Psuedotemporal Analysis of Genetic Regulatory Networks in Neurodegenerative Disorders

Dhamanpreet Kaur, Sebastian Pineda, Samuel Sledzieski

October 23, 2019

1 Introduction

Gene regulatory networks (GRNs) govern the expression of genes via the underlying relationships between transcription factors and signalling molecules. Understanding these networks holds great clinical relevance as it provides insight into effective targets for treating a disease. Substantial work has been done in using computational methods to predict these networks from large-scale transcriptome data [1, 2, 3, 4]. While this data has provided significant information pertaining to gene regulation, these bulk technologies are limited in that they take the average across all cell types in a sample [5]. Single-cell RNA sequencing (scRNA-seq) technology provides the gene expression levels of an individual cell, which would theoretically show more precise gene regulator interactions. The application of network prediction methods remains underdeveloped for the fairly recent advancement of single-cell RNA seq data, as prior studies have used approaches adapted from techniques used on bulk data.

Additionally, genetic regulatory networks have an important time element. While many methods for reconstruction of GRNs infer undirected networks, the direction and weight of these edges have a meaningful biological analogue in the regulatory impacts that genes and their protein products have on each other across time. GRN models built from dynamic processes typically involve methods such as Boolean networks [6, 7], Dynamic Bayesian Networks [1], co-expression networks [8], ordinary differential equations [9], and variational Bayesian inference [10]. A handful of recent studies have developed approaches that first order the cells along a temporal trajectory, using time or pseudotime data to infer GRNs [10, 11]. However, these applications have each used samples explicitly taken at different time points, or have used pseudotime analyses from the area of developmental biology.

This study seeks to explore the application of Dynamic Bayesian networks (DBNs), which have been used with bulk data [12], to infer patterns of gene regulation in neurodegenerative diseases. By using DBNs, it is possible to model the inherently time-dependent regulatory interactions between genes. We will develop a novel pseudotime ordering framework to train a DBN using scRNA-seq data from fully developed brain samples, and apply this framework to data sets from Alzheimer's Disease (AD) and Huntington's Disease (HD) in an attempt to elucidate the regulatory mechanisms behind neurodegenerative disease.

Focusing on AD and HD provides us with the opportunity to study GRNs linked to neurodegeneration. As a well-studied monogenic disease, many of the GRNs perturbed in HD have been characterized, at least at the bulk level, whereas the same cannot be said of AD, a polygenic disease where the pathogenic cause is unknown, where these same mechanisms remain elusive. Despite these differences, it is known that this diseases share many similarities at the molecular level and the ultimately lead to the same fatal consequence. Thus we anticipate that our approach should be applicable to the study of both conditions, and elucidating the regulatory perturbations of one disease will enable us to shed light on the same or similar molecular mechanisms in the other.

2 Specific Aims

Indicate what you're setting out to do and why that step is important; we should include why each step matters Do not make aims linear/sequential

Aim 1: Separate single cell data into states and order states along pseudotime using markers for disease progression. Apply a network-based archetypal analysis approach to cluster and annotate cells from single-cell RNA-seq data. Perform cell-type specific differential expression analysis to identify genes representative of disease progression. Use canonical cell type markers to identify and annotate key cell states or "archetypes" representative of all cells of that type in said state. For the archetypes of each cell type, order the archetypes in psueodotime.

Aim 2: Reconstruct dynamic gene regulatory networks by applying a Dynamic Bayesian Network approach. Using the discretized cell states as pseudo time points, the Bayesian network creates layers to show the temporal progression of the disease and models the dependencies of gene as nodes. The weights of the directed edges from one layer to the next indicates the strength of the activation and inhibition of genes.

Aim 3: Apply graph theoretic techniques to identify regulatory elements correlated with disease progression. Identify high-centrality nodes using standard graph theory measures such as degree centrality, betweenness centrality, or page-rank, and compare with canonical genetic markers for disease progression. Identify potential new disease driver genes based on centrality. Validate edges using gene ontology molecular signatures.

3 Research Strategy

3.A Significance

Even with significant research effort in recent years, the mechanisms and cellular pathways that underlie neurodegenerative disorders are not fully understood. Gene regulatory networks model one of many such mechanisms - the relationships (activation and inhibition) of gene expression by the protein products of other genes. Understanding the way that genes regulate each other can lead to valuable biological insight, including the identification of important pathways in disease progression or driver genes for disease. This in turn can lead to the discovery of new drug targets or treatment methods.

Furthermore, the consolidation of pseudo-temporal ordering for single-cell RNA seq data with a Dynamic Bayesian network may be a novel method for predicting gene regulatory networks. The robust development of this could lead to its application in other diseases and advance our ability to understand widespread gene regulation.

To date, essentially all research into GRNs has been performed on bulk-level RNA-seq data, which does not permit the dissection of cell-type specific regulatory networks. Further, cell-type specific vulnerability and dysregulation has been implicated in virtually all neurodegenerative diseases, meaning that bulk-level analysis provides little useful insight into the underlying pathological mechanisms. Additionally, single-cell studies of such diseases has been limited to transcriptional profiling at a single time-point, which again fails to provide a picture of the critical dynamic processes and molecular mechanisms that drive disease progression. The pseudotemporal study of GRNs using single-cell transciptomic data could potentially address these historic shortfalls in these areas of research.

3.B Innovation

The study of genetic regulatory networks has primarily utilized bulk RNA-seq data, which blurs the signal from genes which are differentially expressed in individual cells. This differential expression can be caused by a number of factors including the cell cycle, cell state, or disease progression. Single-cell RNA sequencing captures gene expression levels for single cells, which provides a higher-resolution picture of gene regulatory networks in an organism as a whole. Additionally, the effect of disease progression on gene regulation has not been well studied.

Our study divides single cell expression profiles into archetypes and performs a novel pseudo-temporal ordering of the cell archetypes based on disease progression. This allows us to train a Dynamic Bayesian Network model by using the gene expression at each pseudo-time point as a layer of the DBN. By constructing a dynamic gene regulatory network, we will perform the first characterization to our knowledge of regulatory networks in neurodegenerative disorders which incorporates disease progression information. We will do so in a cell-type specific manner, which accounts for the differential expression profiles of different cell types and identifies variance in genetic regulation across cell types.

3.C Approach

should include: rationale, plan, expected outcome, potential challenges and solutions (this last one esp needs more work)

3.C.1 Aim 1

We will annotate cell types and cluster cells into archetypes using the ACTIONet framework [13]. Important differentially expressed genes (DEGs) for tracking disease progression will be discovered de novo by performing differential gene expression analysis on our cohort of disease and unaffected individuals. From this, we will be able to build a canonical gene expression profile for disease, both for Alzheimer's Disease and Huntington's Disease. We can build a similar profile for healthy individuals. These expression profiles can be compared with other previously reported DEGs, such as those in Mathys et al. [14] For each archetype, we will compute the median expression profile using the expression levels of all single cells in the archetype. Archetypes can then be ordered along the spectrum from healthy to disease (again in a cell-type specific manner), providing a pseudotime estimation for disease progression cell type based on their distance from the canonical profile. This ordering of archetypes provides a framework within which to learn the GRN.

3.C.2 Aim 2

We will construct a Dynamic Bayesian network in which each node corresponds to a gene regulator and the directed edge between two nodes indicates an activation or inhibition. The pseudo time states of the single cell archetypes will be used as layers of the Bayesian network such that edges can only be directed in line with the sequence of disease progression. Learning the Bayesian network is a two-part process: structure learning and parameter learning. We will explore both constraint-based and hybrid approaches for structure learning. Moreover, the benefit of Bayesian Networks is that while they can operate as an unsupervised learning process, they allow for user input of prior knowledge and build the network given those constraints. We will employ this feature to strengthen the model by adding in known regulatory pathways.

We will verify the ability of model to accurately predict known gene regulatory networks before proceeding to offer new insights. Given the network is generated based on training data, we can further validate the network based on the ability of machine learning methods to distinguish between synthetic data simulated from network and testing data.

3.C.3 Aim 3

We can perform verification of inference and novel biological analysis by applying methods from graph theory to analyze the inferred gene regulatory network. Genes which play a significant regulatory role in disease progression can be identified by their centrality within the network. Centrality of a gene can be quantified in a number of ways, including *degree centrality*, the number of edges incident on a node, *betweenness centrality*, a measure of the importance of a node in information flow through the network, and *page-rank* centrality [15], which weights a nodes importance based on the importance of nodes which point to it.

After identifying genes with high centrality, we can cross-reference these with AD and HD associated molecular signatures from the Molecular Signatures database (MSigDB) [16]. We can also perform a more rigorous validation of our network using the methods described in Iacono et al. [17], which uses MSigDB to compute a p-score for each inferred edge.

4 Resources

We will use internal single-nuclear RNA-seq data from 48 Alzheimer's disease and unaffected control samples from the ROSMAP study of varying age, gender, and pathological progression [18]. We will also use internal single-nuclear RNA-seq data from 6 Huntington's disease and unaffected controls from the NIH NeuroBioBank of varying pathological grades. Validation of central genes and important gene pathways will be done with pathway and gene ontology (GO) gene sets from MSigDB [16, 19].

Computation will be performed on our personal machines, and using the Broad Institute's compute infrastructure when necessary. The data analysis will be performed primarily in the R and Python programming languages.

We will draw on content from Lectures 6 (Expression Analysis), 10 (Regulatory Networks), 11 (Network Structure), and 21 (Single-Cell Genomics) in completing our project. Additionally, we will be mentored by Shahin Mohammadi from the Kellis Lab.

5 Collaboration

We will meet regularly to discuss project progress updates and any necessary adjustments made to the original plan. Sebastian will focus on processing of scRNA-seq data, differential expression analysis and cell type annotation, and biological application of the reconstructed networks and central genes. Sam will focus on cell type annotation, archetype clustering and ordering, and network analysis using graph theory tools. Dhaman will focus on the statistical inference and training of the DBN, as well as validation of the reconstructed network through simulation and benchmarking. We plan to all work collaboratively and share expertise as appropriate.

6 Timeline

Week	**Deadline**	**Task**
14 Oct	17 Oct - Proposal Due	Solidify project outline and identify data sets
21 Oct	-	Begin DE Analysis with AD data
28 Oct	28 Oct - Proposal Reviews	Annotate cell types and identify single cell archetypes
4 Nov	7 Nov - Proposal Response	Construct pseudotime ordering of archetypes
11 Nov	-	Learn networks with DBN, Begin work with HD data
18 Nov	-	Begin graph theory analysis, refine DBN training
25 Nov	25 Nov - Midcourse Report	Finalize AD analysis, apply to HD
2 Dec	8 Dec - Final Report	Write report and wrap up final tests
9 Dec	10 Dec - Final Presentations	Finish presentation

References

[1] P. Li, C. Zhang, E. J. Perkins, P. Gong, and Y. Deng, "Comparison of probabilistic boolean network and dynamic bayesian network approaches for inferring gene regulatory networks," in *BMC bioinformatics*, vol. 8, p. S13, BioMed Central, 2007.

[2] N. A. Barker, C. J. Myers, and H. Kuwahara, "Learning genetic regulatory network connectivity from time series data," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 8, no. 1, pp. 152–165, 2009.

[3] F. Liu, S.-W. Zhang, W.-F. Guo, Z.-G. Wei, and L. Chen, "Inference of gene regulatory network based on local bayesian networks," *PLoS computational biology*, vol. 12, no. 8, p. e1005024, 2016.

[4] B. Yang, W. Zhang, and J. Lv, “A new supervised learning for gene regulatory network inference with novel filtering method.,” *International Journal of Performability Engineering*, vol. 14, no. 5, 2018.

[5] M. W. E. J. Fiers, L. Minnoye, S. Aibar, C. Bravo González-Blas, Z. Kalender Atak, and S. Aerts, “Mapping gene regulatory networks from single-cell omics data,” *Briefings in Functional Genomics*, vol. 17, pp. 246–254, 01 2018.

[6] M. Imani and U. Braga-Neto, “Optimal gene regulatory network inference using the boolean kalman filter and multiple model adaptive estimation,” in *2015 49th Asilomar Conference on Signals, Systems and Computers*, pp. 423–427, Nov 2015.

[7] C. Y. Lim, H. Wang, S. Woodhouse, N. Piterman, L. Wernisch, J. Fisher, and B. Gottgens, “Btr: training asynchronous boolean models using single-cell expression data,” *BMC Bioinformatics*, vol. 17, 2016.

[8] A. T. Specht and J. Li, “Leap: constructing gene co-expression networks for single-cell rna-sequencing data using pseudotime ordering,” *Bioinformatics*, vol. 33, no. 5, 2016.

[9] H. Matsumoto, H. Kiryu, C. Furusawa, M. S. H. Ko, S. B. H. Ko, N. Gouda, T. Hayashi, and I. Nikaido, “SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation,” *Bioinformatics*, vol. 33, pp. 2314–2321, 04 2017.

[10] M. Sanchez-Castillo, D. Blanco, I. M. Tienda-Luna, M. C. Carrion, and Y. Huang, “A Bayesian framework for the inference of gene regulatory networks from time and pseudo-time series data,” *Bioinformatics*, vol. 34, pp. 964–970, 09 2017.

[11] L. F. Iglesias-Martinez, W. Kolch, and T. Santra, “BGRMI: A method for inferring gene regulatory networks from time-course gene expression data and its application in breast cancer research,” *Scientific reports*, vol. 6, p. 37140, 2016.

[12] B. Yu, J.-M. Zu, S. Li, C. Chen, C. Rui-Xin, W. Wang, Y. Zhang, and M.-H. Wang, “Inference of time-delayed gene regulatory networks based on dynamic bayesian network hybrid learning method,” *Oncotarget*, vol. 8, no. 46, 2017.

[13] S. Mohammadi, J. Davila-Velderrain, and M. Kellis, “A multiresolution framework to characterize single-cell state landscapes,” *bioRxiv*, 2019.

[14] H. Mathys, J. Davila-Velderrain, Z. Peng, F. Gao, S. Mohammadi, J. Z. Young, M. Menon, L. He, F. Abdurrob, X. Jiang, *et al.*, “Single-cell transcriptomic analysis of alzheimer’s disease,” *Nature*, p. 1, 2019.

[15] S. Brin and L. Page, “The anatomy of a large-scale hypertextual web search engine,” *Computer networks and ISDN systems*, vol. 30, no. 1-7, pp. 107–117, 1998.

[16] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, and J. P. Mesirov, “Molecular signatures database (MSigDB) 3.0,” *Bioinformatics*, vol. 27, pp. 1739–1740, 05 2011.

[17] G. Iacono, R. Massoni-Badosa, and H. Heyn, “Single-cell transcriptomics unveils gene regulatory network plasticity,” *Genome biology*, vol. 20, no. 1, p. 110, 2019.

[18] P. L. De Jager, Y. Ma, C. McCabe, J. Xu, B. N. Vardarajan, D. Felsky, H.-U. Klein, C. C. White, M. A. Peters, B. Lodgson, *et al.*, “A multi-omic atlas of the human frontal cortex for aging and alzheimer’s disease research,” *Scientific data*, vol. 5, p. 180142, 2018.

[19] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, “Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles,” *Proceedings of the National Academy of Sciences*, vol. 102, no. 43, pp. 15545–15550, 2005.