of our pathway models. This suggests that, at least compared to a shallow and computationally efficient neural net, biologically inspired feature extraction works better than "automated" learning of features through neurons. Moreover, explicit feature extraction through

pathway scores retains biological interpretability, which is lost in training the neural net.

C. Feature Importance of Pathways in OLS Regression Models

We extracted the most important features in the OLS regression models trained on pathway enrichment scores and aggregate pathway z-scores. Within the top 25 features of both models, ranked by absolute value of the coefficient, several pathways with known contributions to senescence and aging appear, including "regulation of reactive oxygen species metabolic processes", "autophagosome assembly", "regular of TORC1 signaling", and "interleukin beta-1 production". These pathways are significant as they involve the major hallmarks of aging curated by biologists.[14]

VI. Conclusion and Future Work

In this paper, we have shown that regression models trained on scores representing biological pathway activity outperform models trained on gene expression data for individual genes as features. This work has several key implications. First, pathway score algorithms represent a viable feature map to reduce the dimensionality of high-dimensional gene expression data while preserving patterns of variation relevant to aging. In the context of the previous application of these methods to binary classification tasks (disease prediction), it appears likely that pathway scores can broadly enhance the predictive power of machine learning algorithms in bioinformatics while retaining biological interpretability. From a biological perspective, our work is significant in that it supports the hypothesis that aging occurs primarily at the level of pathway (as opposed to individual gene) dysregulation.

A major future direction of the work includes expanding the data sources that we train our models on. We currently used a dataset of transcriptomics of fibroblasts; however, the gene expression signatures of aging likely vary from cell type to cell type. Therefore, in the future, we would like to train different models by cell type to understand cell-specific determinants of aging. Additionally, we could use a stacked regression to combine cell-specific models and compare performance against models trained on fibroblasts. Moreover, the pathways-based approaches can be applied to characterize similarities and differences in the determinants of aging across species. We plan on applying our approach to mice and comparing performance to models on humans as a potential way to understand common aging pathways between mice and humans.

Contributions

Rishabh and Yash pair-coded each element of the project and contributed equally to the data preprocessing, model building, and model evaluation. Rishabh and Yash wrote the paper together. Yash focused background research on machine learning methods in bioinformatics, while Rishabh focused on biological underpinnings of pathway dysregulation in aging.

Code and Data Availability

Data and code available on GitHub.

References

[1] Oxford Handbook of Evolutionary Medicine

[2] Van Houtven G, Honeycutt AA, Gilman B, et al. Costs of Illness Among Older Adults: An Analysis of Six Major Health Conditions with Significant Environmental Risk Factors [Internet]. Research Triangle Park (NC): RTI Press; 2008 Sep. Available from: https://www.ncbi.nlm.nih.gov/books/NBK532461/ doi: 10.3768/rtipress.2008.rr.0002.0809

[3] Fleischer, J.G., Schulte, R., Tsai, H.H. et al. Predicting age from the transcriptome of human dermal fibroblasts. Genome Biol 19, 221 (2018). https://doi.org/10.1186/s13059-018-1599-6

[4] Lee E, Chuang HY, Kim JW, Ideker T, Lee D. Inferring pathway activity toward precise disease classification. PLoS Comput Biol. 2008 Nov;4(11):e1000217. doi: 10.1371/journal.pcbi.1000217. Epub 2008 Nov 7. PMID: 18989396; PMCID: PMC2563693.

[5] Han L, Maciejewski M, et al, A probabilistic pathway score (PROPS) for classification with applications to inflammatory bowel disease, Bioinformatics, Volume 34, Issue 6, 15 March 2018, Pages 985–993, https://doi.org/10.1093/bioinformatics/btx651

[6] Pershad Y, Guo M, Altman RB. Pathway and network embedding methods for prioritizing psychiatric drugs. Pac Symp Biocomput. 2020;25:671-682. PMID: 31797637; PMCID: PMC6951442.

[7] Hill DP, Smith B, McAndrews-Hill MS, Blake JA. Gene Ontology annotations: what they mean and where they come from. BMC Bioinformatics. 2008 Apr 29;9 Suppl 5(Suppl 5):S2. doi: 10.1186/1471-2105-9-S5-S2. PMID: 18460184; PMCID: PMC2367625.

[8] Zhang Y, Ma Y, Huang Y, Zhang Y, Jiang Q, Zhou M, Su J. Benchmarking algorithms for pathway activity transformation of single-cell RNA-seq data. Comput Struct Biotechnol J. 2020 Oct 15;18:2953-2961. doi: 10.1016/j.csbj.2020.10.007. PMID: 33209207; PMCID: PMC7642725.

[9] Ng, Andrew. Regularization and model selection. CS229 Lecture Notes. Spring 2021.

[10] UniProt Gene ID Mapping Tool. https://www.uniprot.org/uploadlists/

[11] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005 Oct 25;102(43):15545-50. doi: 10.1073/pnas.0506580102. Epub 2005 Sep 30. PMID: 16199517; PMCID: PMC1239896.

[12] Ng, Andrew. Principal component analysis. CS229 Lecture Notes. Spring 2021.

[13] Pedregosa F, Michel V, Grisel O, Blondel M, Prettenhofer P, Weiss R, et al. Scikit-learn: Machine Learning in Python. Vol. 12, Journal of Machine Learning Research. 2011.

[14] López-Otín C, Blasco MA, Partridge L, Serrano M, Kroemer G. The hallmarks of aging. Cell. 2013;153(6):1194-1217. doi:10.1016/j.cell.2013.05.039