Bayesian Networks in Microbiology

Breanna Shi, Carolyne Foster

University of Minnesota

Abstract:
In this study, we examine Bayesian inference. We consider an application where DNA samples are taken from the Rumen of cattle and sheep. Bayesian networks are a framework we used to illustrate the connections between the microbiome and antimicrobial resistance genes (AMR) that exist in nature. we examine how the hill-climbing algorithm is used to create a Bayesian network from microbiome and AMR counts. We will investigate the significance that Bayesian research has to global communities.

Background:
The microbiome is the totality of all microbial communities on a given organism. Microbial gene sequences can be obtained through 16S sequencing, and shotgun sequencing. 16S sequencing is useful for discovering rare microbiome genes, where shotgun sequencing provides more sequence data. This might allow us, for example, to sequence a certain bacteria's entire genome [4]. A Gene marker can be identified by a computational tool and may prevent a microbe from being misidentified. When we can identify microbes in a community, we can create counts of the microbiome [4]. These counts are what we analyze to build Bayesian networks.
The Bayesian networks are based on the study of 98 cattle and sheep from four different samples. A Bayesian network was produced with connections that appeared in 50 percent of the bootstraps. Bootstraps is a term used for the iterations of graphs tested by the Hill-Climbing algorithm. [2].

Bayesian Statistics:

Bayesian inference is based upon Bayes Theorem,

$$p(y) = \frac{f(\theta)\pi(\theta)}{\int(f(\theta)\pi(\theta)d\theta)}$$

in which probability is determined based on data and assumptions.P(y) is known as the posterior distribution from the definition of conditional probability. $f(y|\theta)$ is known as the likelihood, and $\pi(\theta)$ is the prior.$\int f(y|\theta)\pi(\theta)$ is equal to the probability of drawing your sample/data, or f(y). We let 'y' be our collected data. In the Rumen Project this would be represented by our microbiome and AMR counts. The posterior distribution allows us to find a distribution for for a fixed sample of data, 'y' Bayes theorem allows mathematicians to update their probability as more data becomes known. In instances in which not much data in known, the probability will be heavily dependent on assumptions.

Bayesian Networks:

Bayesian Networks is a graphical model that represents probabilistic dependencies between variables as a directed acyclic graph, where each node corresponds to a variable[5]. The probabilities used in a Bayesian Network is derived from the Markov Property, which states:

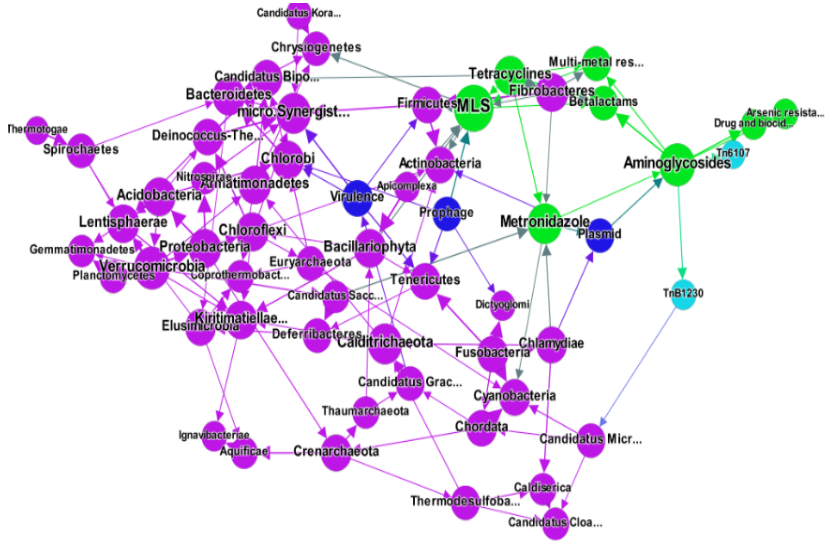
$$P_X(X) = \Pi_{i=1}^{P} P_{X_i}(X_i|\Pi_{X_i})$$

where X is the probability that an event will occur multiplied by all of its conditional probabilities.

Hill-Climbing Algorithm:

The Hill-Climbing algorithm is searching within a neighborhood of its initial guess and following the upward trend until it no longer has a greater maximum value within its neighborhood. The Hill-Climbing algorithm uses localized structure to identify which action or step will lead us towards a maximized solution. For the Bayesian Network, the Hill-Climbing algorithm is recognizing which connections or deletions of edges in a graph lead to a more statistically likely final network.

Results:



Significance:

Studying anti-microbial resistance is important for understanding how antibiotics affect the environment. The Bayesian Network clearly displays that connections between the microbiome and the antimicrobial resistance exists in sheep/cattle. With further study, we may find that AMR is playing a larger role in cattle/sheep health than we now understand. This introduction of anti-microbial resistance has a chain effect in the health of humans who consume these animals and their products.

Citations:

[1] Carlin, B., Louis, T. (2008). The Bayes Approach. In Bayesian Methods For Data Analysis (3rd ed., pp. 15–159). CRC Press.

[2] Doster, E., Jorgeson, B., Noyes, N., Luciano, C., Shi, B. (2020). NCBA Rumen Report. https://docs.google.com/document/d/1COrsFtIU7XOqkSmzNVM40-8vh2yfpb/editheading=h.gjdgxs

[3] Kolaczyk, E., Gabor, C. (2014). Statistical Analysis of Network Data with R. Springer.

[4] Li, H. (2015). Microbiome, Metagenomics, and High-Dimensional Compositional Data Analysis. Annual Review of Statistics and Its Application, 2(1), 73–94. https://doi.org/10.1146/annurev- statistics- 010814-020351

[5] Nagarajan, R., Scutari, M., Lebre, S. (2013). Bayesian Networks in R with Applications in Systems Biology. Springer.