Title: Identifying microbial drivers in biological phenotypes with a Bayesian Network Regression model

Authors: Samuel Ozminkowski[1]* (ozminkowski@wisc.edu) and **Claudia Solis-Lemus**[1]

Institutions: [1]University of Wisconsin-Madison, Madison, WI

Website: https://github.com/samozm/BayesianNetworkRegression.jl

Project Goals: Short statement of goals. (Limit to 1,000 characters)

1. Assess the applicability of a novel Bayesian Network Regression framework to microbiome applications.
2. Implement a fast and efficient sampling algorithm to sample posterior conditional distributions for the model.
3. Release an open-source and publicly available package in the Julia programming language so that domain scientists can utilize this model on their data.

Abstract text: Please limit such that entire document does not exceed 2 pages.

Microbial communities are among the main driving forces of biogeochemical processes in the biosphere. Understanding the composition of microbial communities and how these compositions shape specific biological phenotypes is crucial to comprehend complex biological processes in soil, plants and humans alike. Standard approaches to study the connection between microbial communities and biological phenotypes rely on abundance matrices to represent the microbial compositions. Different experimental settings are defined and then microbial compositions are measured (as abundances) on each experimental setting. Next, the abundance matrices are used as input in a regression-type (or machine-learning) analysis to relate the microbial community to phenotypes of interest.

Given that relative abundances only provide a snapshot of the composition of the community at the specific time of sampling and do not account for correlations between microbes, microbial interaction networks have been recently preferred to represent microbial communities. Yet models to connect a microbial network to a biological phenotype remain unknown. There has only been a handful of new methods that aim to identify associations between network predictors and a phenotype via a regression framework. However, these methods have only been studied for brain connectome networks which, unlike microbial networks, are intrinsically dense. In conclusion, methods to find associations between a sample of microbial networks and a biological phenotype remain unknown.

In this work, we introduce a Bayesian Network Regression (BNR) model that uses the microbial network as the predictor of a biological phenotype. This model intrinsically accounts for the interactions among microbes and is able to identify influential edges (interactions) and influential

nodes (microbes) that drive the phenotypic variability. While the model itself is not new, it has only been studied for brain connectome networks, and thus, its applicability to microbial networks which are inherently more high-dimensional and sparser has not been studied. Furthermore, unlike in brain connectome research, in microbiome research, it is usually expected that the presence of microbes have an effect on the response (main effects), not just the interactions. Here, we develop the first thorough investigation of whether Bayesian Network Regression models are suitable for microbial datasets on a variety of synthetic data that was generated under realistic biological scenarios. We test whether the Bayesian Network Regression model that accounts only for interaction effects (edges in the network) is able to identify key drivers in phenotypic variability (microbes). We show that this model is able to identify influential nodes and edges in the microbial networks that drive changes in the phenotype for most biological settings, but we also identify scenarios where this method performs poorly which allows us to provide practical advice for domain scientists aiming to apply these tools to their datasets. In addition, we implement the method in an open-source publicly available and easy-to-use new Julia package (BayesianNetworkRegression.jl) with online documentation and step-by-step tutorial which will allow scientists to easily apply this model on their own data.

This research directly addresses the DOE SC program goals of developing computational approaches that can integrate large omics data types from multiple and heterogeneous sources, such as those used in the Genomic Science program. Our open-source software will be incorporated into the DOE Systems Biology Knowledgebase, an open-source software platform that serves the systems biology research community.

References/Publications

1. Ozminkowski, Samuel and Claudia Solis-Lemus “Identifying microbial drivers in biological phenotypes with a Bayesian Network Regression model” (in preparation).

Funding statement: This research was supported by the DOE Office of Science, Office of Biological and Environmental Research (BER), grant no. DE-SC0021016.