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

- **Causal Discovery & Causal Inference:** These are a set of methods and models that attempt to causally answer scientific questions using observed data rather than RCTs.
- **Gut Microbiome:** The gut microbiome is an important indicator of human health, and extensive research is ongoing to explore its impact on human health and disease.
- **Causality in the Gut Microbiome:** Most studies report associations, but these are often insufficient to answer the scientific question of interest.

Research Questions

1. **Microbe-Microbe:** How do the microbe-microbe interaction networks between healthy and diseased participants differ?
2. **Microbe-Disease:** Which microbes have a causal relationship to disease status, and how is it quantified?
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

- **T2D:** NIH Human Microbiome Project (HMP2) dataset, filtered to healthy visits with 16S sequencing. Includes 153 insulin-sensitive (IS) and 178 insulin-resistant (IR) samples.
- **PCOS:** Meta analysis dataset from 14 different clinical studies across Asia and Europe. 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. 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 $X \rightarrow Y \leftarrow Z$) 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).

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 SparCC 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 (a variant of PC) where the covariates correspond to the heterogeneity index.
4. **Causal inference**. We estimate the causal effects of microbes using do-calculus, and compare the results with Bayesian Inferential Regression for Differential Microbiome Analysis (BIRDMAn).
5. **Variational autoencoder**. xxx. Formulas.

T2D

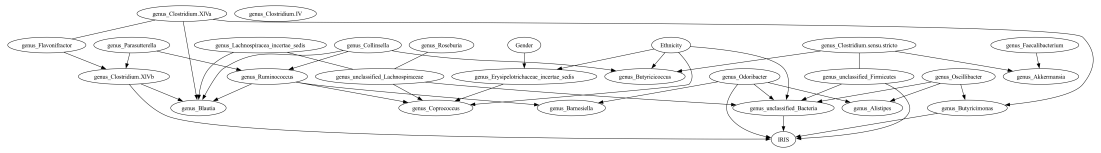

Figure 1. Microbe-Disease Network for T2D.

Model 1: logit(disease status) ∼ microbes directly linked
Model 2: logit(disease status) ∼ microbe + neighbors(microbe) or mediators
BIRDMAn: Bayesian inference with $NegBinomial(\mu, \phi)$ where μ is the mean count and ϕ is the dispersion

Genus	Model 1	Model 2	BIRDMAn	Literature Agreement
Butyricimonas	-2.0070*	-2.26645*	-5.19385*	Yes
Clostridium XIVb	1.54212*	1.80822*	2.15788*	Inconclusive
Odoribacter	-1.46989*	-3.055047*	-2.43796*	Yes
unclassified Bacteria	-0.12991*	-0.12284*	0.12409	N/A
unclassified Firmicutes	-0.69477*	-0.933718*	-1.47437*	N/A

Table 1. Do-Calculus Results for T2D.

Table 1. shows the log odds ratio for Model 1 and Model 2, and the mean CLR from BIRDMAn. Significant values are denoted with an asterisk (*).

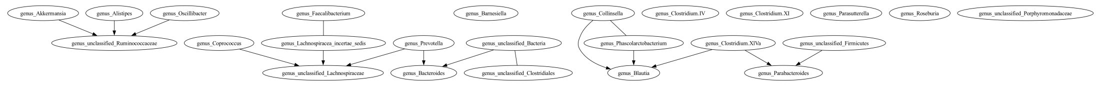

Figure 2. Microbe-Microbe Network for T2D, IS cohort.

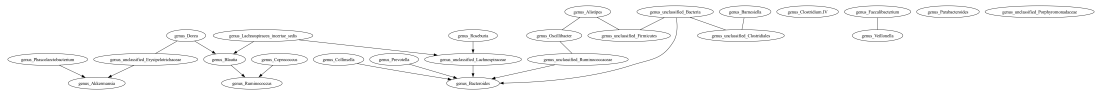

Figure 3. Microbe-Microbe Network for T2D IR cohort.

We generate similar graphs for the two microbe-microbe networks using PC with a max depth of 2 (Figure 2 and 3), and find that the two cohorts share 21 microbes. IS have an additional three: *Clostridium.XI*, *Clostridium.XIVa*, and *Parasutterella*; IR have an additional four: *Dorea*, *Ruminococcus*, *Veillonella*, and *unclassified Ersipelotrichaceae*.

(Insert VAE results).

PCOS

From the microbe-disease 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).

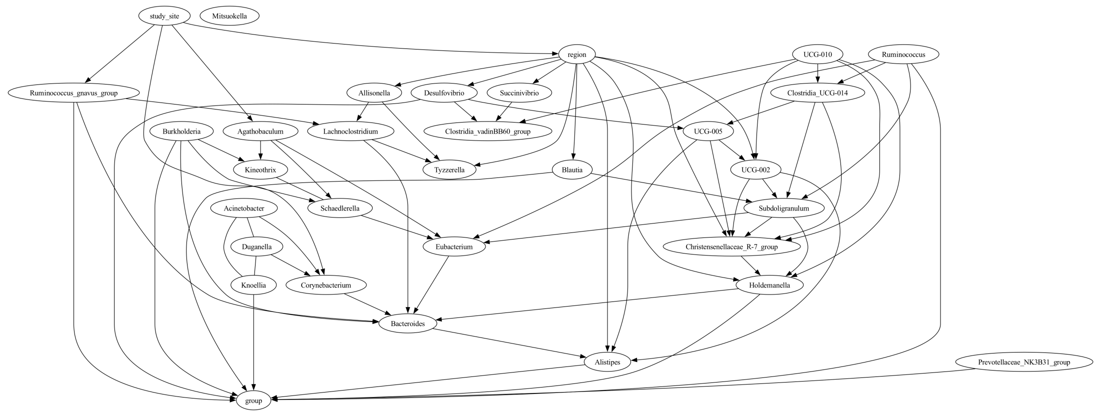

Figure 4. Microbe-Disease Network for PCOS.

Genus	Model 1	Model 2	BIRDMAn	Literature Agreement
Alistipes	0.13272*	0.15346*	0	Unknown
Blautia	0.07461*	0.07008*	0	Unknown
Burkholderia	-7.60599	-0.48578	0	Unknown
Desulfovibrio	-0.79283*	-1.14492*	0	Unknown
Holdemanella	-0.22801*	-0.17267*	0	Unknown
Knoellia	592.26751	1.40864	0	Unknown
Prevotellaceae NK3B31 group	-0.42407	-0.47231*	0	Unknown
Ruminococcus	-0.14137*	-0.13490*	0	Unknown
Ruminococcus gnavus group	0.24152*	0.18259*	0	Unknown

Table 2. Do-Calculus Results for PCOS.

For the microbe-microbe networks, we see that the two cohorts share xxx, but differ in xxx. Please see our website for these visualizations.
(Insert VAE results).

Conclusion & Future Work

1. **Microbe-Microbe:** Healthy and diseased participants share certain microbes, but also differ on which microbes are present and how they interact with other microbes. It is important to consider these microbes as communities of organisms rather than singular entities.
2. **Microbe-Disease:** Using CDNOD and do-calculus, we are able to quantify the effects of microbes on disease status, and they agree with microbiome-specific differential analysis methods such as BIRDMAn, and they also agree with current literature.
3. **Prediction:** TBD

We hope this project shows the potential of causal discovery and causal inference methods in human gut microbiome research, and can be generically applied to other diseases of interest. We would like to thank our mentors, Dr. Biwei Huang & Dr. Jelena Bradic, and Dr. Sam Degregori & the Knight Lab for guidance throughout this project.

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.