MSDS 452: Graphical, Network, and Causal Models

Checkpoint-B

February 12, 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.

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 the root cause 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 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 resistance to a drug 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 conditional independence tests are used to search for evidence in the data to refute the model. If no evidence of dependence is detected, the causal model is accepted.

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 dependent features. The challenge of this optimization problem is not only that the number of possible network structures grows exponentially with the number of nodes ($2^{n(n-1)}$ for directed graph), but also the solution space is discrete. 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 sampler 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).

There are many limitations with causal discovery algorithms. Therefore, the strategy for causal modeling should be utilizing causal discovery to narrow down candidate graphs to be used in subsequent causal inference testing, while blending in human oversight and domain expertise.

3.0 Methods

3.1 Background on RNA sequencing

The transcriptome (a collection of total ribonucleic acids, RNA) of a biological sample is sampled through a technique called RNA sequencing. Typically, there are 30-50 million read fragments of 70-100 bases in length sequenced for a single sample, called a library. Different fragments align to different parts of the genome and occur various times (counts) during the sampling. The random variables, the read counts for each expressed gene (RNA) i in library t, Y_{it}, follows a negative binomial distribution $Y_{it} \sim NB(\mu_i = m_t\lambda_{it}, \phi_i)$, where m_t is the library size, λ_{it} is the proportion of the library that is a particular RNA, and ϕ_i is the dispersion of each gene, giving $var(Y_i) = \mu_i + \phi_i\mu_i^2$. The first term of variance arises from Poisson distribution which captures the variance of the sequencing sampling process, while the second term models the overdispersion commonly seen in biological experiments.

After estimating the parameters of the distributions of each random variable, differential gene expression is identified through a variant of Fisher's exact test with NB distribution instead of hypergeometric distribution. For independence test used in causal model refusal, negative binomial regression can be used.

3.2 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).

The dataset has raw count data of 39376 gene features for 22 individuals with different treatment outcomes: R for resistance to chemotherapy and S for sensitive to chemotherapy.

3.3 Network model

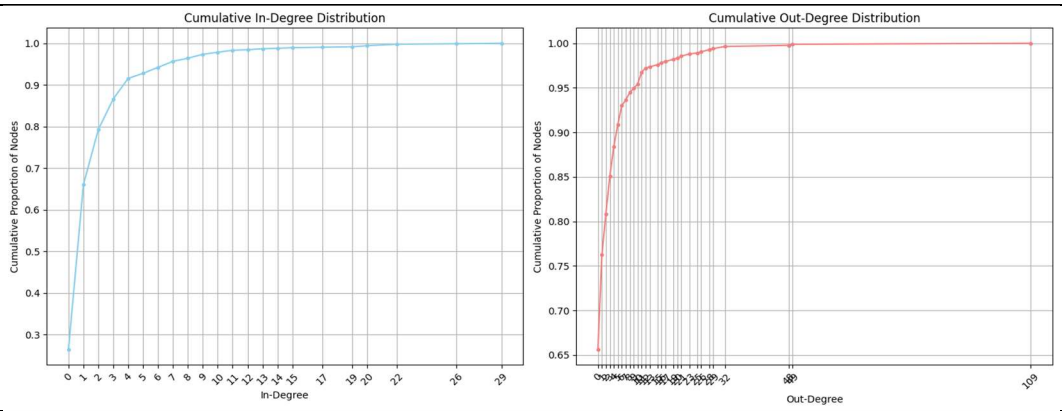
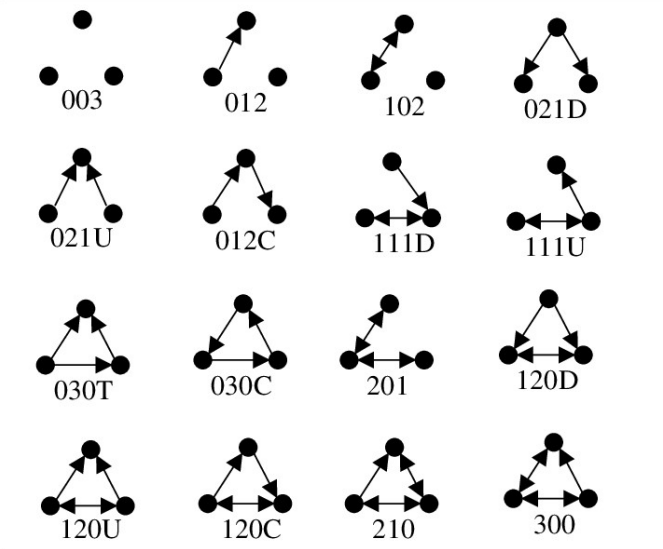
The causal DAG will have nodes to represent genes, and edges to model the relationship between the genes. The differentially expressed genes that are statistically significant from the NB exact test through comparing the resistant vs the sensitive samples are dependent on the status of the resistance outcome. In other words, there is at least one path from these genes to the outcome in the causal DAG. Some of these genes lie upstream of resistance (the cause of resistance); others are downstream of resistance (the effect of resistance). Causal discovery algorithms impose constraints on the DAG with evidence of conditional independence in the data with the assumption that the reverse process of causal inference holds: conditional independence in the joint distribution implies d-separation in the graph, and colliders $X \rightarrow R \leftarrow Y$ are detectable with evidence of dependence and independence alone. The latter is especially relevant to the use case for new drug discovery targeting the biomarker, where identifying the cause of drug resistance (upstream effectors) is more important than the result of drug resistance.

4.0 Preliminary Results

4.2 Features of established gene regulatory networks.

I was first interested in characterizing the graph features of the gene regulatory network (GRN). This can help me establish necessary constraints on optimizing the DAG.

The GRN was extracted from Kyoto Encyclopedia of Genes and Genomes (KEGG), a knowledge base of gene interactions based on manual learning of experimental data in literature. Its characteristics are summarized below:

Number of nodes	829
Number of edges	1569
Edge density	0.0023
Average clustering coefficient	0.0177
DAG?	No, have 5 simple cycles
Average shortest path	2.0
In-degree and out-degree distribution	
Triad census	{'003': 93336828, '012': 1251554, '102': 776, '021D': 15144, '021U': 4285, '021C': 1658, '111D': 4, '111U': 45, '030T': 157, '030C': 1, '201': 0, '120D': 0, '120U': 1, '120C': 1, '210': 0, '300': 0}	 (Uddin, Hossain and Khan 2018)

4.2 Identifying differentially expressed genes (DEG)

The `edgeR` module uses trimmed mean of M-values normalization method to estimate the (relative) effective size of each library (m_t) in an experiment (Robinson and Oshlack 2010). Under the assumptions that most genes in a transcriptome are not DEG and gene of lower expression are more likely noise, this method trims the gene counts with variance in the upper and lower 30% of the data and those with absolute expression levels across libraries in the lower 5% of the data (these are default values and can be adjusted).

The common dispersion parameter ϕ (assuming all was estimated using a quantile-adjusted conditional maximum likelihood procedure to account for the bias introduced by estimating the gene count mean μ and adjust the observed counts up or down depending on whether the corresponding library sizes are below or above the geometric mean, as described by Robinson and Smyth (2008).

The gene-wise dispersion parameter ϕ_i is then estimated through "squeezing" toward the common dispersion through a weighted conditional log-likelihood, which is a weighted combination of the individual and common likelihoods (Robinson and Smyth 2007).

Running the module on the breast cancer dataset returned Benjamini-Hochberg adjusted p-value statistic for each gene. A set of significant DEGs were identified as the top 100 genes with lowest adjusted p-values (results in R notebook).

4.3 Causal discovery

4.4 Model evaluation

To evaluate the results, I will use extracted gene regulatory network from KEGG as reference. The accuracy of the DAG 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. With this goal, I think identifying colliders is more important than building out the entire DAG of 100 nodes. Some of the conditional independence relationships cannot infer edge directions, resulting in graphs of Markov equivalence class.

DAG causal logic does not deal with feedback loops in biological networks. However, this might not pose an issue as the 829-node GRN has only 3 triad cycles and some

multi-edge loops. For 40 nodes, I expect the numbers to be less. Therefore, assuming DAG is a good approximation of the true GRN.

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, and Alicia Oshlack. 2010. "A scaling normalization method for differential expression analysis of RNA-seq data." *Genome Biology* 11:R25.

Robinson, MD, and GK Smyth. 2007. "Moderated statistical tests for assessing differences in tag abundance." *Bioinformatics* 23(21):2881–2887.

Robinson, MD, and GK Smyth. 2008. "Small-sample estimation of negative binomial dispersion, with applications to SAGE data." *Biostatistics* 9,2, pp. 321–332.

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.

Uddin, S., M. E. Hossain, and A. Khan. 2018. "Triad Census and Subgroup Analysis of Patient-Sharing Physician Collaborations." *IEEE Access* vol. 6, pp. 72233-72240.

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.