A MULTI-DATA APPROACH TO GENE REGULATORY NETWORK RECONSTRUCTION

Sean Leonard

Dr. Nikos Chrischoides

Old Dominion University

Biological threats have become relevant over the last year since the onset of COVID19. These threats are addressed by finding the necessary mechanisms of the human body to affect them through external methods: vaccines and drugs, among others. Through the development of biology-focused analytical tools, ParaView, a data visualization tool, can also be extended to face biological threats. Not only can the research times of such treatments be decreased, but also these tools can be provided to a broad range of researchers. By developing the ability to conduct gene regulatory network (GRN) reconstruction in ParaView, researchers are better able to produce new medications faster. GRNs can currently be derived from the information found in cells through machine learning and probabilistic graphical models. The current methods are imperfect and do not factor in supplementary genetic data. Given the supplemental data, a data mining approach can be developed. Unlike previous methods, the proposed work relies upon two independent data types to make conclusions supported by more than one form of data. This paper shows that a dual data approach is possible in GRN reconstruction. This also provides an analysis of the differences between space and Earth gene expression.

Introduction

The COVID19 pandemic has brought about a perspective shift for many people regardless of its impact. COVID19 has shown that our existing methods to study biological threats do not entirely protect us and our ways of life. Given the improvements to vaccine development, effective countermeasures were produced at an unprecedented turnaround time; however, the global economy was still significantly affected. Given the unfortunate reality, developing the ability to research the parts of an affected cell quickly is a challenge that must be addressed to speed up the pipeline for facing current and future biological threats.

COVID19, like many viruses, relies on proteins to support its ability to infect and further replicate within the human body. Proteins are small cogs in the biological machine. Proteins can bind to other proteins, which is how changes are triggered in the body. For instance, a virus has proteins that connect to the cells they will infect, allowing them to enter. The instruction manual to build a protein is called a gene. Genes are stored in an extensive library of other protein-coding instructions called a genome. This genome is made up of a molecule called DNA. A cell leverages its DNA by making copies of genes in the form of a molecule called RNA. These RNA molecules are then used to make proteins. However, an organism's genes are not all available to be read at any given time. That is the primary reason that muscle and liver cells act, behave and look like they do. Capturing the RNA molecules in a cell enables a researcher to ultimately capture the RNA molecules or reads that pertain to the active genes in a cell or sample. The active genes in a cell make up a gene expression profile. Numerous regulatory factors control if a gene is active. By finding the network of

genes that interact with each other, we can discover what a cell is working on at any given time.

Furthermore, we can better predict what happens if we expose that cell to a stimulus. However, this prediction is complicated by the numerous regulatory factors that control a gene's expression. This makes capturing the regulatory link between genes a nontrivial task. Two such factors are promoters and enhancers, regions of DNA that pertain to a target gene and can be manipulated to either encourage or inhibit the expression of a target gene. One such manipulator is broadly called DNA methylation. DNA methylation data captures the presence of methyl groups bound to specific areas in the DNA. The presence of methyl groups often has a negative relationship with gene expression[1-3]. As such, if there is a large amount of DNA methylation in a promoter or enhancer region, a gene may not be expressed even if it would be otherwise. Also within the promoter or enhancer regions are motifs. Motifs are binding points for proteins that help to express genes. By capturing the methylation levels in an enhancer, we can assume that the genes controlled by that enhancer will be affected. Currently, capturing all of the states of all the regulatory factors for a particular gene is impossible. Due to limited technology, high costs, and other factors, data for each regulatory factor cannot be generated for many experiments. This makes deriving a GRN much more complex as the regulatory environment has to be treated like a black box. The only information to work on is often gene expression data.

Currently, existing algorithms take in gene expression data and build a gene regulatory network or GRN[4-8]. At present, gene expression data is captured by breaking open a cell and then taking out all RNA molecules and counting them up. However, this data collection method does not capture multiple moments in time for a cell, as once a cell is broken, it does not continue to produce RNA. This stagnates what a cell was doing, making it unclear how its gene expression profile may have changed. As a result, scientists gather numerous cells under the same or similar conditions and capture their gene expression data. This method attempts to catch more than one cell to develop a more comprehensive picture of how that cell's gene expression profile is behaving. This introduces a limitation to current methods. Unfortunately, because they use gene expression data, false positives are generated. In Figure 1 is a simple example of a problem that is caused by this limitation.

Due to this flaw and many others, the removal of such errors is a crucial focus for GRN reconstruction algorithms. One approach is to develop more sophisticated models that are better able to decipher the links between genes. However, this approach will always be limited by the current data collection's flaw. Another demands more data be created over a period of time; however, this is a costly practice. Thus, we believe that by appropriately integrating the data of regulatory factors, errors are reduced. This works by factoring in several aligned pieces of biological evidence that all actively participate in a gene's regulation. The regulatory factor data in question is DNA methylation data which has been shown to affect the regulation of genes[1-3].

In extension, the development of universal or easy-to-access tools for analyzing and studying current and future problems is highly demanded. Given the nature of today's mass focus to solve a biological threat, the need to develop easily accessible high-quality

analysis tools is higher than ever. Currently, ParaView exists as one such tool that enables many people to visualize their data. However, ParaView is limited due to its lack of niche biological analysis and visualization plugins. By developing ParaView in this area, it will be able to support biologists more when facing current and future threats. Currently, GRN reconstruction represents an algorithmic idea of reverse-engineering the cell's inner machinery. A GRN can tell a researcher the genes which have a role in the regulation of other genes. In conjunction with new improvements in parallel mesh generation algorithms on proteins, the need to flesh out the underlying GRN controlling a target protein's creation is one more piece of the puzzle.

Methodology

Within this work, two plugins for ParaView will be laid out. One is referred to as In-depth Protein Tooltip, and another is referred to as GRN analysis.

In-depth Protein Tooltip

Within ParaView, there exist many tools for viewing various types of scalar and vector data. However, these tools often lack some of the finer details needed for the analysis of protein structure. The development of a plugin that was easy to use and integrated with existing ParaView functionality was considered the primary focus. This tool would then need to be able to study an essential part of protein structure. That being the cavities within proteins. The tool allows a user to hover over areas of selected data points and then generate a tooltip that would calculate the mean, standard deviation, and variance of the selection point's values, along with other metrics. A user can then observe cavities while also quickly and intuitively understanding the distributions of various protein-specific values within these cavities. The development can be broken down into three steps. The first requires the data points within a cavity to be extracted. The second needs the basic statistic values like mean, median, etc., to be calculated. The third and final step calls for the basic statistics to be visualized as a tooltip when a user hovers over a set of selected data points.

GRN Analysis

Currently, ParaView cannot reconstruct a GRN. In addition, the visualization of a GRN is a nontrivial task. As such, the development of a GRN analysis process and a more GRN focused visualization plugin is necessary. By natively incorporating a GRN analysis tool within the ParaView framework, a user needs only gather the data necessary to run the GRN reconstruction algorithm. Initially, a user would be forced to run both GRN reconstruction and then modify the output of that GRN reconstruction to be fed and visualized in another biology-related visualization software like Cytoscape.

The current approaches to GRN reconstruction utilize gene expression data. However, that data is one small piece of the puzzle. Acting upon gene expression is numerous known and unknown biological features, all of which alter it. Developing an approach that can integrate gene expression and relevant biological factors, it would then follow that forging of a regulatory link between two genes is more reliable given the multiple different sources of data arriving at a similar conclusion. Therefore, we present a data mining algorithm that factors in both gene expression data and DNA methylation data to reconstruct a GRN.

This method ultimately looks to link both the enhancers and promoters to a target gene. The motifs within the enhancer and promoter regions are then used to find regulatory links between the target gene and the motifs' genes. This method is conducted at the transcript level due to more than one promoter being present for a gene. However, at the transcript level, each transcript of a gene can pertain to a unique promoter. A gene has multiple transcripts or variations either because they have unique start sites or their instructions are truncated somehow. Thus, a gene has numerous transcripts, which all result in unique but largely similar proteins. The gene regulatory network reconstructed with this method is at the gene level so that genes will have links between genes based on its transcripts' links. Therefore, a gene may have several transcripts, each transcript linking to another gene, and the result would be that gene linking with all of the other genes.

The link between the promoter and a particular transcript is done by looking at the transcription start site (TSS). The TSS is where a gene starts being transcribed from DNA. In the case that promoters are overlapping, the center for each promoter is calculated, and the promoter center closest to the TSS is considered that transcript's promoter. At this point, the motifs and their corresponding gene transcripts are collected. The expression of the motif's transcript is then correlated through Pearson correlation with the target transcript. This correlation is between the motif transcript's gene expression for each sample and the target transcript's gene expression for each sample. If the correlation coefficient is above a certain threshold, negative or positive, that motif transcript is considered a regulator of the target transcript linking the two genes for which those transcripts pertain.

A transcript's promoter is calculated and used as a midpoint for a 200kb(200,000 nucleic acids) regulatory region to link enhancers to their target gene. All of the enhancers within this regulatory region are considered potentially connected to the target transcript and collected. The link between each enhancer and transcript is then tested through Pearson correlation, with the target transcript's expression being correlated with the enhancer's methylation level across all samples. As with the promoter, if the correlation coefficient is above a set threshold, it is considered an active enhancer in regulating that target transcript. Given that an enhancer is deemed to be active, the motifs in the enhancer are linked in the same way that the promoter motifs were linked to the target gene.

Results

In-depth Protein Tooltip

After the start of the implementation of the Indepth Protein Tooltip, it was found out that a simpler alternative was possible. This alternative required the utilization of several filters within ParaView. The alternative method called for the use of the threshold filter, which would then extract all the points within a cavity as they all have an identification value that sets which cavity they are a part of. The following steps are trivial parts of ParaView, with the calculation of basic statistics already being part of ParaView. As for the visualization of a tooltip. After the trick with the threshold filter was found, the need for a tooltip was minimal due to the ease at which a user could access the cavity distribution information once they knew how. The setup for this approach is depicted in Figure 2.

GRN Analysis

The validation and differential analysis of the data used in testing the GRN reconstruction algorithm is shown in Figure 3. After validating the data, a necessary step to remove inactive transcripts and normalize their expression values, the data was then run in the GRN reconstruction algorithm depicted in Figure 3. Icam1 and Trpc1, two genes, were selected, and their direct regulating genes are pictured in Figures 4 and 5. The edge color depicts a positive, green, or negative, red, relation with the target gene. The edge width within these figures is calculated from the correlation between the regulating gene and the target gene's gene expression. That value is raised to the power of four and multiplied by five to increase the differences between different correlations. These two genes were selected due to studies that indicated them as important genes in the microgravity and muscle environment[9-11]. The Icam1 gene encodes an intercellular adhesion protein. It has been shown that altering the shape of a cell can alter its function[12]. This mechanotransduction pathway is of considerable interest when observing cells in microgravity conditions. Currently, cells are under the pressure of gravity which works against them to form a shape that has evolved over the time humans have walked the planet. However, when exposed to microgravity conditions, the shape of cells may be altered. If the shape is altered, it stands to reason that the cell's typical pathways are affected, and a protein coded by Trpc1 and Icam1 would be directly in the middle of that change in the process. The differential analysis made it clear that control (on Earth) and case (in space) data showed a marked difference in their gene expression activity, as shown in Figure 6. The further study into these two methods and the extension and further study of the proposed GRN reconstruction method are points of future work. The ParaView plugin portion of this work remains as future work.

Conclusion

The development of tools to further expand and enable scientists to address new biological threats is a way to improve our response times in the future. Through further study of new environments, the mechanisms that control the human body are slowly revealed. Further investigation into the expression of genes in a microgravity environment visualizes the body's reaction to an extreme setting and is a point for future work and analysis.

Dataset Preparation

The data used in this paper comes from the GLDS-99 dataset within GeneLab (NASA GeneLab Data Systems : /genelab/accession/GLDS-99/). The gene expression or RNA-seq data was preprocessed using TrimGalore for trimming, STAR for mapping, and RSEM for quantification of transcripts. FastQC was used for quality control. For the DNA methylation or WGBS-seq data TrimGalore was used for trimming and Bismark was used for mapping. The methylation values were for CpG sites within the data and each CpG site was required to have three overlapping reads before the methylation value was considered. Putative enhancers/promoters[13] had motif finding run on them by Homer on these regions to gather the potential TFs.

Acknowledgments

I would like to acknowledge the Virginia Space Grant Consortium for funding this work. I would also like to thank Dr. Nikos

Chrisochoides and Dr. Jiangwen Sun. I would finally like to thank Angelos Angelopoulos and Kevin Garner.

References

[1] Aran, D., & Hellman, A. (2013). DNA methylation of transcriptional enhancers and cancer predisposition. *Cell*, *154*(1), 11–13.

[2] Harris, C. J., Scheibe, M., Wongpalee, S. P., Liu, W., Cornett, E. M., Vaughan, R. M., Li, X., Chen, W., Xue, Y., Zhong, Z., Yen, L., Barshop, W. D., Rayatpisheh, S., Gallego-Bartolome, J., 6ayesi, M., Wang, Z., Wohlschlegel, J. A., Du, J., Rothbart, S. B., … Jacobsen, S. E. (2018). A DNA methylation reader complex that enhances gene transcription. *Science*, *362*(6419), 1182–1186.

[3] Dhar, G. A., Saha, S., Mitra, P., & Nag Chaudhuri, R. (2021). DNA methylation and regulation of gene expression: Guardian of Our Health. *The Nucleus*.

[4] Papili Gao, N., Ud-Dean, S. M., Gandrillon, O., & Gunawan, R. (2017). Sincerities: Inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinformatics*, *34*(2), 258–266.

[5] Aibar, S., González-Blas, C. B., Moerman, T., Huynh-Thu, V. A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., van den Oord, J., Atak, Z. K., Wouters, J., & Aerts, S. (2017). Scenic: Single-cell regulatory network inference and clustering. *Nature Methods*, *14*(11),

[6] Matsumoto, H., Kiryu, H., Furusawa, C., Ko, M. S., Ko, S. B., Gouda, N., Hayashi, T., & Nikaido, I. (2017). SCODE: An efficient regulatory network inference algorithm from single-cell RNA-seq during differentiation. *Bioinformatics*, *33*(15), 2314–2321.

[7] Sanchez-Castillo, M., Blanco, D., Tienda-Luna, I. M., Carrion, M. C., & Huang, Y. (2017). A 6ayesian framework for the inference of Gene Regulatory Networks from time and pseudo-time series data. *Bioinformatics*, *34*(6), 964–970.

[8] Babtie, A. C., Chan, T. E., & Stumpf, M. P. H. (2017). Learning regulatory models for cell development from single cell transcriptomic data. *Current Opinion in Systems Biology*, *5*, 72–81.

[9] Numaga-Tomita, T., Oda, S., Nishiyama, K., Tanaka, T., Nishimura, A., & Nishida, M. (2018). TRPC channels in exercise-mimetic therapy. *Pflügers Archiv – European Journal of Physiology*, *471*(3), 507–517.

[10] Bradbury, P., Wu, H., Choi, J. U., Rowan, A. E., Zhang, H., Poole, K., Lauko, J., & Chou, J. (2020). Modeling the impact of microgravity at the cellular level: Implications for human disease. *Frontiers in Cell and Developmental Biology*, *8*.

[11] Bizzarri, M., Monici, M., & Loon, J. J. (2015). How microgravity affects the biology of Living Systems. *BioMed Research International*, *2015*, 1–4.

[12] Ingber, D. E. (2003). Tensegrity I. Cell Structure and hierarchical systems biology. *Journal of Cell Science*, *116*(7), 1157–1173.

[13] Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., Sandstrom, R., Ma, Z., Davis, C., Pope, B. D., Shen, Y., Pervouchine, D. D., Djebali, S., Thurman, R. E., Kaul, R., Rynes, E., Kirilusha, A., Marinov, G. K., Williams, B. A., … Ren, B. (2014). A comparative encyclopedia of DNA elements in the mouse genome. *Nature*, *515*(7527), 355–364.

Figures

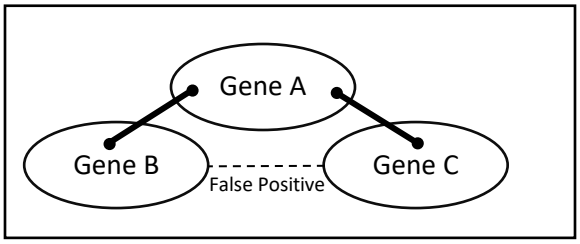
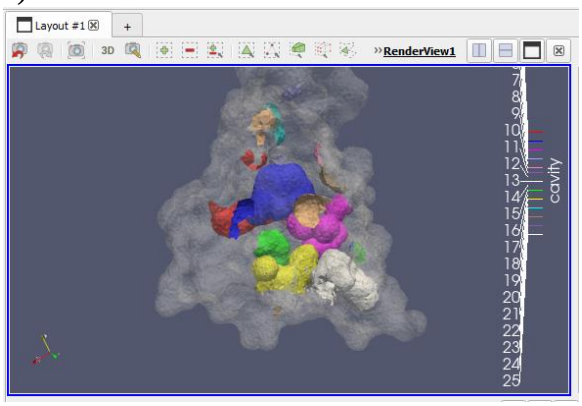
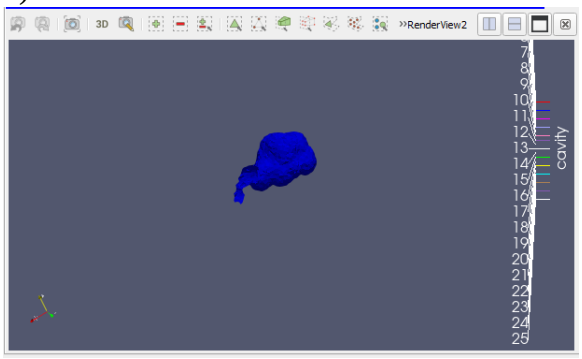


Figure 1. Gene A regulates Gene B and Gene C. As such, Gene B and C could have similar expression levels. In current models, this can result in a false positive link being created between Gene B and C because their gene expression levels are highly correlated. In reality, Gene A should have a link to Gene B and Gene C, and Gene B should not be linked to Gene C.

a)



b)



c)

SpreadSheetView1

Showing DescriptiveStatistic Attribute: Row Data Precision: 6

	Cardinality	M2	M3	M4	Maximum	Mean	Minimum	Variable
0	1482	2638.43	-11911.6	59497.8	5	4.62926	-1.37941	esp_clamped
1	1482	34651.4	-538290	9.35103e+06	19.5136	17.0244	-1.37941	esp_raw

Figure 2. a) a image of the protein with different categories highlighted in different colores b) an extraction of one specific cavity depticted in blue in a) and b). c) The basic statistics for the data with the blue cavity in b) calculated by ParaView.

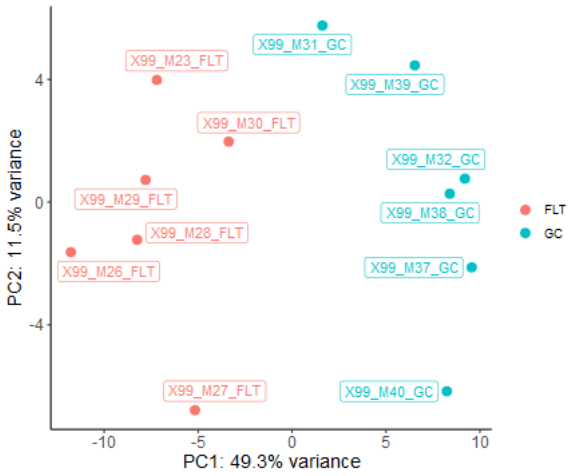


Figure 3. Principal Component Analysis between control and case samples shows a separation between control (GC) and case (FLT) samples.

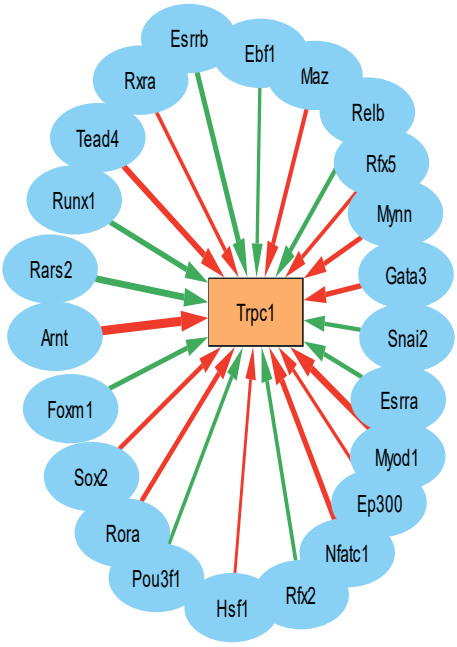


Figure 4. Transcription factors for the Trpc1 gene.

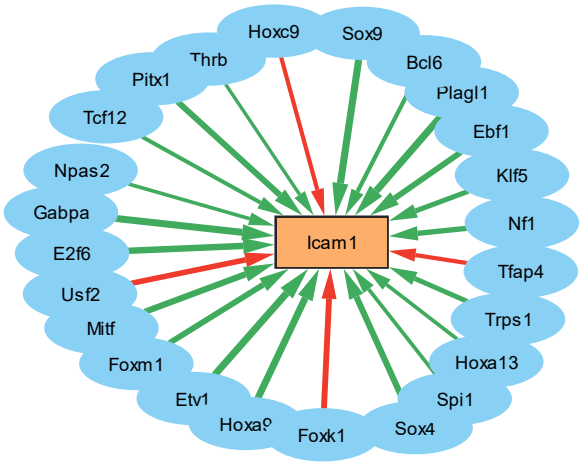


Figure 5. Transcription factors for the Icam1 gene.

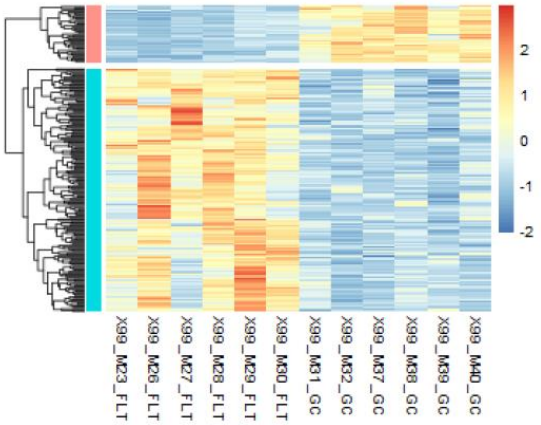


Figure 6. Gene expression activity graph shows a higher activity in case samples (FLT) then control samples (GC).