Undergraduate Research and Discovery Scholar Program
Summer 2020 - Project Proposal
Lebanese American University

Machine Learning Approaches to Gene Regulatory Network Modeling on Glioma Expression Data: A Comparative Study

Mohamed Mehdi MERBAH
Mentor: Georges KHAZEN, PhD

LAU
الجامعة اللبنانيّة الأميركيّة
Lebanese American University

08-2020

Contents

1 Introduction and Motivation

1.1 Problem Description

Transcriptional regulation is a fundamental mechanism that defines the very essence of the development of different life forms, it lies at the core of the central dogma of molecular biology and is an integral part of our current understanding of gene expression regulation at the cellular level. The concept of Gene Regulatory Networks (GRNs) emerged as a systematic approach to model and describe the interactions of different products of transcription and the relationships between the expression rates of different genes; essentially a schema of how a collection of DNA segments within a cell interact with each other through their transcription products. The relatively recent massive gross of high-throughput gene expression data that accompanied the developments of Next-Generation Sequencing (NGS) technologies opened the door for many opportunities when it comes to modeling complex biological systems such as GRNs using computational tools and methods, and hence the application of statistical and machine learning techniques to predict the behavior of these networks.

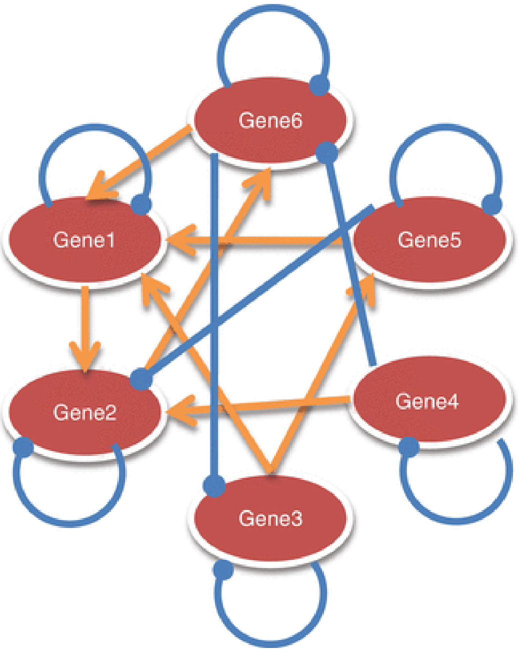

Figure 1: Simplified diagram of a six-node Gene Regulatory Network where each gene is denoted by a node and the edges represent the regulatory relationships

The network model is as illustrated in *Figure 1*, a graphical representation that has regulation target genes (TGs) as nodes and and their regulatory relationships with other genes as their connecting edges. The interaction is often achieved through transcription factors (TFs), proteins encoded for by these genes that serve as either activators or inhibitors to the TGs by binding to them directly or to enhancer or promoter regions in the genome.

The data generated from NGS technology ranging from RNA-sequencing (RNA-seq) measuring RNA transcript levels, to Chromatin immunoprecipitaion experiments coupled with high-throughput DNA-sequencing (ChIP-seq) to identify protein binding and interaction with DNA, are often transformed and then integrated and fed as input into sophisticated computational models that then output interaction scores or gene expression correlation scores, the latter allow us then to construct such graphical interaction models to interpret the network.

Reconstruction of accurate GRN has been and continues to be quite a challenging problem in systems biology, but solving it could potentially be the key to solving much greater problems in biomedicine. Cancer is essentially a disease that originates from a malfunction in the regulatory mechanisms due to mutations in key genes responsible for regulating the cell growth cycle; and so we could see how understanding the GRNs that govern expression levels of oncogenes (genes that have potential to cause cancer) and tumor-supressor genes (anti-oncogenes) is of high importance. Moreover, although Gliomas -a type of brain cancer-represent 33 percent of all brain tumors (Gliomas, JHM), there hasn't been much focus on on the GRN inference for the disease.

1.2 Preliminary Literature Review

The abundance of resources and publicly available data derived from sequencing experiments allowed for many studies on applied computational methods in systems biology to take place. The high-throughput transcriptome datasets generally employed in GRNs reconstruction could be from sequencing experiments where gene expression levels are measured in samples from different individuals or time-course data to compare expression levels over time. These high-throughput measuring techniques paved the way to the use of computational techniques to statistically reconstruct such interactions, a process sometimes called *reverse engineering* (Vân Anh Huynh-Thu and Guido Sanguinetti, 2019).

A variety of reverse engineering strategies were proposed over the years that rely on computational methods, each based on a mathematical model with a specific set of parameters encompassing the different variables involved when studying gene expression. Hache et al. (2009) accomplished a systematic evaluation of the performances of different reverse engineering methods applied on artificial gene expression data: that is, data generated from simulating a known GRN using software; the study compared the performance of Relevance Networks, Bayesian Networks, Dynamic Bayesian Networks, undirected Guassian Graphical models and Neural Networks. Results revealed that Neural Networks performed the best in terms of sensitivity to gene regulatory interactions compared to the other methods in the study. With similar results that highlight the accuracy of machine learning methods in GRNs modeling, Swait et al. (2010) compared three continuous deterministic methods commonly used: The S-system (SS), artificial neural networks (ANNs), and the general rate law of transcription (GRLOT) model;

The study evaluated the prediction accuracy of the methods with reference to dynamic GRN models they created from static models data. Results showed that the ANN method generally produces the most accurate and robust models for the least computational cost. Moreover, in a comprehensive evaluation of machine learning methods in GRN reconstruction, Madhamshettiwar et, al. (2012) conducted a comparative study of eight unsupervised and a single supervised learning method (SIRENE) on ovarian cancer datasets. Major difference in prediction accuracy were observed, yet the supervised method outperformed all of the unsupervised ones.

To further explore the performance of different machine learning approaches in modeling, another comprehensive study (Maetschke et al., 2014) compared supervised, semi-supervised and unsupervised methods on simulated expression data generated using GeneNetWeaver - a tool for generating *in silico* simulated expression data and benchmarks for performance profiling of network inference methods; results indicated that the unsupervised methods achieved low prediction accuracies and were considerably outperformed by the semi-supervised and supervised methods. They also highlight that unsupervised methods are suitable only for simple small networks composed of either activating or inhibitory interactions but not both. The study further performed evaluations with added TF-binding data in an attempt to improve prediction accuracy scores, but to no success as the results remained the same.

The current state of the literature supports the superiority of supervised and semi-supervised machine learning models over unsupervised ones and seems to be a consensus across all the different comparative studies.

2 Proposed Study

Our proposed study aims to run a comparative analysis of different machine learning models applied on gene expression data obtained from Glioma patients to infer GRNs. The choice of disease was mainly directed by the lack of available studies on GRNs modeling in brain tumors and scarcity of experimental exploration of how network modeling on this type of cancer might give insight into the regulatory mechanisms of the disease and also due to the advancements and innovation in specialized machine learning packages since the last widely inclusive comparative study was done by Maetschke et al. (2014).

The mRNA-seq gene expression data for this study were obtained from the Chinese Glioma Genome Atlas (CGGA), a database curated for data storage and analysis to explore brain tumors datasets over 2,000 samples from Chinese cohorts. The expression profiles collected are from 693 individuals with Low-Grade Gliomas (LGG) (443) and Glioblastoma (GBM) (249), a dataset used in a study aimed at localizing seizure-susceptible brain regions associated with LGG by Wang et al. (2015) and another study (Liu et al., 2018) on using radiomics data to develop a prediction model of progression-free survival in patients with LGG as it was used as a training dataset. It is worth mentioning that the datasets also include patient clinical data that encompass other variables such as age, gender and treatment status.

The study will mainly consist of a continuous review of methods and protocols applied on similar types of data to reconstruct interaction and regulation networks and running them on the available data. Hence, the use of a multitude of available "off-the-shelf" prediction and inference tools (WCGNA, GENIE3...etc) used previously on simulated datasets in many studies that recreated the DREAM challenge (a network reconstruction and inference competition) models and used their benchmarks to assess their predictions; aswell as other less packaged and more customizable mehtods fitted for our problem.

Initial data analysis would allow us to set a threshold-score for correlation to infer a direct relationship between individual genes; in essence, a score that is a baseline to decide if two genes do have some sort of regulatory relationship.

Most of the computational tools and packages used in the study would be in Python or R due to the abundance of software packages and libraries that target this specific type of data analysis, and so these would be the main languages used throughout the study. Running the analysis however could require access to remote servers to run the heavy computational jobs that cannot be handled by desktop machines.

3 Risk Assessment

Although the project does not involve any wet-lab work that might entail risks related to costs or data collection as the data is essentially from online resources dedicated to the subject of the study, there are still some technical and logistical risks that should be taken into consideration for planning.

The major risks that may arise during the execution of the project would be mainly time-related. The computational resources offered by the university to handle the computational load required to train the machine learning models on large datasets like those of gene expression have been reported to be down for maintenance since earlier this year; and so there could be a delay in obtaining the results due to that. This delay in data analysis could be compensated by investing the time in mastering more intricate methods and trying to widen the scope of possible applications, though it would shift the project schedule to some extent.

The element of uncertainty in university plans as the coming academic year goes due to the current disruption caused by the global pandemic might also a contributing factor to the delay of progression of the project when it comes to planning. As the execution of fully-online or hybrid learning plan is still under study, some central timepoints like exam dates and schedules are yet to be determined, and therefore limit the potential for more elaborate planning.

These risks could, however, be accounted for by contingency planning in case of a delay in the project schedule and a proactive approach to restructuring and rescheduling of the tasks as we move along.

The time-bound risks may be the major factors, but are manageable. The pressure is alleviated to some extent as this project is not my senior study and therefore offers more of a great learning opportunity of how Bioinformatics research is conducted.

4 Expected Results and Impact

As the preliminary literature review had illustrated, most supervised methods of inference outperformed the unsupervised ones in terms of prediction accuracy and sensitivity to gene interaction in GRN modeling. When it comes to accuracy and robustness results we do expect our results to match those of the aforementioned studies, and hence the superior performance of supervised methods due to prior knowledge in the training sets with associated interaction labels. Although it would be interesting to see the interactions of oncogenes and their regulatory activity in LGG and GBM and the sorts of insights that we might extract from the models; and hence, shed some light on gene regulatory activity in brain cancer and the associated GRNs that govern it.

Cancer being mainly a disease of altered gene expression regulation, seeing how perturbations to a system of gene interactions would affect the rate of expression of previously known cancer genes or ones that regulate them (such as tumour suppressor genes) would be quite insightful in understanding how alterations to transcriptional regulation contributes to tumour development. Assessing the accuracies of machine learning methods for GRN inference then, could contribute to the investigation on how to control the transcriptional activation of gene expression in cancer.

An extension to the study after benchmarking the performance of the studied methods would a more detailed research project to identify how expression patterns of key oncogenes and/or genes responsible for treatment drug resistance within GRNs of different cancer types; this would allow us to see how ML based methods compare to recently developed innovative computational methods like the one constructed by Zhang et al. (2019) in a study aimed at generating insights into dynamic adaptation and regulatory network mechanisms in cancer drug resistance.

Ultimately, the goal of the study is to provide an insight into what type of ML models are best fit for GRN inference (GRNI) and more specifically GRNI of cancer gene expression data, we do not aim to construct our own mathematical or learning model but rather assess the performance of known models with a certain degree of optimization to fit the context of the problem and improve their accuracy.

5 Project Plan and Duration

The project plan could be described as a phased implementation with overlapping and sometimes continuously running tasks that will span over the duration of both the Fall 2020 and Spring 2021 academic semesters. The following is the general work breakdown structure that highlights the main components of the implementation plan:

1. Readings and Literature Review
    (a) Continuous review of literature on GRN reconstruction
    (b) Review documentation of methods to be applied
    (c) Review documentation of packages and libraries to be used
2. Data Analysis
    (a) Data wrangling and preparation
    (b) Learning methods and how to apply them
    (c) Applying methods and assessing prediction quality
    (d) Generating results and visual summaries
3. Reporting
    (a) Compiling results and describing them
    (b) Discussing the results and reflecting on previous work
    (c) Drafting a conclusion
    (d) Producing a research report

6 Conclusion

Gene Regulatory Network inference is quite an interesting and equally challenging task, and applying modern machine learning methods to infer gene regulatory interactions would be a very insightful opportunity to dive into prediction modeling and cancer genomics by studying gene expression data for such a complex disease like Glioma. Having the chance to take carry out a study of such prestige and relevance the recent advancements in biomedical research is certainly a unique chance to enter the academic research world from its widest door and nourish my passion for science.

References

Bradner, J. E., Hnisz, D., amp; Young, R. A. (2017). Transcriptional Addiction in Cancer. Cell, 168(4), 629-643. doi:10.1016/j.cell.2016.12.013

CGGA - Chinese Glioma Genome Atlas. (2016). Retrieved August, 2020, from http://cgga.org.cn/

Fantine Mordelet, Jean-Philippe Vert. (2008) SIRENE: supervised inference of regulatory networks, Bioinformatics, Volume 24, Issue 16, 76–82, https://doi.org/10.1093/bioinformatics/btn273

Gliomas. (n.d.). Retrieved August, 2020, from https://www.hopkinsmedicine.org/health/conditions-and-diseases/gliomas

Huynh-Thu V.A., Sanguinetti G. (2019) Gene Regulatory Network Inference: An Introductory Survey. In: Sanguinetti G., Huynh-Thu V. (eds) Gene Regulatory Networks. Methods in Molecular Biology, vol 1883. Humana Press, New York, NY. https://doi.org/10.1007/978-1-4939-8882-2_1

Hache, H., Lehrach, H. Herwig, R. Reverse Engineering of Gene Regulatory Networks: A Comparative Study. J Bioinform Sys Biology 2009, 617281 (2009). https://doi.org/10.1155/2009/617281

Liu, X., Li, Y., Qian, Z., Sun, Z., Xu, K., Wang, K., Wang, Y. (2018). A radiomic signature as a non-invasive predictor of progression-free survival in patients with lower-grade gliomas. NeuroImage: Clinical, 20, 1070-1077. doi:10.1016/j.nicl.2018.10.014

Madhamshettiwar, P.B., Maetschke, S.R., Davis, M.J. et al. (2012) Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets. Genome Med 4, 41. https://doi.org/10.1186/gm340

Stefan R. Maetschke, Piyush B. Madhamshettiwar, Melissa J. Davis, Mark A. Ragan (2014) Supervised, semi-supervised and unsupervised inference of gene regulatory networks, Briefings in Bioinformatics, Volume 15, Issue 2, 195–211, https://doi.org/10.1093/bib/bbt034

Swain, M.T., Mandel, J.J. Dubitzky, W. (2010) Comparative study of three commonly used continuous deterministic methods for modeling gene regulation networks. BMC Bioinformatics 11, 459. https://doi.org/10.1186/1471-2105-11-459

Wang, Y. (2013). Gene Regulatory Networks. Encyclopedia of Systems Biology, 801-805. doi:10.1007/978-1-4419-9863-$7_3$64

Yinyan Wang, Tianyi Qian, Gan You, Xiaoxia Peng, et al. (2015). Localizing seizure-susceptible brain regions associated with low-grade gliomas using voxel-based lesion-symptom mapping, Neuro-Oncology, Volume 17, Issue 2, 282–288, https://doi.org/10.1093/neuonc/nou

Zhang, J., Zhu, W., Wang, Q., Gu, J., Huang, L. F., amp; Sun, X. (2019). Differential regulatory network-based quantification and prioritization of key genes underlying cancer drug resistance based on time-course RNA-seq data. PLOS Computational Biology, 15(11). doi:10.1371/journal.pcbi.1007435