MSDS 452: Graphical, Network, and Causal Models

Checkpoint-A

January 29, 2026

Causal inference for drug resistance biomarker discovery

Yingyu Mao

Abstract

I will explore the application of Causal Bayesian Networks (BN) and Causal Structural Discovery (CSD) to infer the root causes of drug resistance from observational transcriptomic data. Using a breast cancer study focused on chemotherapy resistance, the research aims to distinguish between genes that are causes (upstream) and effects (downstream) of resistance by establishing specific causal relationships. I also plan to compare traditional heuristic algorithms, such as Markov chain Monte Carlo (MCMC) methods, against modern continuous optimization frameworks like NOTEARS to find optimal network structures.

Results will be evaluated against established biological interaction databases using metrics such as precision, recall, Structural Hamming Distance (SHD), and Structural Intervention Distance (SID). Ultimately, this research seeks to identify super nodes within gene regulation networks to reveal potential therapeutic targets for combined treatment regimes.

Keywords: Biomarker Discovery, Drug Resistance, Causal Bayesian Networks, Causal Structural Discovery

1.0 Introduction

Biomarkers are measurable indicators (such genes, proteins, molecules, cellular, or systematic changes) in biology samples that reveal normal biological processes, disease states, or treatment responses. In the realm of modern translational medicine, biomarkers facilitate clinical diagnostics and therapeutic monitoring. In precision medicine, they provide the molecular resolution necessary to tailor treatments to the individual, ensuring higher efficacy and reduced toxicity. In drug discovery, some of them are even direct targets of new therapeutics.

Identifying these disease-specific biomarkers become critical for diagnosis, prognosis, understanding the disease mechanism and formulating corresponding therapeutic plans. For instance, B-type natriuretic peptide (BNP) is used as a standard to diagnose heart failure and assess the severity of ventricular dysfunction. BRCA1/2 mutations indicate a high risk for breast/ovarian cancer and predict sensitivity to PARP inhibitors (Turk and Wisinski 2018). In the context of understanding drug resistance mechanisms of malignancies, biomarkers can be genes or a class of genes that enable genome plasticity in cancer cells or allow the cells to bypass canonical signal transduction or metabolic pathways. In this paper, I will explore causal inference methodologies to facilitate biomarker discovery and provide understandings in drug resistance mechanisms.

2.0 Literature Review

The conventional methodology on biomarker discovery for drug resistance in cancer came out of correlation based experimental design and statistical analysis. The main theme was to identify a set of genes that are differentially expressed in the resistant population in comparison to the sensitive population (Miri, et al. 2023, Lombard, et al. 2021, Barrón-Gallardo, et al. 2022). These can serve as features for building a prediction model for disease outcomes. Alternatively, follow-up controlled experiments were performed to further verify the true biomarker (secondary screening). Later, perturbation-based screening studies (e.g. CRISPR screen) provided direct experimental evidence for genes that confer drug resistance or sensitivity within various disease ex vivo models, some of which are translatable to patients (Doench, et al. 2016 , Zhong 2024, Szlachta, et al. 2018).

Conducting controlled experiments is the classic reductionist methodology to eliminate confounders and establish causal relationships. However, there can be situations where only observational data are available, such as in clinical settings where a perfectly controlled experiment is improbably due to patient heterogeneity, or the ex vivo model (tumor-derived cell culture) is unsustainable for the scale of perturbation experiments. How do we infer the root cause of a drug resistance just from observational data?

Causal inference aiming to infer a true mechanism of cause and effect from correlations provides a solution. At the heart of this statistical framework is the Causal Bayesian

Network, a combination of Bayesian probability and graph theory that uses a Directed Acyclic Graph (DAG) to map out the underlying data generation process. In these networks, variables are nodes, and causal relationships are directed edges. Being acyclic avoids the paradox of circular reasoning.

Traditionally, researchers relied on human intuition and domain expertise to draw the DAG. Once the model was set, the researcher would use tools like propensity score matching or instrumental variables to calculate the strength of the effect.

As datasets grew more complex for human intuition to map (such as omics datasets in biology), the field evolved toward Causal Structural Discovery (CSD), where the goal is to find the most optimal Bayesian network with a given set of nodes that represent correlated features. The challenge of this optimization problem is that the number of possible network structures grows exponentially with the number of nodes ($2^{n(n-1)}$ for directed graph). Markov chain Monte Carlo (MCMC) methods such as Metropolis–Hastings or Gibbs sampling algorithm was used to search the solution space of the Bayesian network that models aberrant signal transduction network (Neapolitan and Jiang 2014). Neighborhood sample and Hit-and-Run sampler were also tested in finding drug resistance driver genes by studying the transcriptomes of resistant versus sensitive cell lines (Azad and Alyami 2021).

As an alternative to local heuristics approach, the CSD field is moving toward continuous optimization, exemplified by frameworks like NOTEARS (Zheng, et al. 2018). By turning the acyclicity of a graph into a continuous representation, these methods then use classic gradient descent to find the optimal causal structures.

3.0 Methods

3.1 Dataset

I selected a breast cancer study that focused on the transcriptome (gene expression levels) for identifying biomarkers related to chemotherapy resistance (Barrón-Gallardo, et al. 2022). The original goal was to build a predicatory model based on genes identified with differential gene expression (DEG) analysis (such as edgeR (Robinson, McCarthy and Smyth 2010)) by comparing the population achieved pathological complete recovery and that with residual disease after chemotherapy. Machine learning models built upon these features achieved 83-86% accuracy of treatment outcomes in another follow-up confirmation study (Chen, et al. 2022).

3.2 Network model

The causal Bayesian network (BN) will have nodes to represent genes, and edges to model the relationship between the genes. The differentially expressed genes are correlated to drug resistance. Some of these genes lie upstream of resistance (the cause of resistance); others are downstream of resistance. The former are the nodes of interest

in building the BN. Causal inference can distinguish these two groups of genes by establishing three forms of relationships among the resistance (R), upstream genes (X), and downstream genes (Y) of R: $X \rightarrow R \leftarrow Y$, $X \rightarrow R \rightarrow Y$, $X \leftarrow R \leftarrow Y$.

Once the set of causal genes is established, I plan to use heuristics algorithms with MCMC as well as the newer continuous optimization to find the optimal BN.

4.0 Expected Results

I expect to learn traditional causal inference methods through identifying the upstream genes of resistance, and causal structure algorithms with the filtered set of genes.

To evaluate the results, I will need to extract gene/protein interaction networks from the published databases such as KEGG, CORUM, or STRING. The accuracy of the BNs can be assessed by calculating the precision, recall, or F1 of their edges with those established through learning literatures of experimentally identified edges. Structural hamming distance (SHD) and structural intervention distance (SID) are also insightful metrics (Chevalley, et al. 2022).

5.0 Conclusions

Drug resistance could be achieved through the combined effect of several signaling pathways. The ideal scenario is to identify the super nodes of this gene regulation network. This would reveal the next therapeutic target for combined treatment regimes.

I'm still researching ways to distinguish the type of causal link between genes, either activation or inhibition. It might be irrelevant in the DAG of a BN. I think I need more understanding of the causal inference methods to resolve this confusion.

DAG causal logic also does not deal with feedback loops in biological networks. When interpreting the result, this should be taken into consideration. Experimental confirmation might still be needed on a smaller set of super nodes.

References

Azad, AKM, and SA Alyami. 2021. "Discovering novel cancer bio-markers in acquired lapatinib resistance using Bayesian methods." *Brief Bioinform.* 22(5):bbab137.

Barrón-Gallardo, C, M Garcia-Chagollán, AJ Morán-Mendoza, R Delgadillo-Cristerna, MG Martínez-Silva, MM Villaseñor-García, A Aguilar-Lemarroy, and LF Jave-Suárez. 2022. "A gene expression signature in HER2+ breast cancer patients related to neoadjuvant chemotherapy resistance, overall survival, and disease-free survival." *Front Genet.* 13:991706.

Chen, J, L Hao, X Qian, L Lin, Y Pan, and X Han. 2022. "Machine learning models based on immunological genes to predict the response to neoadjuvant therapy in breast cancer patients." *Front Immunol.* 13:948601.

Chevalley, Mathieu, Yusuf Roohani, Arash Mehrjou, Jure Leskovec, and Patrick Schwab. 2022. "CausalBench: A Large-scale Benchmark for Network Inference from Single-cell Perturbation Data." *arXiv* 2210.17283.

Doench, JG Fusi, N, M Sullender, M Hegde, EW Vaimberg, KF Donovan, I Smith, Z Tothova, et al. 2016 . "Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9." *Nat Biotechnol.* doi: 10.1038/nbt.3437.

Lombard, AP, W Lou, CM Armstrong, LS D'Abronzo, S Ning, CP Evans, and AC Gao. 2021. "Activation of the ABCB1 Amplicon in Docetaxel- and Cabazitaxel-Resistant Prostate Cancer Cells." *Mol Cancer Ther.* 20(10):2061-2070.

Miri, A, J Gharechahi, Mosleh I Samiei, K Sharifi, and V Jajarmi. 2023. "Identification of co-regulated genes associated with doxorubicin resistance in the MCF-7/ADR cancer cell line." *Front Oncol.* 13:1135836.

Neapolitan, R, and X Jiang. 2014. "Inferring Aberrant Signal Transduction Pathways in Ovarian Cancer from TCGA Data." *Cancer Inform* 13(Suppl 1):29–36.

Robinson, MD, DJ McCarthy, and GK Smyth. 2010. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." *Bioinformatics* 26(1):139-40.

Szlachta, K, C Kuscu, T Tufan, SJ Adair, S Shang, AD Michaels, MG Mullen, et al. 2018. "CRISPR knockout screening identifies combinatorial drug targets in pancreatic cancer and models cellular drug response." *Nat Commun.* 9(1):4275.

Turk, AA, and KB Wisinski. 2018. "PARP inhibitors in breast cancer: Bringing synthetic lethality to the bedside. Cancer." *Cancer* 124(12):2498-2506.

Zheng, Xun, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. 2018. "DAGs with NO TEARS: Continuous Optimization for Structure Learning." *arXiv* 1803.01422.

Zhong, C., Jiang, WJ., Yao, Y. et al. 2024. "CRISPR screens reveal convergent targeting strategies against evolutionarily distinct chemoresistance in cancer. ." *Nat Commun* 15, 5502.