MACHINE LEARNING ASSISTED GENE EXPRESSION RESTORATION

Arnob Saha Ankon
1024052111

Arnab Bhattacharjee
1024052106

1 Introduction and Motivation

Gene expression can be used to determine whether a cell is healthy or diseased. This gene expression depends on the cell's state, which is influenced by its previous state. We can categorize the cell's state into the source state and the target state. The changes between the source and target states occur because of the activation and repression of some of the genes present in the cell. The identification of the genes that drive the transformation from the source state to the target state is important, as this transformation enables early detection of diseases like cancer and tumors. However, identifying these genes remains an unresolved challenge.

We can represent the complex relationships between genes and transcription factors that regulate and determine the behavior of the cell using a gene regulatory network (GRN) [2,3]. In GRNs, the nodes represent genes, and the edges represent the regulatory relationships between genes. Accurate identification of target genes requires knowledge of the GRN in both the source and target cell states. GRNs are extremely important as they help us understand the biological processes within the cell. One key application of GRNs is the comparison of healthy versus diseased tissues [4, 5]. Such studies help us identify genes that may be responsible for diseases.

Some approaches to inferring the relationship between the source state and target state include differential expression analysis [7], machine learning and deep learning-based perturbation [8] algorithmic approaches [1], or manual perturbation of specific transcription factors (TFs). In the paper GAGER (GRN Assisted Gene Expression Restoration) [1], the authors designed an algorithm to identify specific genes whose manipulation transforms the source state into the target state of the cell by comparing the GRNs of the source and target cell states. The algorithm mainly applies a series of perturbations to convert the source cell state into the target cell state. It aims to infer the differences between the regulatory edges of the GRNs of the source and target cell states. From this, it identify transcription factors that potentially change the cell state from source to target.

One limitation of their work is the assumption of a simple linear relationship between genes, while the relationship between genes may actually be more complex. They try to perturb TFs greedily to reduce the expression differences between the source and target cell states. We aim to explore gene expression restoration further by assuming that the relationships between genes are complex.

2 Proposed Methodology

1. **Inference of the Gene Regulatory Network (GRN):** We will use the tool SCENIC [6] to construct GRNs from the scRNA-seq dataset, as described in the paper [1]. If possible, we will also use the tool CellOracle [8] for the same task.

2. **Construction of the Difference Network:** We will construct the difference network using the algorithm outlined in GAGER [1].

3. **Estimation of Gene Expression and Regulatory Coefficients:** We will assume a complex, nonlinear relationship between genes rather than a simple linear relationship to calculate the expression and regulatory coefficients of the difference network using a neural network.

4. **Selection of Target Genes and Generation of GRNs:** We will select relevant transcription factors and genes, update their expression values to infer GRNs, and finally compare the produced results with the target cell state.

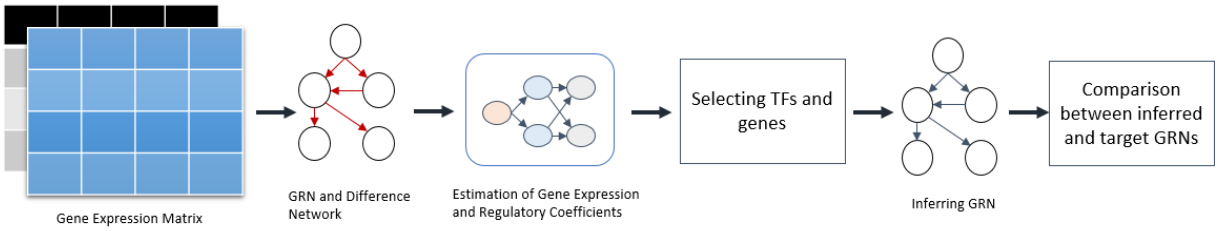

Figure 1: Overview of the proposed methodology: steps include inferring the Gene Regulatory Network (GRN) and the difference network, estimating gene expression and regulatory coefficients, selecting target genes, and comparing the inferred GRN with the target GRN.

3 Datasets

We will experiment on the following datasets:

1. **Saccharomyces cerevisiae RNA-seq data:** Gene expression data from Omnibus GSM3564448 (link) and Omnibus GSM4039308 (link).

2. **Hypoplastic Left Heart Syndrome (HLHS) RNA-seq data:** Gene expression data from Omnibus GSE146341 (link).

3. **Progression of Neurodegeneration in Mouse Model RNA-seq data:** Gene expression data from Omnibus GSE103334 (link).

4 Evaluation

1. We will compare the estimated gene expression and regulatory coefficients between our approach and the method in GAGER [1].

2. We will compare the lists of selected TFs and genes between our method and the one in GAGER, identifying similarities and differences.

3. The performance of the selected TFs and genes will also be evaluated by comparing the inferred GRN from our method, the approach in GAGER, and the target GRN, focusing on the differences between them.

4. To assess scalability and computational efficiency, we will compare the runtime of our method with that of the method in GAGER, considering the time required to estimate gene expression and regulatory coefficients and infer the GRN.

References

[1] Chowdhury, M.Z.U.S., Any, S.S., Samee, M.A.H. and Rahman, A. GAGER: gene regulatory network assisted gene expression restoration. bioRxiv, pp.2024-11.

[2] Mangan, S., Alon, U. Structure and function of the feed-forward loop network motif. Proceedings of the National Academy of Sciences 100(21), 11980–11985 (2003)

[3] Davidson, E.H., Erwin, D.H. Gene regulatory networks and the evolution of animal body plans. Science 311(5762), 796–800 (2006)

[4] Gaiteri, C., Ding, Y., French, B., Tseng, G.C., Sibille, E. Beyond modules and hubs: the potential of gene coexpression networks for investigating molecular mechanisms of complex brain disorders. Genes, brain and behavior 13(1), 13–24 (2014)

[5] Zickenrott, S., Angarica, V., Upadhyaya, B., Del Sol, A. Prediction of disease–gene–drug relationships following a differential network analysis. Cell death & disease 7(1), 2040–2040 (2016)

[6] Aibar, S., Gonza´lez-Blas, C.B., Moerman, T., Huynh-Thu, V.A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., et al. Scenic: single-cell regulatory network inference and clustering. Nature methods 14(11), 1083–1086 (2017)

[7] Love, M.I., Huber, W., Anders, S. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. bioRxiv (2014) https://doi.org/10.1101/002832, https://www.biorxiv.org/content/early/2014/11/17/002832.full.pdf

[8] Kamimoto, K., Hoffmann, C.M., Morris, S.A. Celloracle: Dissecting cell identity via network inference and in silico gene perturbation. BioRxiv, 2020–02 (2020)