Causal Discovery of Gene Regulatory Networks on Tissue-Specific Human Cancer Gene Expression data

Yamini Mathur*, Kumara Sri Harsha Vajjhala*, Sai Charishma Valluri*
*Iowa State University
Department of Computer Science
Atanasoff Hall, 2434 Osborn Dr, Ames, IA 50011
{yamini,harshavk,svalluri}@iastate.edu

***Abstract*—In this project, we predict the causal relationships between genes in gene expression data and thereby identify the Gene Regulatory Network(GRN). To identify the Gene Regulatory Network, we utilized two methods: the Path Consistency Algorithm based on Conditional Mutual Information(PCA-CMI) and the Inductive Causation(IC) Algorithm. We implemented the PCA-CMI algorithm by Zhang et al. [1] and extended it to include edge orientations. We also carried out experiments on the IC Algorithm using the Causality Python library [2], and compared the results with the results obtained from the PCA-CMI Algorithm on the DREAM3 In-Silico Network Inference Challenge Dataset [3]. Our results indicate that the PCA-CMI Algorithm significantly outperforms the results of the IC Algorithm. We further tested our PC Algorithm on the GTEX-TCGA Human Cancer Dataset [4] and obtained compelling results.**

I. Introduction

Gene expression controls important information about cellular functions and the interaction between genes. The process of interaction between the genes that regulate is called gene regulation, in which cells increase or decrease the production of specific genes [5]. It is also known that modeling biological networks describe relationships primarily between Deoxyribonucleic Acid(DNA), RNA, metabolites, and proteins. Gene Regulatory Networks can be estimated from gene expression data using causal inference [6]. This can be of great benefit to solve various biological problems by discovering Gene Regulatory Networks, which decide how genes are activated or repressed, and have information about which cells are creating proteins at a particular time [7]. Recent advances in high-throughput biological data collection have provided novel platforms for understanding Gene Regulatory Networks(GRNs), thus creating an enormous interest in mathematically modeling biological networks. Extraction of cause and effect relations in GRNs will open up research involving probable causes of diseases such as cancer and AIDS.

Causal inference has been applied to various biological system problems such as predicting causal networks of gene pathways to identify disease pathways and disease-causing genes [8]. Causal Inference in GRNs has been applied and tested for various types of data such as Breast cancer datasets, plants, insights into the transcription of cancerous cells [9], and others. Further investigation on related works has also been undertaken, such as comparing the effectiveness of different methods to infer causal GRNs. This investigation concludes that Causal Entropy and Transfer Entropy performed well and fusion data of various types from various sources will likely produce significant results [10]. Furthermore, the "Dialogue for Reverse Engineering Assessments and Methods" (DREAM) project has conducted several competitions for various GRN-related inference problems [11]. The significant challenge in estimating GRNs from expression data is that these network inference problems, for now, are mostly restricted to low-dimensional data due to high computational complexity [12]. So the high dimensionality of data is currently a great challenge in this area [6]. Another challenge includes the possible non-linear dependencies as conditional independence tests are computationally expensive [13].

The rest of the paper is organized as follows. Section II provides a background of datasets used. Section III presents other related research conducted in this area. Section IV presents a detailed description of the two methods; PC CMI algorithm and IC method. Section V describes the experimental setup and the experiments we conducted in this study. Section VI presents the obtained results and a comparison between the methods employed. Section VII provides a conclusion of the study and explores future work.

II. Datasets

We utilize two datasets, one simulated and the other, a real-world dataset. The Dream3 In-Silico Network Inference challenge dataset was a simulated gene expression dataset with interventional (gene-knockdown) data. The In-Silico data was generated by continuous differential equations, which approximated the actual gene regulatory interactions. Further, a reasonable amount of gaussian noise was added to the approximations. The dataset consists of data for two organisms; Yeast and Ecoli. There are 10, 50, and 100 node expression data for each organism. The dataset includes different experiments performed on the simulated gene expression data. The first Trajectories experiment uses time-series-based data with environmental perturbations on observational data. The other two experiments are called heterozygous mutants and homozygous null mutants, which contain gene knock-down

and gene knock-out interventional data. The second dataset used is the GTEX-TCGA Human Cancer Dataset, a well-studied real-world dataset of numerous cancer types (TCGA) and non-cancer (GTEX) gene expression data. It consists of RNA-seq expression profiles of both cancer and non-cancer tissues. This dataset contains 10 tissues, out of which we picked the gene expression data for Breast and Lung for both cancer and non-cancer samples. The dataset consists of 1054 samples for Breast cancer, of which 965 data samples contain a tumor, and the remaining 89 samples are normal. The Lung cancer data has 804 samples, out of which 491 samples have a tumor, while the remaining 313 samples are normal. The benchmark gene regulatory networks were obtained from The grand gene regulatory network database.

III. Related works

Various approaches have been introduced to identify the Gene Regulatory Network (GRN) from the gene expression data. Some of them used MI-based approaches, while others used well-known Euclidean distance and Pearson correlation coefficient [14].

Several mutual information-based approaches have been successfully applied to infer GRNs such as ARCANE [15], and CLR [14]. These approaches mainly focus on computing the pairwise Mutual Information values between all the possible pairs of genes, resulting in an MI matrix. It is well known that mutual information (MI) provides a general measurement for dependencies in the data, in particular positive, negative, and nonlinear correlations. Another advantage of MI-based methods is their ability to deal with thousands of variables (genes) in the presence of a limited number of samples [16]. Despite these advantages, MI-based methods were helpful only when investigating pairwise regulations in a GRN. They are unable to discover the joint regulations of a gene by two or more genes [1]. In contrast, Conditional Mutual Information (CMI) is capable of detecting the joint regulations by exploiting the conditional dependency between genes of interest [1].

Another recent work on identifying the causal relationships in gene regulatory networks was again done by Zhang et al [17]. They extended the PCA-CMI approach that we reproduced in this project, to incorporate a novel conditional independence test they proposed, called the conditional mutual inclusive information (CMI2). Going for a new measure for conditional independence test was because the existing CMI measure underestimates the regulation strength between genes given a conditioning set, resulting in a higher number of false negatives. CMI2 is calculated using the Kullback-Leibler divergence between the possible distributions when an edge is added and not added between two vertices in a graph. Assuming that the underlying data is gaussian, CMI2 can be efficiently calculated. Their causal discovery approach based on CMI2, called CMI2NI, outperformed the traditional PC and PCA-CMI approaches and was tested on both Dream 3 and Human cancer TCGA datasets.

IV. Methodology

In this section, we will introduce some definitions of information theory, including Entropy, Joint Entropy, Mutual Information(MI), Conditional Mutual Information (CMI), and the algorithms PCA-CMI and IC for inferring GRNs.

A. Information theory

We used the formulation of MI and CMI from the information theory in [1]. MI is used as a criterion for measuring the dependence between two variables (genes) X and Y. Similarly, CMI is used for measuring the Conditional Independence between variables (genes) X and Y given another set of variables (genes) Z.

For a discrete variable X, the entropy *H(X)* is the measure of average uncertainty of variable X and can be defined by

$$H(X) = -\sum_{x \in X} p(x) \log p(x), \tag{1}$$

where *p(x)* is the probability of each discrete value *x* in X.

The joint entropy *H(X,Y)* of X and Y can be denoted by

$$H(X,Y) = -\sum_{x \in X, y \in Y} p(x,y) \log p(x,y), \tag{2}$$

where *p(x,y)* is the joint probability of x in X and y in Y.

MI measures the dependency between two variables. For discrete variables X and Y, MI is defined in terms of their entropies as

$$I(X,Y) = H(X) + H(Y) - H(X,Y), \tag{3}$$

where H(X,Y) is joint entropy of X and Y. High MI value indicates that there may be a close relationship between the variables (genes), while low MI value implies their independence.

CMI measures conditional dependency between two variables (genes) given other variable(s) [gene(s)]. The CMI of variables X and Y given Z is defined in terms of entropies as

$$H(X,Y|Z) = H(X,Z) + H(Y,Z) - H(Z) - H(X,Y,Z), \tag{4}$$

where H(X,Z), H(Y,Z), H(X,Y,Z) are joint entropies. Similarly, high CMI indicates that there may be a close relationship between the variables X and Y given variable(s) Z.

B. PCA-CMI Algorithm

After we obtain MI and CMI values through 3 and 4, the path consistency algorithm (PCA) is used to remove the edges with (conditional) independent correlation from the graph. The inference of GRNs will be performed by deleting the edges with independent correlation recursively, i.e. from low to high order independent correlation until there is no edge that can be deleted. We implemented the PCA-CMI algorithm as described in [1]. The following gives a brief description of the algorithm.

Step-0:(Initialization)

Input the gene expression data, and set the parameter θ for deciding the independence. Generate the complete network G

for all the genes (i.e., clique graph of all genes). Set $L = -1$

Step-1:

$L = L+1$. For a non-zero edge $G(i,j) \neq 0$, select adjacent genes connected with both genes i and j. Compute the count T of the adjacent genes (not including genes i and j).

Step-2:

If $T < L$, stop. If $T \geq L$, select out L genes from these T genes and let them as $K = [k_1, ..., k_L]$. The number of all selections for K is C_T^L. Compute the L-order $CMI(i,j|K)$ for all C_T^L selections (note:calculate $MI(i,j)$ if K is empty), and choose the maximal one denoting $I_{max}(i,j|K)$. If $I_{max}(i,j|K) < \theta$, set $G(i,j) = 0$. Return to Step-1.

The process of PCA-CMI in detail is described as follows. First, generate a complete graph according to the number of genes. Second, for adjacent gene pairs i and j, compute MI (zero-order CMI) I(i,j). If the gene pair i and j have an MI value less than the threshold, it represents independent correlation; then we delete the edge between genes i and j. Third, for adjacent gene pairs i and j, select the adjacent gene k of them and compute first-order $CMI(i,j|K)$. If the gene pair i and j have a CMI value less than the threshold, which represents their independent correlation, delete the edge between them. The next step is to compute higher-order CMI until there are no more adjacent edges. The resultant true network would be the GRN.

C. IC Algorithm

Inductive Causation (IC) Algorithm provides a way to identify which equivalence class a joint distribution is compatible with, given the conditional independence it contains. The input to the algorithm is the joint distribution p($\{$V$\}$) on the set $\{$V$\}$ of variables, and the output is a graphical pattern that reflects all and no more conditional independencies than the ones in p($\{$V$\}$) [18]. The algorithm is as follows.

1) For each pair of variables a and b in $\{$V$\}$ search for a set S_{ab} such that conditional independence between a and b given $S_{ab}(a \perp b|S_{ab})$ holds in p($\{$V$\}$). Construct an undirected graph linking the nodes a and b if and only if S_{ab} is not found.
2) For each pair of non-adjacent nodes a and b with a common adjacent node c check if c belongs to S_{ab}. If it does, then continue. If it does not, then add arrowheads pointing at c to the edges (i.e., $a \rightarrow c \leftarrow b$).
3) In the partially oriented graph that results, orient as many edges as possible subject to two conditions: (i) Any alternative orientation would yield a new v-structure. (ii) Any alternative orientation would yield a directed cycle.

We used the existing library in [2] that already implemented the IC algorithm for the algorithm implementation purpose.

D. Edge Orientation

We implemented the below four rules for the Edge Orientation as described in [18].

1) Orient $b - c$ into $b \rightarrow c$ if there is $a \rightarrow b$ such that a and c are not adjacent.
2) Orient $a - b$ into $a \rightarrow b$ whenever there is a chain $a \rightarrow c \rightarrow b$
3) Orient $a - b$ into $a \rightarrow b$ whenever there are two chains $a - c \rightarrow b$ and $a - d \rightarrow b$ such that c and d are nonadjacent.
4) Orient $a - b$ into $a \rightarrow b$ whenever there are two chains $a - c \rightarrow d$ and $c \rightarrow d \rightarrow b$ such that c and d are nonadjacent and a and d are adjacent.

We oriented the edges of the undirected graphs of the PCA-CMI Algorithm and IC Algorithm using these edge orientation rules.

V. Experiments

Our experiment setup comprised a Google colab notebook with a Linux operating system on Intel(R) Xeon(R) CPU at 2.20GHz, with a RAM size of 12 GB. We divided the task of experimentation into three parts:

1) Comparing the PCA-CMI and IC algorithm's performance on the simulated gene expression data of *Yeast* and *E.coli* organisms provided by Dream 3 dataset
2) Comparing the edge-oriented causal equivalence graphs of both PCA-CMI and IC with the gold standard DAGs for *Yeast* and *E.coli* in Dream 3 dataset and
3) Contrasting the gene regulatory network obtained for cancer samples with non-cancer samples for Breast and Lung tissues in the GTeX-TCGA Human cancer dataset.

For experiments (1) and (2), we ran our PCA-CMI implementation on the 10 and 50 node datasets of *Yeast* and *E.coli* with 8 different θ values - (0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09), and compared the maximum value among these with the metric of IC algorithm implementation, which used a Robust Regression for conditional independence testing. In experiment 3, we first pre-processed the custom dataset of GTeX-TCGA provided by MetaOmGraph [4]. In the pre-processing step, we first log-transformed the data to reduce the variability in the distribution. Then, we removed mitochondrial genes, which are not involved in the regulation process. On this pre-processed data, we used a random forest classifier to classify cancer and non-cancer samples for the breast and lung tissues and obtained the list of the most important genes involved in the classification process. Using the top ten important genes from this list, we reduced the data dimensions to these ten genes. The PCA-CMI algorithm was then applied separately to the tumor and non tumor samples.

In order to compare the graphs obtained by the algorithms with the gold standard gene regulatory networks, we first converted the predicted graphs (skeleton or an equivalence class) to adjacency matrices. An adjacency matrix A is a 2-dimensional square matrix, where each row and column represent a vertex in the graph. In an undirected graph, an edge between vertex V1 and V2 is represented by a value 1 in both the cells *A[V1][V2]* and *A[V2][V1]*. Whereas, in a directed graph, a directed edge between vertex *V1* and *V2* is represented by a value 1 in cell *A[V1][V2]* and a value 0 in

cell *A[V2][V1]*.

Similarly, we converted the gold standard graphs for each experiment into adjacency matrices. We then utilized an approach similar to the evaluation metrics for the binary classification problem in machine learning to compare the two corresponding matrices. We then converted the 2-dimensional adjacency matrices into 1-dimensional arrays and compared each corresponding array entry. Comparing the number of zeros and ones we predicted correctly in the adjacency matrices gives information on how many edges we were able to predict correctly, since every edge orientation is represented by a one (in both directions for an un-directed graph and one direction for a DAG) and every non-edge between vertices is represented by a zero on both cells of the vertices in the matrix.
The evaluation metrics we used were the F1 score, which is the harmonic mean between precision (positive prediction rate) and recall (sensitivity), true positive rate, and false-positive rate. The overall accuracy of the prediction was measured too, but it is not a good comparison metric, as both the predicted and benchmark matrices are skewed towards the class zero, which make up around 90% of the matrix, thereby resulting in an overall accuracy close to that value.

VI. Results

A. PCA-CMI and IC algorithm skeleton identification

The 10-node Dream 3 dataset results for both PCA-CMI and IC algorithm indicate that PCA-CMI performs slightly better than IC across all the different perturbation types and organisms. The table below shows the maximum F1 score of PCA-CMI compared to the F1 score of the IC algorithm on the 10-node Dream 3 dataset for Yeast and E.coli.

TABLE I: Comparison of the PCA-CMI and IC algorithms for skeleton discovery on Dream 3 dataset for gene regulatory netoworks of size 10

	PCA-CMI	IC
Experiment	F1 Score	F1 Score
Ecoli1-trajectories	0.59	0.55
Ecoli1-heterozygous	0.56	0.49
Ecoli1-null-mutants	0.60	0.47
Ecoli2-trajectories	0.52	0.52
Ecoli2-null-mutants	0.60	0.49
Ecoli2-heterozygous	0.56	0.44
Yeast1-null-mutants	0.66	0.48
Yeast1-heterozygous	0.62	0.42
Yeast1-trajectories	0.46	0.49
Yeast2-trajectories	0.70	0.58
Yeast2-null-mutants	0.57	0.40
Yeast2-heterozygous	0.66	0.39
Yeast3-heterozygous	0.63	0.49
Yeast3-null-mutants	0.45	0.36
Yeast3-trajectories	0.75	0.54

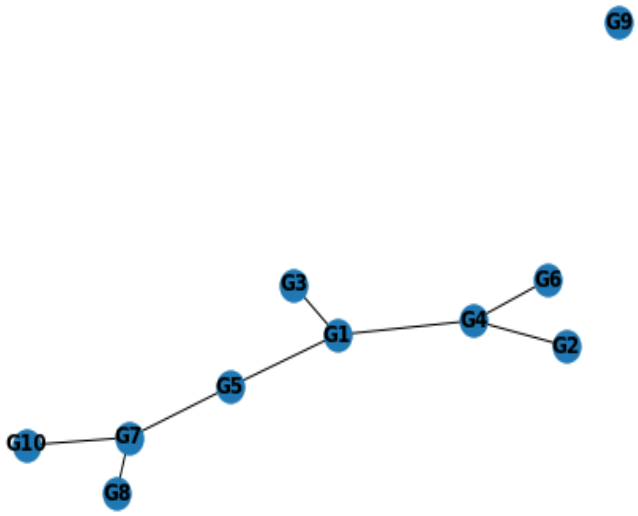
For the Dream 3 dataset of size 50, we observed F1 scores and True positive rates around 0.5, indicating that around 50% of the edges in the gold standard were being identified accurately by PCA-CMI. However, we could not run the IC algorithm on this dataset due to the library not handling the data size.

TABLE II: F1 scores and True positive rate(TPR) of PCA-CMI algorithm on Dream 3 dataset for gene regulatory netoworks of size 50

	PCA-CMI	PCA-CMI
Experiment	F1 Score	TPR
Ecoli1-null-mutants	0.49	0.49
Ecoli1-nonoise-heterozygous	0.52	0.51
Ecoli1-nonoise-proteins	0.52	0.51
Ecoli1-trajectories	0.50	0.53
Ecoli1-heterozygous	0.49	0.49
Ecoli2-nonoise-proteins	0.49	0.50
Ecoli2-nonoise-heterozygous	0.49	0.50
Ecoli2-heterozygous	0.50	0.50
Ecoli2-null-mutants	0.52	0.53
Ecoli2-trajectories	0.49	0.52
Yeast1-null-mutants	0.53	0.54
Yeast1-nonoise-heterozygous	0.51	0.51
Yeast1-heterozygous	0.47	0.46
Yeast1-nonoise-proteins	0.51	0.51
Yeast1-trajectories	0.50	0.52
Yeast2-nonoise-proteins	0.52	0.52
Yeast2-null-mutants	0.52	0.52
Yeast2-nonoise-heterozygous	0.52	0.52
Yeast2-trajectories	0.53	0.54
Yeast2-heterozygous	0.51	0.51
Yeast3-null-mutants	0.53	0.53
Yeast3-heterozygous	0.50	0.50

Below is an example of the GRN skeleton which our PCA-CMI algorithm predicted.

Fig. 1: Yeast 1 GRN skeleton predicted by PCA-CMI



B. PCA-CMI edge orientation

We tested the edge orientation algorithm on the Dream 3 dataset of size ten networks. The results indicated a slight decrease in the true positive rate of the number of edges identified compared to the skeleton identification results. These were

Fig. 2: Yeast 1 gold standard skeleton

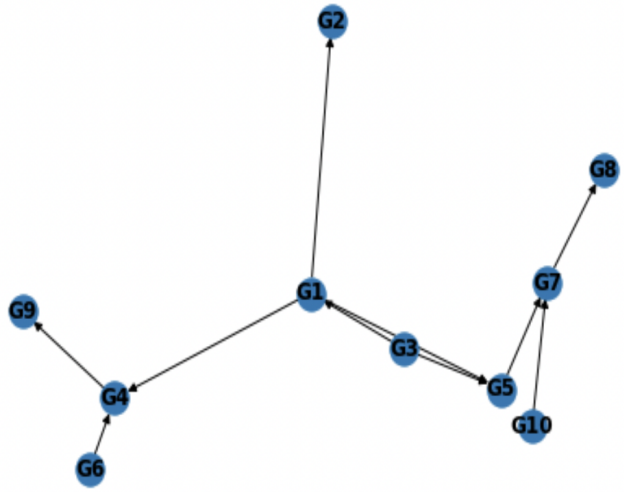
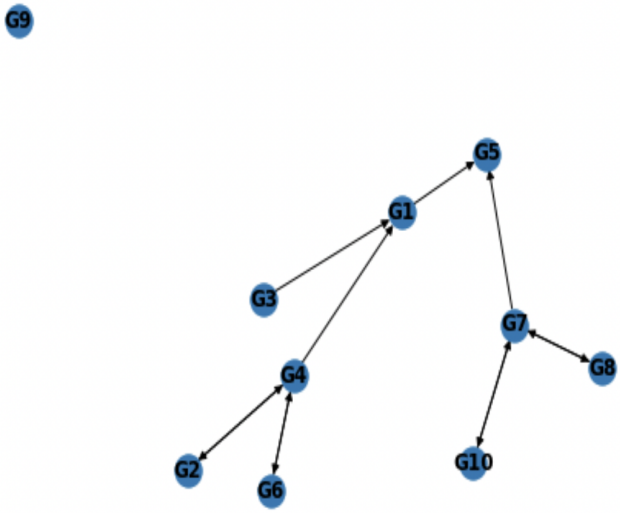


Fig. 3: Yeast 1 GRN equivalence class predicted by PCA-CMI



expected because 1) an unoriented edge in our equivalence class is represented by a bi-directional edge indicating a one on both the cells of the adjacency matrix, while it would be having a value of zero on one of the ends in the benchmark DAG. 2) Our edge orientation algorithm could have some false positives.

TABLE III: Comparison of the PCA-CMI (skeleton) and PCA-CMI (edge-oriented) F1 scores compared to the benchmark on Dream 3 dataset for gene regulatory netoworks of size 10

	PCA-CMI (skeleton)	PCA-CMI (edge-oriented)
Experiment	F1 Score	F1 Score
Ecoli1-trajectories	0.59	0.58
Ecoli1-heterozygous	0.56	0.47
Ecoli1-null-mutants	0.60	0.59
Ecoli2-trajectories	0.52	0.49
Ecoli2-null-mutants	0.60	0.58
Ecoli2-heterozygous	0.56	0.51
Yeast1-null-mutants	0.66	0.66
Yeast1-heterozygous	0.62	0.61
Yeast1-trajectories	0.46	0.44
Yeast2-trajectories	0.70	0.70
Yeast2-null-mutants	0.57	0.57
Yeast2-heterozygous	0.66	0.66
Yeast3-heterozygous	0.63	0.63
Yeast3-null-mutants	0.45	0.36
Yeast3-trajectories	0.75	0.75

Figure 3 is an example of the edge oriented GRN which our PCA-CMI algorithm predicted.

C. Contrasting the GRN obtained for tumor vs non-tumor samples for breast and lung tissues in the GTeX-TCGA human cancer dataset

We observed some interesting patterns in the causal graphs obtained for tumor samples when compared to the non-tumor samples. For the breast tissue in figure 4, we could see that the gene SCARA5 does not involve in any regulatory activity in a normal tissue, while it has a lot of regulatory connections in the non-tumor tissues. This could lead to some interest in studying the gene further to ascertain that it is involved in breast cancer regulatory activity.

However, for the lung tissue in figure 5, we did not observe much difference between the causal graphs for LUAD, Lung_normal and LUSC classes. It could mean that these genes might not be having a causal relationship with the cancer outcome.

Fig. 4: The causal graphs for breast cancer(left) and normal breast tissue(right) obtained by running the PCA-CMI algorithm on GTeX-TCGA dataset

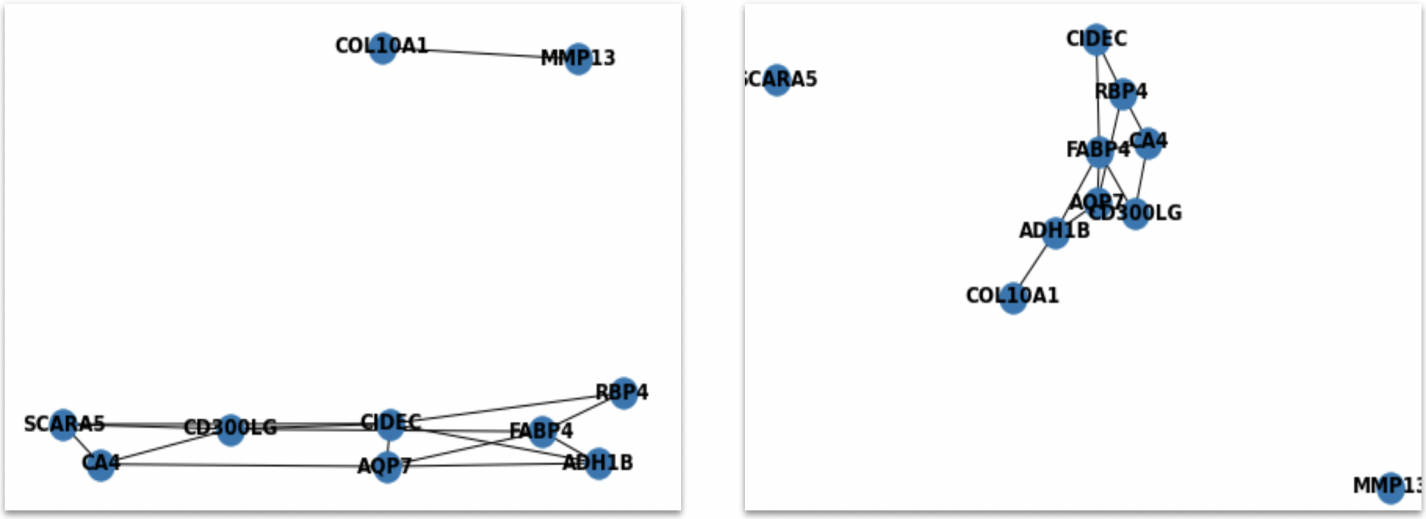
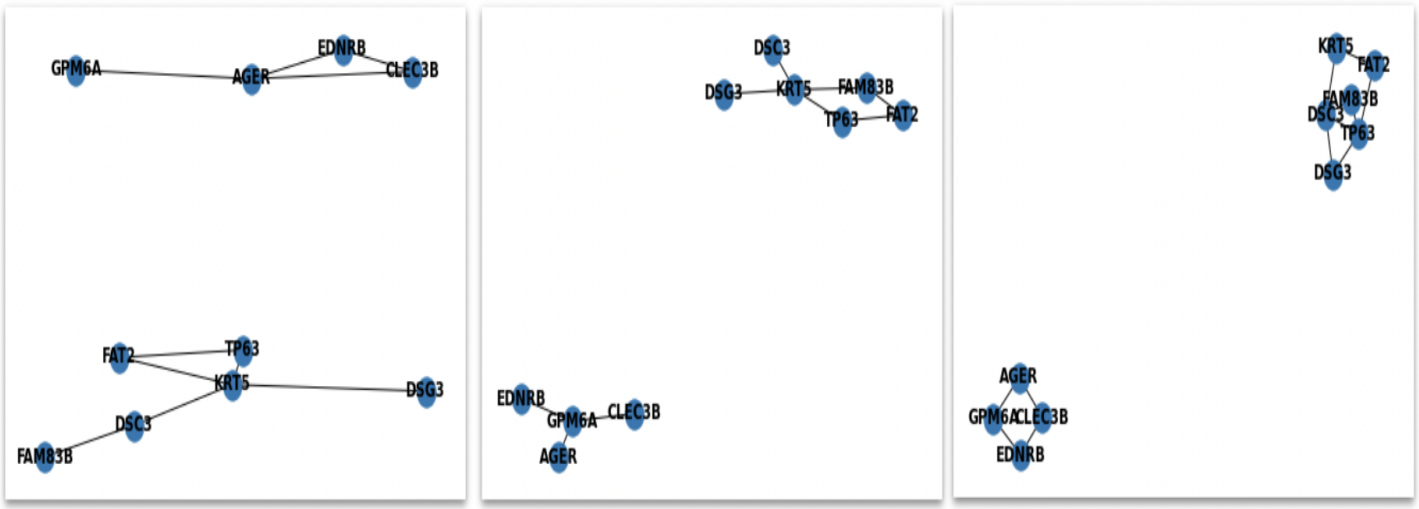


Fig. 5: The causal graphs for Lung adenocarcinoma (LUAD - left) and normal lung tissue(center) and Lung squamous cell carcinoma(LUSC - right) obtained by running the PCA-CMI algorithm on GTeX-TCGA dataset



VII. Conclusion

In this paper, we have implemented the PCA-CMI algorithm proposed by Zhang et al. and extended it to include edge orientation. We compared the performance of our PCA-CMI algorithm with a standard implementation of the Inductive Causation (IC) algorithm on the Dream 3 simulated dataset, which consists of various gene knockout and gene knock-down perturbations. The results we obtained indicated that the PCA-CMI algorithm outperforms the IC on most of the experiments and has a shorter execution time. We then applied the PCA-CMI approach to the real-world GTeX-TCGA human cancer dataset for breast and lung tissues and found interesting regulatory interactions between tumor and non-tumor tissues. In future work, we could investigate a way to utilize a deep-learning-based conditional independence test, as CMI is not accurate for larger conditional set Z. Moreover, we could perform more experiments on the real-world human cancer dataset and compare it with the known gene regulatory networks. It would provide a deeper insight into how the causal discovery performs on noisy, non-gaussian data.

Our source code and results can be found here: https://github.com/vksriharsha/Causal-Inference-on-Gene-Expression-Data

REFERENCES

[1] X. Zhang, X.-M. Zhao, K. He, L. Lu, Y. Cao, J. Liu, J.-K. Hao, Z.-P. Liu, and L. Chen, "Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information," *Bioinformatics*, vol. 28, no. 1, pp. 98–104, 2012.

[2] Akelleh. Causality. [Online]. Available: https://github.com/akelleh/causality

[3] (2021, Mar) About dream. [Online]. Available: https://dreamchallenges.org/about-dream/

[4] U. Singh, M. Hur, K. Dorman, and E. S. Wurtele, "MetaOmGraph: a workbench for interactive exploratory data analysis of large expression datasets," *Nucleic Acids Research*, vol. 48, no. 4, pp. e23–e23, 01 2020. [Online]. Available: https://doi.org/10.1093/nar/gkz1209

[5] T. A. Long, S. M. Brady, and P. N. Benfey, "Systems approaches to identifying gene regulatory networks in plants," *Annual review of cell and developmental biology*, 2008. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2739012/

[6] P. Bode, "Trends in bioinformatics: Causal inference on gene expression data," *Hasser Plattner Instituit*. [Online]. Available: https://hpi.de/fileadmin/user_upload/fachgebiete/plattner/teaching/TrendsBioinformatics/TiB2018/final_PhilippBode.pdf

[7] S. S. Ahmed, S. Roy, and J. Kalita, "Assessing the effectiveness of causality inference methods for gene regulatory networks," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 17, no. 1, pp. 56–70, 2020.

[8] C.-H. Peng, Y.-Z. Jiang, A.-S. Tai, C.-B. Liu, S.-C. Peng, C.-T. Liao, T.-C. Yen, and W.-P. Hsieh, "Causal inference of gene regulation with subnetwork assembly from genetical genomics data," *Nucleic acids research*, Mar 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3950678/

[9] S. Bionetworks, "Sage bionetworks," *Synapse*. [Online]. Available: https://www.synapse.org/#!Synapse:syn2813589/wiki/401435

[10] S. S. Ahmed, S. Roy, and J. Kalita, "Assessing the effectiveness of causality inference methods for gene regulatory networks," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 17, no. 1, pp. 56–70, 2020.

[11] A. Aalto, L. Viitasaari, P. Ilmonen, L. Mombaerts, and J. Gonçalves, "Gene regulatory network inference from sparsely sampled noisy data," *Nature News*, Jul 2020. [Online]. Available: https://www.nature.com/articles/s41467-020-17217-1#citeas

[12] T. D. Le, T. Hoang, J. Li, L. Liu, and S. Hu, "Parallelpc: an r package for efficient constraint based causal exploration," 2015.

[13] J. D. Ramsey, "A scalable conditional independence test for nonlinear, non-gaussian data," 2014.

[14] I. Priness, O. Maimon, and I. Ben-Gal, "Evaluation of gene-expression clustering via mutual information distance measure," *BMC Bioinformatics 8*, p. 111, 2007.

[15] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. D. Favera, and A. Califano, "Aracne: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context," *BMC Bioinformatics 7*, 2006.

[16] P. E. Meyer, F. Lafitte, , and G. Bontempi, "minet: a r/bioconductor package for inferring large transcriptional networks using mutual information," *BMC Bioinformatics 9*, p. 461, 2008.

[17] X. Zhang, J. Zhao, J.-K. Hao, X.-M. Zhao, and L. Chen, "Conditional mutual inclusive information enables accurate quantification of associations in gene regulatory networks," *Nucleic Acids Research*, vol. 43, no. 5, pp. e31–e31, 12 2014. [Online]. Available: https://doi.org/10.1093/nar/gku1315

[18] Chicharro, Daniel, and S. Panzeri, "Algorithms of causal inference for the analysis of effective connectivity among brain regions," *Frontiers in Neuroinformatics*, vol. 8, pp. 98–104, 2014.