Discovering Causal Relationships in Microbial Time Series Data

Mathilde Diekema

Supervisors: Thomas Bataillon
Leendert Vergeynst
Williane Vieira Macêdo

Department of Bioinformatics
Aarhus University

This thesis is submitted for the degree of
Master of Science in Bioinformatics

June 2023

Acknowledgements

I want to express my sincere gratitude to my supervisors, Thomas Bataillon, Leendert Vergeynst, and Williane Vieira Macêdo, for providing their invaluable guidance, comments, and suggestions throughout the project. I would especially like to thank Williane for her help with experimental questions and Thomas and Leendert for continually motivating me to move forward with their excellent overview, new ideas and essential questions. I extend my thanks to the research group who contributed new ideas and a great atmosphere. Everyone has been incredibly kind and helpful. Finally, I would like to thank my boyfriend Emil for all his love and support.

Abstract

A revolution in DNA sequencing technologies has enhanced our ability to gain deeper insights into microbial community interactions enabling the exploration of new frontiers in environmental and medical microbiology. However, correlation-based association analysis used in most microbiome studies is noninformative of causal relationships between internal and external factors. To address this limitation, we explored the potential of a recently developed causal discovery method, PCMCI, to reconstruct the causal dependency graphs underlying microbial interactions. The PCMCI framework was tested together with the conditional independence test ParCorr on real-world data and two artificial time series datasets with known dependencies that mimic the properties of real-world microbial data.

The real-world dataset was highly relevant regarding the microbial community in the anaerobic digestion of complex wastewater. However, the underlying experiment that created the data was not adapted to causal discovery in time series. To improve the data for future experiments, the results suggested a higher time resolution, larger sample size, and a system under steady-state conditions to obey the assumptions under which the underlying causal dependencies can be inferred.

The artificial data were simulated with a stochastic generalized Lotka-Volterra model that only allowed for weak interactions between the species in order for the time series not to be unstable. In the first experiment with synthetic data, three interacting species were simulated. For this dataset, PCMCI was not able to recover the underlying causal structure. The second synthetic dataset was simulated with an artificially created variable Z that depended on two noninteracting microbial species. This model system allowed for more significant interactions between the variables, and PCMCI captured the underlying causal structure with a true positive rate of 100 %. For both synthetic datasets, the false positive rate fulfilled the requirement of the significance level of 5 %, indicating a well-calibrated test due to fulfilled assumptions. As expected by the linear multiplicative noise term in the generalized Lotka-Volterra model, the type dependencies of the noise were heteroskedastic. This suggests using an adapted version of the ParCorr conditional independence test customized heteroskedastic data for future studies on causal discovery on microbial data with PCMCI.

Table of contents