Causal Discovery on Gut Microbial Data for Disease Risk Prediction

Mariana Paco Mendivil [1] Candus Shi [2] Nicole Zhang [3] Mentor: Dr. Biwei Huang [4] Mentor: Dr. Jelena Bradic [5]

[1]mpacomendivil@ucsd.edu [2]c6shi@ucsd.edu [3]nwzhang@ucsd.edu [4]bih007@ucsd.edu [5]jbradic@ucsd.edu

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Background

- **Association vs. Causation:** In many domains, researchers are often concerned with finding the underlying structure that generates the data we observe. Traditionally, we have used statistical methods and models that allow us to make associative conclusions about the hypothetical population from an observed sample. However, associations are often insufficient to answer our scientific questions, and randomized experiments can be expensive, slow, and unethical to conduct in different domains. Causal discovery and causal inference are a set of methods and models that attempt to causally answer these scientific questions given the limitations of the observed data [1].
- **Gut Microbiome:** The gut microbiome has been shown to be an important indicator of human health, and extensive research has been conducted to explore its impact on human health and disease. Given its diverse composition, the gut microbiome is a complex area of study as it can have heterogeneous effects for different populations [2].
- **Causal Discovery and Inference in the Gut Microbiome:** Previous research has explored causal discovery in gut microbiome studies, notably using algorithms like PC-stable to construct causal networks and implement do-calculus for estimating microbe-microbe and microbe-outcome causal effects [3]. More generally in the field of causal discovery, recent advancements include CD-NOD, an algorithm specifically designed for heterogeneous data, which is particularly valuable for gut microbiome research where samples often come from different studies [4].

Research Questions

1. **Microbe-Microbe:** How do the microbe-microbe interaction networks between the healthy and diseased participants differ?
2. **Microbe-Disease:** What microbes have a causal relationship to disease status?
3. **Prediction:** Is it possible to predict disease status with the current composition of the dataset given causal representation learning techniques? How do they differ with the microbes learned in question 2?

Data

To answer the questions above, we apply our framework to gut microbial data that investigated T2D and an individual participant data meta-analysis dataset that investigated PCOS.

- **T2D:** For T2D, we use the NIH Human Microbiome Project (HMP2) dataset [5], filtered to healthy visits with 16S sequencing. Includes 153 insulin-sensitive (IS) and 178 insulin-resistant (IR) samples.
- **PCOS:** For PCOS, we use a dataset aggregated from 14 different clinical studies across Asia and Europe [6], filtered to individual-level samples with 16S sequencing. Includes 435 healthy controls (HC) and 513 PCOS patients.

Causal Discovery

Causal discovery attempts to recover the true causal structure of a system given observed data. One way to model this causal structure is through a directed graphical model. A widely-used general-purpose causal discovery algorithm is the Peter-Clark (PC) algorithm [7]. It follows these key steps:

1. Start with a **complete undirected graph** (each node connected to all other nodes).
2. **Remove edges** based on statistical independence and conditional independence tests.
3. **Identify v-structures** (patterns like $X \to Y \leftarrow Z$) to infer causal directions.
4. **Apply Meek's rules** to orient additional edges while preserving v-structures.

The result is a **CPDAG (Completed Partially Directed Acyclic Graph)**, which represents a set of causal structures consistent with the observed data, also known as the Markov Equivalence Class (MEC).

Why Use PC?

- Works for different data types (as long as independence tests match the data distribution).
- Efficient for large datasets.
- Assumes the **causal Markov condition**, the **faithfulness** assumption, and **no hidden confounders**,.

Methods

In this study, we use causal discovery algorithms and compare them with predictive modeling to explore the causal relationships between the gut microbiome and two diseases: T2D and PCOS. Due to the high-dimensionality of the data and small sample sizes, we first select features through sparse estimation methods and sure-screening to reduce the number of microbes.

1. **Filter out rare OTUs.** Remove microbes where all samples have less than 1% relative abundance.
2. **Feature selection and sure screening.** For the microbe-microbe network, we use two methods, SparCC [8] and graphical lasso to reduce the number of edges between pairs of microbes. For the microbe-disease network, we use logistic lasso regression to reduce the number of features that are not helpful in predicting disease.
3. **Causal discovery algorithms.** For the microbe-microbe network, we implement PC-stable with a max depth of 2. For the microbe-outcome network, we implement CD-NOD where the covariates correspond to the heterogeneity index.
4. **Variational autoencoder.** xxx. Formulas.

T2D

(Report results for microbe-microbe network). From the microbe-outcome network (Figure 1), the following five genera are causal to T2D ('IRIS' node): *Butyricimonas*, *Clostridium XIVb*, *Odoribacter*, *unclassified Bacteria*, and *unclassified Firmicutes*. To further investigate their individual effects, we implement do-calculus through logistic regression models on T2D given the neighbors of the genus of interest (Table 1).

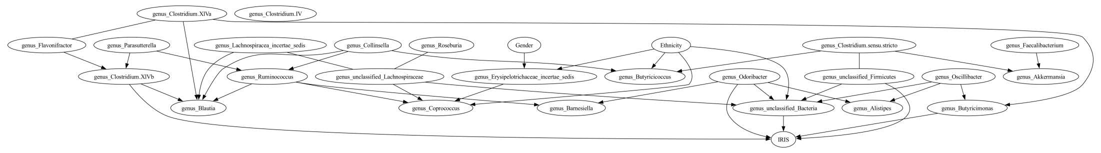

Figure 1. Microbe-Disease Network for T2D.

Genus	Odds Ratio	P-Value	Literature Agreement
Butyricimonas	0	0	Unknown
Clostridium XIVb	0	0	Unknown
Odoribacter	0	0	Unknown
unclassified Bacteria	0	0	Unknown
unclassified Firmicutes	0	0	Unknown

Table 1. Do-Calculus Results for T2D.

(Insert VAE results).

PCOS

(Report results for microbe-outcome network). From the microbe-outcome network (Figure 2), the following nine genera are causal to PCOS ('group' node): *Alistipes*, *Blautia*, *Burkholderia*, *Desulfovibrio*, *Holdemanella*, *Knoellia*, *Prevotellaceae NK3B31 group*, *Ruminococcus*, and *Ruminococcus gnavus group*. We find their individual causal effects with do-calculus (Table 2).

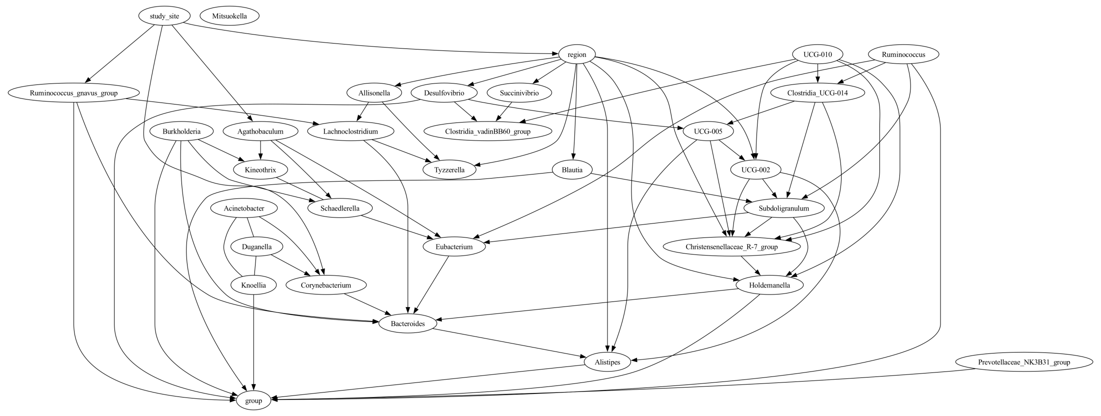

Figure 2. Microbe-Disease Network for PCOS.

Genus	Odds Ratio	P-Value	Literature Agreement
Alistipes	0.153458	4.68e-05	Unknown
Blautia	0	0	Unknown
Burkholderia	0	0	Unknown
Desulfovibrio	0	0	Unknown
Holdemanella	0	0	Unknown
Knoellia	0	0	Unknown
Prevotellaceae NK3B31 group	0	0	Unknown
Ruminococcus	0	0	Unknown
Ruminococcus gnavus group	0	0	Unknown

Table 2. Do-Calculus Results for PCOS.

(Insert VAE results).

Conclusion & Future Work

Answer questions 1, 2, and 3. Explain BIRDMAn. Point to website for more results. Fix references to et al for many authors.

References

[1] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal Inference in Statistics—A Primer*. John Wiley & Sons Ltd, 2016.

[2] Daniel P. Baars, Marcos F. Fondevila, Abraham S. Meijnikman, and Max Nieuwdorp. The central role of the gut microbiota in the pathophysiology and management of type 2 diabetes. *Cell Host & Microbe*, 32(8):1280–1300, 2025/02/09 2024.

[3] Musfiqur Sazal, Vitalii Stebliankin, Kalai Mathee, Changwon Yoo, and Giri Narasimhan. Causal effects in microbiomes using interventional calculus. *Scientific Reports*, 11(1):5724, 2021.

[4] Biwei Huang, Kun Zhang, Jiji Zhang, Joseph D Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data with independent changes. *CoRR*, abs/1903.01672, 2019.

[5] Wenyu Zhou, M. Reza Sailani, Kévin Contrepois, Yanjiao Zhou, Sara Ahadi, Shana R. Leopold, Martin J. Zhang, Varsha Rao, Monika Avina, Tejaswini Mishra, Jethro Johnson, Brittany Lee-McMullen, Songjie Chen, Ahmed A. Metwally, Thi Dong Binh Tran, Hoan Nguyen, Xin Zhou, Brandon Albright, Bo-Young Hong, Lauren Petersen, Eddy Bautista, Blake Hanson, Lei Chen, Daniel Spakowicz, Amir Bahmani, Denis Salins, Benjamin Leopold, Melanie Ashland, Orit Dagan-Rosenfeld, Shannon Rego, Patricia Limcaoco, Elizabeth Colbert, Candice Allister, Dalia Perelman, Colleen Craig, Eric Wei, Hassan Chaib, Daniel Hornburg, Jessilyn Dunn, Liang Liang, Sophia Miryam Schüssler-Fiorenza Rose, Kim Kukurba, Brian Piening, Hannes Rost, David Tse, Tracey McLaughlin, Erica Sodergren, George M. Weinstock, and Michael Snyder. Longitudinal multi-omics of host–microbe dynamics in prediabetes. *Nature*, 569(7758):663–671, 2019.

[6] Yanan Yang, Jiale Cheng, Chongyuan Liu, Xiaopo Zhang, Ning Ma, Zhi Zhou, Weiying Lu, and Chongming Wu. Gut microbiota in women with polycystic ovary syndrome: an individual based analysis of publicly available data. *eClinicalMedicine*, 77, 2025/01/19 2024.

[7] Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10, 2019.

[8] Jonathan Friedman and Eric J Alm. Inferring correlation networks from genomic survey data. *PLoS Comput Biol*, 8(9), 2012.

Please visit our website at https://nzhang20.github.io/Causal-Discovery-on-Gut-Microbial-Data-for-Disease-Risk-Prediction/ or at the above QR code.