Application of Bayesian Techniques to Multi-omic Longitudinal Data

Daniel Ruiz-Perez [1] **Jose Lugo-Martinez** [2] **Ziv Bar-Joseph** [2] **Giri Narasimhan** [1]

Abstract

We report updates on the development of a computational pipeline for the inference of heterogeneous interactions from longitudinal studies, improving on previous work on a pipeline for homogeneous interactions. Metagenomics, metatranscriptomics, and metabolomics data from iHMP IBD were used and modeled using Dynamic Bayesian Networks. The data were interpolated and temporally aligned to account for differential rates of change, missing data, and irregular sampling. A metabolic framework is imposed, and the inferred validations are being validated both computationally and experimentally.

1. Introduction and Motivation

Microbes living inside the human body interact with each other and their host through different metabolites, suggesting a complex "social network" (Ackerman, 2012). Microbiomes are dynamic in nature, which makes longitudinal data necessary to understand their behavior (Gerber, 2014). Moreover, the *internal clocks* of different individuals and factors such as age or gender influence the metabolic speed of many biological processes. Thus, to analyze time-series data across individuals, we need to compensate for it and also deal with non-uniform sampling, noisy and missing data. Finally, the reduction in the price of sequencing techniques is now making it possible to generate multi-omics data (e.g., metatranscriptomics, metabolomics, and metagenomics), thus allowing us to link many pieces of the puzzle. An in-depth knowledge of this network would facilitate the development of new and effective drugs and treatments. For this task, Dynamic Bayesian Networks (DBNs) will be used to infer temporal biological interactions. DBNs are probabilistic graphical models consisting of a directed acyclic graph where at each time slice the nodes correspond to random variables of interest and directed edges correspond to their conditional dependencies. They are ideally suited to model heterogeneous dynamic systems and infer temporal interactions between their constituents.

[1]Bioinformatic Research Group, Florida International University, Miami, FL, USA [2]Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA, USA. Correspondence to: Giri Narasimhan <giri@cs.fiu.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 108, 2020.

2. Methods

Figure 1 illustrates the computational pipeline developed in (Lugo-Martinez et al., 2019), where the example uses the time series of abundance values for the microbial taxon *Gammaproteobacteria* from five samples of an infant gut data set sampled at multiple time points (La Rosa et al., 2014). Figure 1 shows: **a)** Raw relative abundance values for each sample; **b)** Cubic B-spline curves for each time series following previous work (Bar-Joseph et al., 2012), which enable principled estimation of unobserved time points and interpolation at uniform intervals. **c)** Temporal alignment of each individual sample against a selected reference sample, which takes care of different metabolic rates and time lags; **d)** Removing samples with higher alignment error, since not every individual may follow the same process; **e)** Learning a DBN structure and parameters, and inferring biological relationships in the learned DBN; and **f)** comparing the original and predicted relative abundance across four different taxa. The authors of (Lugo-Martinez et al., 2019) modified the Matlab package CGBayesNet (McGeachie et al., 2014) to allow for intra-edges in the structure learning (connections within the same time point) and implemented balancing penalty functions such as Akaike Information Criterion and Bayesian Information Criterion (Penny, 2012). Also, dynamic restrictions encoded as an adjacency matrix establish the edge types allowed during structure learning were added. Within the same time slice, clinical variables can influence microbial taxon abundance, which in turn can influence the expression of a gene in its genome, which in turn can be involved in metabolic pathways to impact the metabolites produced, which in turn can be consumed by other taxa in the next time slice, impacting their abundance.

In the project reported here, the tools from (Lugo-Martinez et al., 2019) were modified for multi-omic data and are undergoing the process of evaluation and testing. For this project, we are using the data generated by the Integrative Human Microbiome Project, which followed 132 individuals with Inflammatory Bowel Disease over a period of one

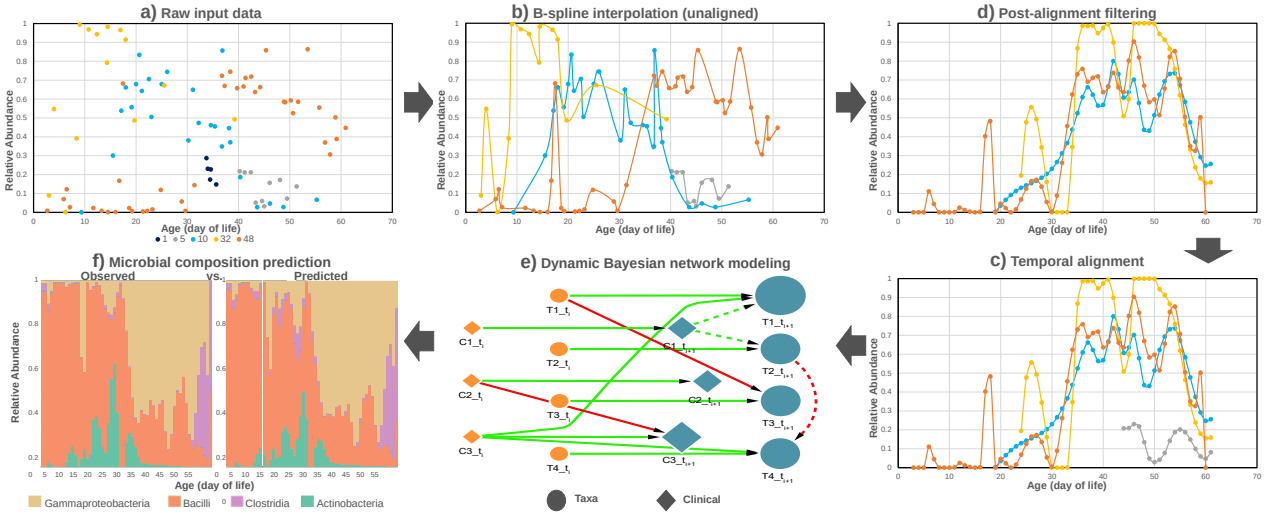

Figure 1. Computational pipeline developed.

year (Lloyd-Price et al., 2019). Cross validation prediction error will be reported in addition to the error of a continuous prediction to assess how fast the predictions deviate from the ground truth. The inferred interactions are being validated computationally, experimentally, and using literature-based methods. In-house scripts will query known databases and assess the biological relevance of a connection from a taxon to a metabolite or a gene to a metabolite. Then the interactions with the highest bootstrap confidence and overall importance will be selected and tested experimentally.

3. Results and Discussion

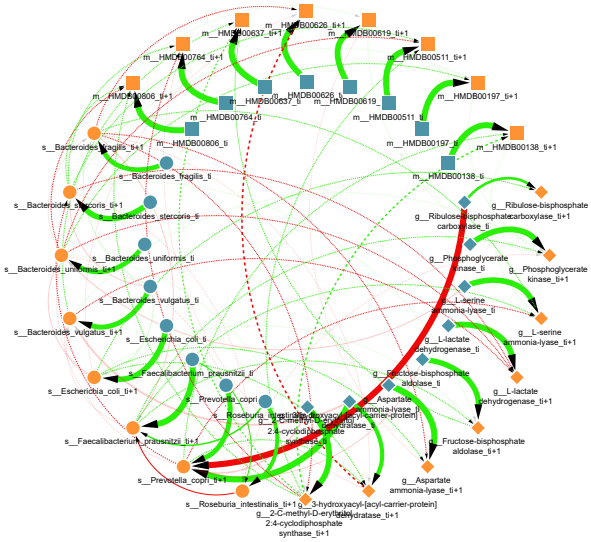

Figure 2. Heterogeneous DBN for iHMP IBD data set

The intra-edges and penalty modifications of CGBayesNet were previously tested in (Lugo-Martinez et al., 2019) for interactions among taxa and compared against the state of the art methods, showing an increase in prediction accuracy among all data sets tested. Similar validations are underway for this project.

Figure 2 shows a bootstrapped DBN incorporating a subset of metabolites, genes and taxa and the interactions between them. The shape of a node represents its type, while edge color represents the sign of the regression coefficient between the nodes (green is positive and red is negative). Preliminary validation using a larger subset of nodes was executed using the tool MIMOSA (Noecker et al., 2016) to calculate the metabolic potential of each taxon, and KEGG genome libraries to validate taxa-metabolite and taxa-gene interactions respectively. A significant (based on a Poisson-Binomial distribution) number of the edges predicted by the DBN were present in the validation tools and databases used.

The work reported in this abstract represents novel and valuable research on an integrated analysis of multiomic longitudinal data.

References

Ackerman, J. The ultimate social network. *Scientific American*, 306(6):36–43, 2012.

Bar-Joseph, Z., Gitter, A., and Simon, I. Studying and modelling dynamic biological processes using time-series gene expression data. *Nat Rev Genet*, 13:552–564, 2012.

Gerber, G. K. The dynamic microbiome. *FEBS Lett*, 588 (22):4131–4139, 2014.

La Rosa, P. S., Warner, B. B., Zhou, Y., Weinstock, G. M., Sodergren, E., Hall-Moore, C. M., Stevens, H. J., Bennett, W. E., Shaikh, N., Linneman, L. A., Hoffmann, J. A., Hamvas, A., Deych, E., Shands, B. A., Shannon, W. D., and Tarr, P. I. Patterned progression of bacterial populations in the premature infant gut. *Proc Natl Acad Sci*, 111 (34):12522–12527, 2014.

Lloyd-Price, J., Arze, C., Ananthakrishnan, A. N., Schirmer, M., Avila-Pacheco, J., Poon, T. W., Andrews, E., Ajami, N. J., Bonham, K. S., Brislawn, C. J., et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*, 569(7758):655, 2019.

Lugo-Martinez, J., Ruiz-Perez, D., Narasimhan, G., and Bar-Joseph, Z. Dynamic interaction network inference from longitudinal microbiome data. *Microbiome*, 7(1): 54, 2019.

McGeachie, M. J., Chang, H.-H., and Weiss, S. T. CGBayesNets: Conditional gaussian bayesian network learning and inference with mixed discrete and continuous data. *PLoS Comput Biol*, 10(6):1–7, 2014.

Noecker, C., Eng, A., Srinivasan, S., Theriot, C. M., Young, V. B., Jansson, J. K., Fredricks, D. N., and Borenstein, E. Metabolic model-based integration of microbiome taxonomic and metabolomic profiles elucidates mechanistic links between ecological and metabolic variation. *MSystems*, 1(1):e00013–15, 2016.

Penny, W. D. Comparing dynamic causal models using AIC, BIC and free energy. *NeuroImage*, 59(1):319–330, 2012.