Signed Causal Bayesian Networks for Microbiomes

Anonymous Author(s)

Abstract

Inferring causality is the process of connecting causes with effects. Identifying even a single causal relationship from data is more valuable than observing dozens of correlations in a data set. Microbe-microbe and host-microbe interactions play a vital role in both health and disease. In this study, we investigate how to learn a causal structure from data from microbiome studies and its potential interpretation about events and processes in the microbial community under study. We report evidence that causal structure can extract colonization patterns even though the analysis only uses data with no temporal information.

1 Introduction and Motivation

Causation is an important type of relationship to be explored with biological data. Thus, it makes sense to see if causal Bayesian networks can identify relationships that are suggestive of causation, leading to lab experiments for validation. Bayesian networks (BNs) were used by Zhang et al. to understand changes in gene regulatory networks (1), and Sazal et al. used BNs to understand relationships among taxa in microbiomes (2). By modeling metabolic reactions and their involvement in multiple subnetworks of "metabosystems", Shafiei et al. used BNs to infer differential prevalence of metabolic subnetworks within microbial communities (3).

A *microbiome* is a community of microbes including bacteria, archaea, protists, fungi and viruses that share an environmental niche (4). Microbiomes can be modeled as a *social network* because of the complex set of potential interactions between its various taxonomic members (5; 6). To understand potential interactions between taxa in a microbial community, the construction of co-occurrence networks (CoN) was proposed by Fernandez et al. (5) and Faust et al. (7). The results suggested that the reason groups of taxa frequently co-infect cohorts of subjects or did the opposite, i.e., co-avoided cohorts of subjects, was because of underlying interactions between them. Unfortunately, that is as far as CoNs are able to go in terms of inferring complex relationships in microbiomes.

In this paper, we investigate how to infer directional relationships between microbial taxa in a microbiome. In humans, normal microbial colonization starts from birth and with the passage of time these communities become relatively stable (8). Some microbes recruit others suggesting an order of colonization in many microbial communities. Understanding colonization and its order can provide a window into how infections take hold. We show that causal structure or *signed Causal Bayesian Networks* (scBNs), a variant of BNs obtained by combining BNs with Co-occurrence networks can help tease apart some of these directed relationships and provide a glimpse into the complex and dynamic world of microbial communities. Work is underway to investigate how to infer other causal relationships from the same data. In particular, our work will highlight the microbial players involved in recruiting other microbes, the key players in causing disease, their relative importance in the disease process, the role of beneficial microbes in alleviating disease symptoms, the role of metabolites in disease, the identification of potential targets for treatment, and much more.

Submitted to 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Do not distribute.

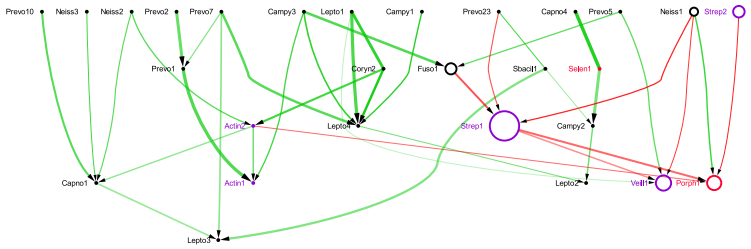

Figure 1: **Signed Causal Bayesian Network**. Nodes represent microbial taxa, edges represent the relationships among taxa, red and green edges represent negative and positive correlation respectively.

2 Methods

Causal structures (CS) are a class of *Probabilistic Graphical Models* (PGMs) (9; 10) where each node represents a random variable from a set, $\mathbf{X} = \{X_i, i = 1, \ldots, n\}$ with n random variables. These structures are represented as a graph $G = (V, E)$, where each vertex in V represents a random variable from $\mathbf{X}$, and E is the set of edges on V. The graph G is also known as a *causal Bayesian network* on X. Although undirected edges are used in cases where the direction cannot be reliably determined or when both directions appear to be valid, the graph G is often "manipulated" to be a Directed Acyclic Graph (DAG). Each random variable X_i has an associated probability distribution. A directed edge in E between two vertices represents direct stochastic dependencies. Therefore, if there is no edge connecting two vertices, the corresponding variables are either marginally independent or conditionally independent (conditional on the rest of the variables, or some subset thereof). To learn a causal structure we adopted a conditional independence test based method proposed by Spirtes et al. (11), later modified by Colombo and Maathuis to make it *order independent*, and known as *PC-Stable* algorithm (12). PC-stable consists of three steps - adjacency search in order to learn the "skeleton", identifying important substructures called v**-structures**, and detecting and orienting other arcs. In Step 1, the algorithm starts with a complete undirected graph and then performs a series of conditional independence tests to eliminate as many edges as possible. The remaining undirected graph is referred to as the *skeleton*. Step 2 is key to inferring a directional model, and uses the concept of v-structures (13). Step 3, three rules (12) are applied repeatedly to orient remaining undirected edges (i.e., arcs not in v-structures). Finally, we added sign information from CoNs to the edges in the causal structure.

We used oral data sets (16S rRNA sequences) generated as part of the Human Microbiome Project (HMP) from eight different sites within the oral cavity from 242 healthy adults (129 males, 113 females) (14; 4). The samples included: saliva, buccal mucosa (cheek), keratinized gingiva (gums), palatine tonsils, throat, tongue dorsum, and supra- and sub-gingiva dental plaque (tooth biofilm above and below the gum). Abundance of individual taxa were computed after amplification of a specific hypervariable region of the bacterial 16S rRNA gene, followed by sequencing, grouping reads into common Operational Taxonomic Units (OTUs) and quantification (15). Mothur (16) was used to compute the microbial abundance profile.

3 Results and Discussion

Figure 1 shows a signed causal structure learned from keratinized gingiva data set. The results with all oral data sets showed a surprising connection to the order in which microbes colonize the human mouth. Two significant observations were as follows. (1) The directed edges of scBN for the oral microbiome data set were consistent with the colonization order. A total of 716 edges were generated for the oral microbiome scBN with the colonization order known for 78 edges. Only 2 edges in the scBN were inconsistent with the known direction, Resulting in an accuracy of 97.4%. (2) scBN Edges with negative correlations were consistent with the colonization order (early to late colonizers). All directed edges between two taxa from two colonization groups were negatively correlated. Thus, the scBNs help us to infer potential relationships and dependencies within a microbiome, and the colonization order without time information. scBNs could help in understanding the other dependencies among the entities of a microbial community.

References

[1] Zhang XF, Ou-Yang L, Yan H. Incorporating prior information into differential network analysis using non-paranormal graphical models. Bioinformatics. 2017;33(16):2436–2445.

[2] Sazal MR, Ruiz-Perez D, Cickovski T, Narasimhan G. Inferring Relationships in Microbiomes from Signed Bayesian Networks. In: 2018 IEEE 8th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS). IEEE; 2018. p. 1–1.

[3] Shafiei M, Dunn KA, Chipman H, Gu H, Bielawski JP. BiomeNet: a Bayesian model for inference of metabolic divergence among microbial communities. PLoS computational biology. 2014;10(11):e1003918.

[4] Peterson J, Garges S, Giovanni M, McInnes P, Wang L, Schloss JA, et al. The NIH Human Microbiome Project. Genome research. 2009;19(12):2317–2323.

[5] Fernandez M, Riveros JD, Campos M, Mathee K, Narasimhan G. Microbial "social networks". BMC Genomics. 2015;16(11):1.

[6] Fernandez M, Aguiar-Pulido V, Riveros J, Huang W, Segal J, Zeng E, et al. Microbiome Analysis: State of the Art and Future Trends. Computational Methods for Next Generation Sequencing Data Analysis. 2016;p. 401–424.

[7] Faust K, Sathirapongsasuti JF, Izard J, Segata N, Gevers D, Raes J, et al. Microbial co-occurrence relationships in the human microbiome. PLoS Comp Bio. 2012;8(7):e1002606.

[8] Houghteling PD, Walker WA. Why is initial bacterial colonization of the intestine important to the infant's and child's health? Journal of pediatric gastroenterology and nutrition. 2015;60(3):294.

[9] Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference. Amsterdam, Netherlands: Elsevier; 2014.

[10] Koller D, Friedman N. Probabilistic graphical models: principles and techniques. Cambridge, MA: MIT press; 2009.

[11] Spirtes P, Glymour CN, Scheines R. Causation, prediction, and search. Cambridge, MA: MIT press; 2000.

[12] Colombo D, Maathuis MH. Order-independent constraint-based causal structure learning. The Journal of Machine Learning Research. 2014;15(1):3741–3782.

[13] Bui AT, Jun CH. Learning Bayesian network structure using Markov blanket decomposition. Pattern Recognition Letters. 2012;33(16):2134–2140.

[14] Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, et al. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486(7402):207.

[15] Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA. Sparse and compositionally robust inference of microbial ecological networks. PLoS computational biology. 2015;11(5):e1004226.

[16] Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing Mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and environmental microbiology. 2009;75(23):7537–7541.