

Explainable Graphical Model-Based Transcriptomic Clock from Single Cell Gene Expression Data

Dániel Sándor , Péter Antal 

Budapest University of Technology and Economics
Department of Artificial Intelligence and Systems Engineering
Budapest, Hungary
Email: {sandor, antal}@mit.bme.hu

Abstract—Predicting the biological age of an organism is a critical task, and multiple accurate models currently exist. However, to extend both lifespan and more critically, healthspan, it is essential to identify points of intervention and understand how to address them. Effective treatments, whether through medication or other methods, require the ability to diagnose the underlying causes of reduced lifespan, like the up or down-regulation of certain genes.

In this study, we evaluate the performance of a modern structure-learning algorithm in constructing a graphical model-based clock for predicting and analyzing biological age. Using a single-cell gene expression atlas, we develop a probabilistic graphical model capable not only of predicting biological age but also of identifying potential gene-level interventions. We compare the performance of various Bayesian and non-Bayesian algorithms for age prediction, assessing them based on their predictive accuracy and ability to elucidate complex biological processes associated with aging in *Caenorhabditis elegans*.

Index Terms—structure learning, probabilistic graphical models, biological clock, gene expression

I. INTRODUCTION

Biological aging clocks predict the biological age of an organism, which is a measurement of health [22]. They usually measure cellular aging within the organism. Biological age differs from chronological age, however it is strongly correlated. This metric is influenced by a multitude of factors, like diet, exercise, or environmental stress. Of course, measurement of biological age is a difficult task, thus we try to model it by the condition of the organism. It is an important measurement because if we can get an accurate picture of the condition, we might gain insights into the possible interventions or treatments.

Many different methods have emerged to predict the biological age, primarily but not necessarily using molecular level data, e.g. epigenetics, telomere length, or transcriptomic predictors [3]. However, non-molecular approaches are also available [8], and if they can be combined we can hope to get a more accurate descriptor (for recent reviews see: [23], [24]). To combine aspects we need models that support multimodal transfer learning to efficiently learn across domains. In this

PROJECT NO. 2024-2.1.2-EKÖP-KDP-2024-00005 HAS BEEN IMPLEMENTED WITH THE SUPPORT PROVIDED BY THE MINISTRY OF CULTURE AND INNOVATION OF HUNGARY FROM THE NATIONAL RESEARCH, DEVELOPMENT AND INNOVATION FUND, FINANCED UNDER THE EKÖP_KDP-24-1-BME-15 FUNDING SCHEME.

study we focus on the most widely used transcriptomic clocks [1], [4] that use RNA-seq data, to model biological age based on the expression levels of genes.

We find that most state-of-the-art methods using predictive approaches are not explainable and are hard to validate with experiments [4], [7], [8]. We attempt to construct a Probabilistic Graphical Model of aging in *Caenorhabditis elegans*, to create an explainable model, which also has predictive capabilities using Bayesian inference. The choice of *c. elegans* was made, because of the extensive data available on the organism including cell type-specific aging patterns [1]. Thus validation of the learned biological clocks becomes easier.

Our main contributions are the following:

- We create a biological clock, that has predictive value on novel, established data from single-cell RNA-seq experiments on *c. elegans*.
- We demonstrate the power of structure learning using scalable structure learning of Directed Acyclic Graphs.
- We validate the explainable model based on existing findings on *c. elegans* aging.

The structure of the paper is the following: In section II we present the related works of structure learning and biological clocks. In section III we present the methodologies of data and model, with an emphasis on the structure learning algorithm. In section IV we discuss our findings on the explainable network-based transcriptomic clocks. We end in section V with a conclusion on the usability of biological clocks.

II. RELATED WORK

Recently, multiple aging atlases have been created for many different organisms [1], [5]–[7]. In all these studies, subjects' genes are sequenced and examined in different age groups. Different studies highlight certain aspects, but many have claimed to identify relevant genes that are up or down-regulated with time. Thus they can be associated with the process of aging.

A. Transcriptomic markers of aging

Roux et al. [1] find that aging in *c. elegans* depends heavily on the cell type examined, as different cells differ in their stress signature. They find that *gei-3*, a gene responsible for enabling DNA binding transcription factors, is upregulated with age in a broad range of cell types, thus they conclude that it is

universally associated with biological age. de Magalhaes et al. [7] examined human genes, using mice as model organisms, and created a network of aging based on associated genes related to DNA metabolism. Aging Atlas [5] is a collection of multiple genetic and other biomarkers related to aging, its notable RNA-seq module is responsible for the findings on the human FOXO3 and CLOCK genes, both of which change significantly with aging. However, these markers are only predicted by associations, and so far few causal approaches exist to find relevant regulators in the aging process.

B. Biological clocks

Biological clocks measure the aging of the organism with mixed predictive performance and use a wide array of modalities [2], [3], [8], [21]. In addition to RNA-seq, the most prominent predictors of biological age are epigenetic markers [2], [10]. The length of the telomere is also mentioned in models with high predictive performance [9]. But extreme cases also exist, such as the prediction of biological age from photos [8]. Some promising approaches based on DNA methylation include XAI-AGE [21], where Prosz et al. have developed an explainable clock with predictive performance comparable to state-of-the-art black-box models. Transcriptomic clocks differ from most of these, in the sense that they can measure variables, that can directly influence aging through gene expression levels.

C. Transcriptomic clocks

Shokhirev and Johnson [11] model multiple clocks based on different markers, using regression techniques to weigh the effect of each gene. This creates extremely accurate clocks with an R^2 value of 0.96. The study demonstrates that simple linear models are often enough to predict biological age; however, this result also relies on the association of genes with age, without regarding causality. Holzschek et al. [12] use neural networks to predict age, with a smaller dataset, however, they manage to incorporate a priori knowledge in their system. In humans, they claim to achieve a mean absolute error of 4.7 years. Our most important comparison is BiT age [4], which is based on the Elastic Net Regression, which also achieved an R^2 value of 0.96, where they claim that it is close to the limit of predictive performance.

D. Network-based modeling

While explainable models do exist in the field of biological aging clocks, these are mostly focused on the post-hoc explainability of blackbox models [13], [21]. To use graphical model-based solutions we find methods in the field of causal discovery. The goal of causal discovery is to learn the structure of independences between our variables. With the help of this, we can construct a regulatory network of genes, where we can incorporate age as a variable, that genes can influence. An important point to make is that we are using a prior, where age is determined from the genetic patterns, thus age cannot have outgoing edges, as it would violate this principle i.e. processes in the organism cause biological age, and biological age cannot

be the cause of genes being expressed differently. However modeling a DAG still makes sense, as necessary interventions can be made at different variables with varying effects.

Learning gene regulatory networks by causal discovery has proven to work in multiple scenarios [14]–[16]. The most prominent real benchmark dataset is the Sachs protein signaling [17] in the field. We present two methods Generative flow networks (GFlow) [19] and BayesDAG [18], both of which are able to solve the Sachs dataset with acceptable accuracy. The choice of Bayesian methods is necessary if we want to conduct further analysis. By sampling directly from the posterior, we can find certain and uncertain parts of our networks, and improve them with prior knowledge.

III. METHODS

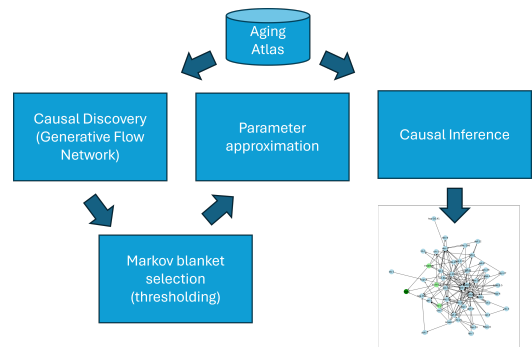


Fig. 1: Overview of methodologies.

We perform end-to-end examination of the aging atlas from Roux et al. [1]. Our primary goal is not to achieve the highest predictive accuracy of age. However, predictive power is an important requirement for the model. We want to draw causal conclusions from the model, that is why we use Causal Discovery and Causal Inference along the process. To tighten the scope we do not analyze the whole dataset, but focus on its most visible changes (the aging of amphid neurons). An overview of the process can be seen in Fig. 1

A. Aging atlas

Our goal was to demonstrate that causal prediction can be drawn from the collected data. We examined the results of the aging atlas and tried to narrow the search, for parts where meaningful conclusions can still be drawn but the number of variables is small enough, that learning a Generative Flow Network over DAG space was still feasible. The training of the network was conducted on the Single-cell Variational Inference transformed 'denoised' dataset.

B. Gene selection

The original paper mentions more than 4000 genes by name, and that is too large in the state space, for most DAG learning strategies to be applicable. As we wanted to prove the most important results of the paper we curated the list of genes to the 50 we deemed most important. The complete list of genes and the reasoning behind their choices can be found in

Appendix A. The most important principles in choice were the following: We chose known indicators of healthspan in *c. elegans*, like *skn-1*, *daf-16*, *hlh-30*, *fkh-7*, *fkh-9*, *daf-12*, *dpy-27*, *gei-3*. We included the *hsp-70* group (*hsp-70*, *hsp-16.41*, *F44E5.4*, *F44E5.5*, *hsp-110*) as they are mentioned to be important cytosolic chaperones, which are upregulated in neurons. Ultimately we included every gene that is described in detail and affects the aging process. Thus our goal is not the selection of relevant genes regulating aging, but finding the best predictors and explaining their relationships.

C. Amphid neurons

Amphid neurons were key samples in the learning process. Roux et al. find that they exhibit the most changes with aging from all cell clusters of *c. elegans*. To denoise the dataset we filtered for the samples from amphid neurons, thus making it easier for an estimator to find the connections between variables. Furthermore, all genes mentioned to be up or downregulated in these cells are included in the gene set.

D. Generative flow network

We use the Generative Flow Network [19] as a base estimator for the DAG posterior. The method uses an iterative approach to building a DAG by adding edges and measuring the probabilities of DAG states by the flow going through them in each state. Flow is defined over the graph of DAGs, where the amount of flow going through a DAG state is proportional to the state's posterior probability. We sample the DAG posterior to get an accurate distribution, on which we can learn the parameters. In the cases where point estimates are given, the estimate is calculated by averaging 64 samples from the posterior, in no cases were the acyclicity constraint violated by averaging.

E. Selecting the Markov Blanket

The Markov Blanket of age is equivalent to its parents in this case, as the prior forbids any outgoing edges from the variable. The only relevant question is where to threshold the edge probabilities to get a desirable Markov Blanket set (MBS). The thresholding is done on the posterior probabilities of the edges in the result of the GFlow algorithm. The desirable MBS is non-empty and contains less than the total number of variables, for explainability and predictive performance.

In most of our experiments, the threshold was set to 0.5, as this contained all edges that were more likely to be in the DAG than not, based on the edge posteriors. To include more variables we conducted experiments with MBS threshold of 0.46 (a threshold determined by the iterative lowering until we got the desirable amount of variables), however, these do not seem to enhance the predictive probabilities of the models. In cases where we trained the model on the full dataset, the threshold is set to 0.6 to account for the noise introduced with the new samples.

F. Parameter learning

On the thresholded MBS we conduct parameter learning, to create a Bayesian Network. In most cases it is simple as the target variable has a small number of parents in the BN, thus the conditional probability tables (CPD) will in turn also be small. As most approaches only support discrete values, for the parameter learning we discretize the parents of the age (which is discrete in the original dataset, into quartiles). We use maximum likelihood estimation to find the best parameter for each value of the CPD.

G. Causal inference

In practice the inference is simple [20]. For each sample of the test dataset, we take the discretized version and predict age, by taking the weighted average of the resulting distributions. If we do not care about point estimates we can generate the posterior over the age variable, which is a more meaningful measure, as this incorporates any uncertainties the model might carry.

IV. RESULTS

In this section, we discuss our findings. The two setups for experimentation only differed in the number of samples used: In the full data setup we used 37938 samples to train our structure and parameter models, and in the amphid neurons setup we used 79 samples for training.

A. Results of structure learning

In the full data case a more dense DAG was found after averaging the models, however, to get a clearer picture we used a higher threshold of 0.6 the resulting DAG structure can be seen in Fig. 3. To validate the structure we turn to the Aging Atlas. Roux et al. report, that *lin-1*, which is one of the parents of age in the DAG, is a transcription factor that is associated with increased movement and healthspan, thus it makes sense that it affects biological age, however, it is not the most indicative of lifespan according to Roux et al. The other parent of age is *F46A8.13*, which is one of the top-ranked genes, that is upregulated across cell types in aging, however, interestingly this is only in the 17th place in the dataset. Although not all genes in the ranking are covered by our examination, this might suggest that further refinement of the method is required.

In the case of the amphid neurons, the threshold can be left at 0.5, this keeps only edges, that most of the DAGs in the sampled posterior contain. The point estimate of the DAG can be seen in Fig. 4 In this case the only parent of the age variable is *F44E5.4*, which is one of the cytosolic chaperones from the *hsp-70* family, responsible for protein refolding. Its parent is *hsp-16.2*, also from the same family. Roux et al. show, that these genes are heavily upregulated with age in all neuron classes, but especially in amphid neurons. Further examination of the *hsp-70* family's regulatory network might be necessary to prove the regulatory relationships between them and aging, but right now we can show, that this family influences the cell type's aging, as we can see from the predictive performances as well.

B. Predictive performance

When it comes to the accuracy of the clocks, the best practice we can do to evaluate them is to examine their correlations with chronological age. We do not expect them to be fully identical, as biological age is influenced by a multitude of factors, but this is the best indicator for the performance of aging clocks. To compare these we share the R^2 value for each clock, and compare them to existing approaches.

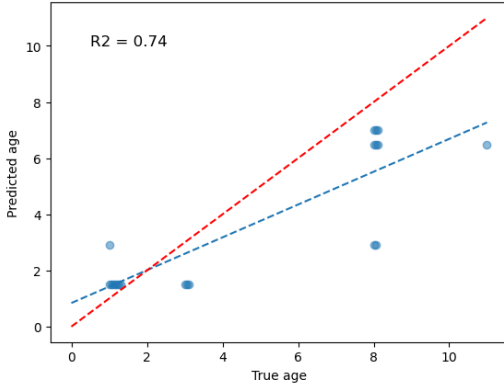


Fig. 2: Correlation of predicted biological age and true chronological age in amphid neurons.

For the prediction of the age variable we use causal inference, and to make it efficient we only include evidence belonging to the Markov blanket of the variable. In each case, we discretized the test data, based on the quartiles of training data, and predicted on the test set. The test data in all cases is a randomized 10% of the relevant dataset. We present the values in Table I

TABLE I: R^2 values of different clocks

Method	R^2
Full Dataset GFlow + MLE	0.46
Amphid Neurons GFlow + MLE	0.74
BiT age reported	0.96

We can see that all our methods significantly correlate with the chronological age variable, which is a remarkable result if we consider the discrete dataset that was used for parameter learning. From our clocks, the amphid neuron case performs best, with the 0.5 threshold, increasing or decreasing the threshold all yield worse results, as more variables introduced more noise in the inference. We can see the model’s performance on the whole test set in Fig. 2. We find that most test samples are predicted to be younger than their chronological age, but potentially all this difference can be explained by the discretization of the test dataset, and the uncertainty introduced by the small size of the dataset.

V. CONCLUSION

To summarize we presented the idea of an explainable network-based aging clock based on transcriptomic data. We

created the clock, which although lags in predictive power from the theoretical limit, is widely usable for causal discovery and further research in aging-related questions. Both the found genes (F44E5.4, lin-1, F46A8.13) are good predictors of aging, and are worthy of further genetic experiment-based examination, together with the hsp-70 family. It is also mentioned by Roux et al. and confirmed by our finding of the structure of the gene regulatory network.

APPENDIX A LIST OF GENES

The complete list of genes for the construction of the clocks is the following: *gei-3*, *daf-16*, *skn-1*, *hlh-30*, *fkh-7*, *fkh-9*, *daf-12*, *nhr-23*, *nhr-25*, *dpy-27*, *hsp-70*, *hsp-16.41*, *hsp-16.2*, *hsp-4*, *cup-2*, *pdi-6*, *hsp-3*, *hrdl-1*, *cnx-1*, *crt-1*, *ctc-3*, *nduo-6*, *pck-1*, *gpd-2*, *tpi-1*, *rpl-3*, *rps-9*, *rps-28*, *mec-12*, *let-607*, *stc-1*, *clec-166*, *tag-353*, *unc-73*, *flp-24*, *ceh-36*, *crh-1*, *daf-19*, *F44E5.4*, *F44E5.5*, *F46A8.13*, *F57F5.1*, *Y94H6A.10*, *xbp-1*, *nmad-1*, *efl-2*, *elt-7*, *lin-1*, *hsp-110*

REFERENCES

- [1] Roux, A. E., Yuan, H., Podshivalova, K., Hendrickson, D., Kerr, R., Kenyon, C., & Kelley, D. (2023). Individual cell types in *C. elegans* age differently and activate distinct cell-protective responses. *Cell Reports*, 42(8).
- [2] Poganič, J. R., Zhang, B., Baht, G. S., Tyshkovskiy, A., Deik, A., Kerepesi, C., ... & Gladyshev, V. N. (2023). Biological age is increased by stress and restored upon recovery. *Cell Metabolism*, 35(5), 807-820.
- [3] Jylhävä, J., Pedersen, N. L., & Hägg, S. (2017). Biological age predictors. *EBioMedicine*, 21, 29-36.
- [4] Meyer, D. H., & Schumacher, B. (2021). BiT age: A transcriptome-based aging clock near the theoretical limit of accuracy. *Aging cell*, 20(3), e13320.
- [5] Aging Atlas: a multi-omics database for aging biology. *Nucleic acids research*, 2021, 49.D1: D825-D830.
- [6] Kedlian, V. R., Wang, Y., Liu, T., Chen, X., Bolt, L., Tudor, C., ... & Zhang, H. (2024). Human skeletal muscle aging atlas. *Nature aging*, 1-18.
- [7] de Magalhaes, J. P., & Toussaint, O. (2004). GenAge: a genomic and proteomic network map of human ageing. *FEBS letters*, 571(1-3), 243-247.
- [8] Zalay, O., Bontempi, D., Bitterman, D. S., Birkbak, N., Shyr, D., Haug, F., ... & Aerts, H. J. (2023). Decoding biological age from face photographs using deep learning. *medRxiv*.
- [9] Vaiserman, A., & Krasniakov, D. (2021). Telomere length as a marker of biological age: state-of-the-art, open issues, and future perspectives. *Frontiers in genetics*, 11, 630186.
- [10] Horvath, S., & Raj, K. (2018). DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nature reviews genetics*, 19(6), 371-384.
- [11] Shokhirev, M. N., & Johnson, A. A. (2021). Modeling the human aging transcriptome across tissues, health status, and sex. *Aging cell*, 20(1), e13280.
- [12] Holzschek, N., Falckenhayn, C., Söhle, J., Kristof, B., Siegner, R., Werner, A., ... & Kaderali, L. (2021). Modeling transcriptomic age using knowledge-primed artificial neural networks. *npj Aging and Mechanisms of Disease*, 7(1), 15.
- [13] Qiu, W., Chen, H., Kaerberlein, M., & Lee, S. I. (2023). ExplainABLE BioLogical Age (ENABL Age): an artificial intelligence framework for interpretable biological age. *The Lancet Healthy Longevity*, 4(12), e711-e723.
- [14] Feng, K., Jiang, H., Yin, C., & Sun, H. (2023). Gene regulatory network inference based on causal discovery integrating with graph neural network. *Quantitative Biology*, 11(4), 434-450.
- [15] Foraita, R., Friemel, J., Günther, K., Behrens, T., Bullerdiek, J., Nimzyk, R., ... & Didelez, V. (2020). Causal discovery of gene regulation with incomplete data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 183(4), 1747-1775.

- [16] Sándor D. & Antal P. (2024). Systematic evaluation of continuous optimization approaches for causal discovery of gene regulatory networks. *Proceeding of 31th Minisymposium of the Department of Measurement and Information Systems*, 6, 50-54.
- [17] Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., & Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721), 523-529.
- [18] Annadani, Y., Pawlowski, N., Jennings, J., Bauer, S., Zhang, C., & Gong, W. (2023). Bayesdag: Gradient-based posterior inference for causal discovery. *Advances in Neural Information Processing Systems*, 36, 1738-1763.
- [19] Deleu, T., Góis, A., Emezue, C., Rankawat, M., Lacoste-Julien, S., Bauer, S., & Bengio, Y. (2022, August). Bayesian structure learning with generative flow networks. In *Uncertainty in Artificial Intelligence* (pp. 518-528). PMLR.
- [20] Pearl, J. (2010). Causal inference. *Causality: objectives and assessment*, 39-58.
- [21] Prosz, A., Pipek, O., Börsök, J., Palla, G., Szallasi, Z., Spisak, S., & Csabai, I. (2024). Biologically informed deep learning for explainable epigenetic clocks. *Scientific Reports*, 14(1), 1306.
- [22] Levine, M. E. (2013). Modeling the rate of senescence: can estimated biological age predict mortality more accurately than chronological age?. *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences*, 68(6), 667-674.
- [23] Li, Z., Zhang, W., Duan, Y., Niu, Y., Chen, Y., Liu, X., ... & Chen, X. (2023). Progress in biological age research. *Frontiers in public health*, 11, 1074274.
- [24] Bafei, S. E. C., & Shen, C. (2023). Biomarkers selection and mathematical modeling in biological age estimation. *npj Aging*, 9(1), 13.