11-12-2020

Temporal and Causal Inference with Longitudinal Multi-omics Microbiome Data

Daniel Ruiz-Perez
Florida International University, druiz072@fiu.edu

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

TEMPORAL AND CAUSAL INFERENCE WITH LONGITUDINAL MULTI-OMICS MICROBIOME DATA

A dissertation submitted in partial fulfillment of the

requirements for the degree of

DOCTOR OF PHILOSOPHY

in

COMPUTER SCIENCE

by

Daniel Ruiz-Perez

2020

To: Dean John L. Volakis
College of Engineering and Computing

This dissertation, written by Daniel Ruiz-Perez, and entitled Temporal and Causal Inference with Longitudinal Multi-omics Microbiome Data, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this dissertation and recommend that it be approved.

Ziv Bar-Joseph

Kalai Mathee

Fahad Saeed

Ananda Mondal

Giri Narasimhan, Major Professor

Date of Defense: November 12, 2020

The dissertation of Daniel Ruiz-Perez is approved.

Dean John L. Volakis
College of Engineering and Computing

Andrés G. Gil
Vice President for Research and Economic Development and Dean of the University Graduate School

DEDICATION

To my parents.

ACKNOWLEDGMENTS

I want to thank my advisor and committee members, everybody in the BioRG group, my family, my girlfriend, and my friends for all your support.

ABSTRACT OF THE DISSERTATION

TEMPORAL AND CAUSAL INFERENCE WITH LONGITUDINAL MULTI-OMICS MICROBIOME DATA

by

Daniel Ruiz-Perez

Florida International University, 2020

Miami, Florida

Professor Giri Narasimhan, Major Professor

Microbiomes are communities of microbes inhabiting an environmental niche. Thanks to next generation sequencing technologies, it is now possible to study microbial communities, their impact on the host environment, and their role in specific diseases and health. Technology has also triggered the increased generation of multi-omics microbiome data, including metatranscriptomics (quantitative survey of the complete metatranscriptome of the microbial community), metabolomics (quantitative profile of the entire set of metabolites present in the microbiome's environmental niche), and host transcriptomics (gene expression profile of the host). Consequently, another major challenge in microbiome data analysis is the integration of multi-omics data sets and the construction of unified models. Finally, since microbiomes are inherently dynamic, to fully understand the complex interactions that take place within these communities, longitudinal studies are critical. Although the analysis of longitudinal microbiome data has been attempted, these approaches do not attempt to probe interactions between taxa, do not offer holistic analyses, and do not investigate causal relationships.

In this work we propose approaches to address all of the above challenges. We propose novel analysis pipelines to analyze multi-omic longitudinal microbiome data, and to infer temporal and causal relationships between the different entities involved. As a first step, we showed how to deal with longitudinal metagenomic data sets by building a pipeline, PRIMAL, which takes microbial abundance data as input and outputs a dynamic Bayesian

network model that is highly predictive, suggests significant interactions between the different microbes, and proposes important connections from clinical variables. A significant innovation of our work is its ability to deal with differential rates of the internal biological processes in different individuals. Second, we showed how to analyze longitudinal multi-omic microbiome datasets. Our pipeline, PALM, significantly extends the previous state of the art by allowing for the integration of longitudinal metatranscriptomics, host transcriptomics, and metabolomics data in additional to longitudinal metagenomics data. The predictive capability of PALM is on par with that of the PRIMAL pipeline while discovering a web of interactions between the entities of far greater complexity. An important innovation of PALM is the use of a multi-omic Skeleton framework that incorporates prior knowledge in the learning of the models. Another major innovation of this work is devising a suite of validation methods, both in silico and in vitro, enhancing the utility and validity of PALM. Finally, we propose a suite of novel methods (unrolling and de-confounding), called METALICA consisting of tools and techniques that make it possible to uncover significant details about the nature of microbial interactions. We also show methods to validate such interactions using ground truth databases. The proposed methods were tested using an IBD multi-omics dataset.

TABLE OF CONTENTS

LIST OF TABLES

TABLE PAGE

LIST OF FIGURES

FIGURE PAGE

LIST OF FREQUENTLY USED MATHEMATICAL SYMBOLS

X, X_i, Y, Z	Random variables
x, y, z	Values of random variables
$P(\cdot), F(\cdot)$	Probability, probability density function
$(\cdot \mid \cdot)$	Conditional probability
$\cdot \perp\!\!\!\perp \cdot$	Conditional independence
$E(\cdot), \hat{\cdot}$	Expected value, and estimated value
$\mu, \sigma^2_{\cdot}, \sigma(\cdot\cdot), \delta$	Mean, variance, covariance, and standard deviation
ρ	Partial correlation
$\mathcal{N}(\mu, \sigma^2)$	Normal distribution
$D(\cdot \mid \cdot)$	Set of conditional probability distributions
$C(\cdot \mid \cdot)$	Set of linear Gaussian conditional densities
n, m	Number of variables, and number of samples
$Pa^G(\cdot)$	Set of parents of a node in G
$Pa^G_{p_x}(\cdot)$	Set of p_x strongest parents of a node in G
$MB(\cdot)$	Markov blanket of node
Δ, Γ	Set of discrete and continuous nodes
$\phi(\cdot), \epsilon$	Penalty term, and error term
N'	Equivalent sample size
$N(\cdot)$	Number of cases that satisfy condition
Θ	Set of parameters of the model
Θ_i	$P(X = X_i)$
$\Theta_{i,j,k}$	$P(X_i = j \mid Pa^G(X_i) = k)$
$\Theta_{i.k}$	$\{\Theta_{i,j,k} \mid j = 1, \ldots, r_i\}$
$n_i, \lvert\Omega_{X_i}\rvert$	Number of states of node X_i
$s^j_r(t)$	Spline curve for taxa j in reference r at time t
$\tau(\cdot), \tau$	Alignment function, and time delay

r, r^*	Reference function, best reference function
a, b	Alignment stretch and shift parameters
α, β	Alignment starting and ending time points
G, V, E	Graph, set of vertices, and set of edges
G_X, V_X, E_X	G, V, and E of the network learned using dataset X

CHAPTER 1

INTRODUCTION

1.1 Motivation and Goals

Microbiomes are communities of microbes inhabiting an environmental niche. The study of microbial communities offers a powerful approach for inferring their impact on the host environment, and their role in specific diseases and health. *Metagenomics* involves analyzing sequenced reads from the whole metagenome in a microbial community in order to determine a detailed abundance profile of microbial taxa [139]. More recently, additional types of biological data are being generated in microbiome studies, including *metatranscriptomics*, which involves surveying the expression of the genes in the complete metatranscriptome of the microbial community [12], *metabolomics*, which involves profiling the concentrations of the entire set of small molecules (metabolites) present in the microbiome's environmental niche [182], and *host transcriptomics*, which provides information about the expression levels of host genes [23].

A major challenge in microbiome data analysis is the integration of multi-omics data sets [124]. Most multi-omic studies focus on a separate analysis of each omics data set without building a unified model [14]. There have been some attempts [202, 99, 201, 203, 92] and tools to facilitate the analysis [16, 150], but there is still much room for improvement regarding reproducibility, flexibility, and biological validity [124, 22, 184]

Deep learning approaches for integrating multi-omics [98, 111] have also been developed, but they are either not interpretable, or limited to predicting just one of the omics. This, together with their computational cost prevents these models from being useful at providing insights into the interplay of the different omics entities. Even Partial Least Squares models have been used to facilitate this integration [49], but they have their own set of limitations depending on the underlying data generation model, and are prone to provide spurious results when applied to high-dimensional data [145].

In addition, microbiomes are inherently dynamic, and so to fully understand the complex interactions that take place within these communities, longitudinal microbiome data appears to be critical [57]. Many attempts have been made to analyze data from longitudinal studies [86, 92, 203]; however, these approaches do not attempt to study interactions between taxa. An alternative approach involves the use of dynamical systems such as the generalized Lotka-Volterra (gLV) models [172, 58], however the large set of parameters in these models diminishes their utility for probabilistic inference.

Finally, it is unclear if the interactions inferred by some of these methods represent a true and direct causal interaction, and not merely the result of a statistical correlation resulting from some indirect causal relationship or model overfitting. Microbiomes are complex environments with many subtle relationships, while the inference of community profiles relies on noisy data from error-prone technologies, and has to contend with a host of hidden confounders that may be hard or impossible to measure, let alone be identified. The jump to infer causality is a natural next step in inferring multi-omic interactions, and the lack of research in this area is striking. Most of the causal microbiome literature focuses on the causal impact of the microbiome to health or disease, but not on the causal interactions between these microorganisms [69, 97, 151, 137].

1.2 Research Contributions

The contributions of this dissertation include a computational pipeline that enables the integration of data across individuals for the reconstruction of dynamic models from time series microbiome data. This pipeline was then extended to allow for the integration of metagenomics, metatranscriptomics, host transcriptomics, and metabolomics longitudinal data under a unified framework. Where possible, the interactions predicted by it were validated both *in silico* and *in vitro*. Finally, a suite of methods for the analysis and augmentation of causal networks was developed with the goal of addressing some of the most fundamental issues in causal inference.

- **Longitudinal microbial network inference** (PRIMAL). We developed a computational pipeline that enables the integration of data across individuals for the reconstruction of dynamic models from time series microbiome data. Our pipeline starts by aligning the data collected for all individuals. The aligned profiles are then used to learn a dynamic Bayesian network which represents causal relationships between taxa and clinical variables. Testing our methods on three longitudinal microbiome data sets we show that our pipeline improves upon prior methods developed for this task. We also discuss the biological insights provided by the models which include several known and novel interactions.

- **Longitudinal multi-omic network inference** (PALM). A key challenge in the analysis of longitudinal microbiome data is the inference of causal interactions between microbial taxa, their genes, the metabolites they consume and produce, and host genes. To address these challenges we developed a computational pipeline PALM that first aligns multi-omics data and then uses dynamic Bayesian networks (DBNs) to reconstruct a unified model. Our approach overcomes differences in sampling and progression rates, utilizes a biologically-inspired multi-omic framework, reduces the large number of entities and parameters in the DBNs, and validates the learned network. Applying PALM to data collected from inflammatory bowel disease (IBD) patients, we show that it accurately identifies known and novel interactions. Targeted experimental validations further support a number of the predicted novel metabolite-taxa interactions.

- **Longitudinal causal multi-omic network inference** (METALICA) In an effort to improve the state of the art in inferring meaningful multi-omic interactions, we addressed some of the most fundamental issues in causal inference. We developed a suite of tools and techniques that discover strong interactions by uncovering hidden multi-omic confounders by the method we called *de-confounding*, find the actual reason two taxa interact with each other and how they do it by a process we called

unrolling, and finally automatically validate the feasibility of such interactions using ground truth databases. We applied our methods to networks learned by causal algorithms such as Tigramite and TETRAD, which we augmented with our restriction framework and alignment techniques, among other improvements. The dataset used was an IBD multi-omic dataset, and the findings were used to compare the inferences of the various methods against the ones of PALM.

1.3 Road Map for the Dissertation

After setting up the stage for the drive in this introductory chapter, the rest of the journey is organized as follows.

In Chapter 2, we will introduce the reader to all notations, definitions, and necessary terminologies. In the process, we will also define all basic concepts, assumptions, and general methods that are going to be used throughout this dissertation.

Chapter 3 provides the mathematical formulation necessary to understand Bayesian networks, causality algorithms, and key bioinformatics techniques used. It also contains a biological literature survey on microbial interaction inference and a description of the different omics used in this dissertation.

Chapter 4 contains a detailed explanation of dynamic Bayesian networks (DBNs), from the model construction to the parameter inference. In addition,we it contains a summary of the contributions performed in this dissertation to the area of DBN learning.

Chapter 5 contains the work carried out on the area of longitudinal microbial network inference. It contains the necessary specific background, and a comprehensive description of the datasets and methods used and developed. Finally, the results of the application of the pipeline developed are presented and discussed in detail.

Chapter 6 contains a description of PALM, the pipeline developed to integrate multi-omic longitudinal data and carry out microbial interactions inference. We set the stage with the appropriate background, a description of the dataset used, and the computational

contributions. We then present the results of the execution of such pipeline, and validate its findings both in-silico and in-vitro. We finally discuss the findings in detail.

Chapter 7 describes the work on the inference of causal interactions from multi-omic data sets. We introduce some important problems in causal inference and present our proposed methods to help address them in the multi-omic context. We then test our methods on the networks learned by three state-of-the-art methods, and use them to perform a comparison among the tools.

We close the dissertation in Chapter 8 with a summary of the dissertation, together with conclusions and suggestions for future work.

CHAPTER 2

NOTATION AND TERMINOLOGY

2.1 Variables

A **random variable** (or just variable) is a numerical description of the outcome of a statistical experiment. It may assume multiple values from its domain. Random variables are denoted by uppercase letters and their values are denoted by lowercase letter. For example, when random variables X, Y, Z are instantiated, their values are denoted by corresponding lower case letters such as x, y, z respectively. For our purposes, these variable can be discrete- or continuous-valued. A discrete variable can only take on values from a finite or infinite set (for example, a random variable representing disease status can take on one of two values {*healthy, diseased*}). Continuous random variables may assume any value in some interval on the real number line (for example, *relative abundance of a particular microbe*). **Response (dependent) variables** are related to the outcome of a study or experiment, and are a function of the **explanatory (independent) variables** of the system. **Confounding** is the situation where the effects of two or more explanatory variables are not separated. Because of this, any relation between an explanatory variable and the response variables (or other explanatory variables) may be due to some other variable not accounted for in the study. A **confounding variable** (or confounder) is any variable that causes spurious associations. It is an extraneous variable that was not appropriately controlled for [70], that is associated with both the response and explanatory variables being contemplated [181]. A **hidden (lurking, or latent) variable** is an explanatory variable that was not considered in the study, but that affects a response variable [175, 159]. When a hidden variable also acts as a confounder, it is called a **hidden confounder**.

2.2 Probability

Probability is the measure of how likely it is for a random variable to take a specific value (**event**). The event $X = 0.3$ means that the random variable X took the value 0.3. The probability of an event is always between 0 and 1, and the probability measure is denoted by P. The **conditional probability** is the specific probability of an event given that some other event has occurred. Conditional probability of an event $X = x$ given another event $Y = y$ is denoted as $P(X = x \mid Y = y)$.

To understand this dissertation it is important to understand the Bayesian approach to probability. The **Bayesian probability** of an event $X = 3$ is a person's degree of belief in that event. While the classical probability is a physical property of the world (e.g., the probability of getting a three on a die throw), the Bayesian probability is a property of the person who assigns that probability (e.g., your degree of belief that you will get a three when you throw the die) [64].

2.3 Basic Statistical Concepts and Notation

- The **expected value** of a variable X, denoted by $E(X)$, is intuitively the *mean* of a large number of independent realizations of X. It is calculated by

$$E(X) = \sum_{x} P(X = x). \tag{2.1}$$

- The **variance** of a variable X, denoted by $Var(X)$ or σ_X^2, measures how the values of X are spread out from their mean. It's calculated by

$$\sigma_X^2 = \frac{\sum x_i - \mu}{n - 1} = E[(X - \mu)^2]. \tag{2.2}$$

where μ is the mean of X.

- The **standard deviation** σ_X of random variable X is the square root of its variance.
- The **covariance** of X and Y, denotes as σ_{XY}, is a measure of the degree to which two variables X and Y vary together. Formally,

$$\sigma_{XY} = E[(X - E(X))(Y - E(Y))], \tag{2.3}$$

- The **correlation** or dependence between two variables is any statistical association between them. It usually refers to the degree to which two variables are linearly related. There are many ways of calculating correlation, but the simplest is probably to directly use the covariance between the variables the following way:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}. \tag{2.4}$$

  However, there are many different measures of correlation, including the Pearson correlation coefficient [127], Spearman correlation coefficient [169], Local Similarity Analysis [144], SparCC [53], Maximal Information Coefficient [138], and Bray–Curtis distance [20].

2.4 Probability Distributions

The **probability distribution** of a discrete random variable X is the set of probabilities of all the different outcomes of X. If the outcomes of X are 1, 2, and 3 then one possible probability distribution of X is $P(X = 1) = 0.2$, $P(X = 2) = 0.45$ and $P(X = 3) = 0.35$. The sum of all probabilities from a probability distribution has to be 1.

The probability distribution of a continuous variable X is expressed as a **probability density function** (PDF), denoted by f_X.

Because the probability of a continuous variable X taking a specific value is 0, the PDF is used to specify the probability of the value of the random variable falling within a particular range. This probability is given by the integral of f_X over that specific range. The area under the entire curve represents the probability of X taking *any* value, and is therefore defined as

$$P[-\infty \leq X \leq \infty] = \int_{-\infty}^{\infty} f_X(x)\, dx = 1. \tag{2.5}$$

The probability distribution most commonly used in this dissertation is the **Gaussian** or **Normal distribution**, denoted as $X \sim \mathcal{N}(\mu, \sigma^2)$ for the variable X. Its calculated as:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \tag{2.6}$$

where μ is the mean, σ is the standard deviation, and σ^2 is the variance of the distribution.

When we are interested in modelling the probability distribution, or PDF of more than one random variable, we use the notion of a **joint probability distribution**, or **joint PDF**, which gives the probability that each of the random variables take some value in a particular range.

2.5 Bayes' Theorem

Arguably the most important theorem in this dissertation, **Bayes' theorem** describes the probability of an event conditioned on relevant prior knowledge for that event. Formally, it is written as follows.

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}. \tag{2.7}$$

Note that $P(A \mid B)$ is called the **posterior**, $P(B \mid A)$ is referred to as the **likelihood**, $P(A)$ is the **prior**, and $P(B)$ is the **marginal likelihood**.

2.6 Conditional Independence

Two variables X and Y are **independent** (denoted by $X \perp\!\!\!\perp Y$) if the probability of one variable is the same as its probability when conditioned on the other. Formally,

$$X \perp\!\!\!\perp Y \iff P(X \mid Y) = P(X). \tag{2.8}$$

Note that this is a symmetric relation, meaning that $X \perp\!\!\!\perp Y \iff Y \perp\!\!\!\perp X$. Two variables X and Y are **conditionally independent** given another variable Z, if fixing Y does not add any new information about X when Z is instantiated. Formally,

$$X \perp\!\!\!\perp Y \mid Z \iff P(X \mid Y, Z) = P(X \mid Z) \iff P(Y \mid X, Z) = P(Y \mid Z). \tag{2.9}$$

2.6.1 Conditional Independence Tests

Two of the most common conditional independence tests are the **exact t** test and the **Fisher's Z** test, which are based on correlation. There are multitude of other tests based on other concepts, and their applicability highly depends on the underlying data distribution.

If X and Y are two continuous random variables and $\mathbf{Z}$ is a set of continuous variables, then conditional independence tests between X, Y given $\mathbf{Z}$ can be written using the partial correlation coefficient $\rho_{XY|\mathbf{Z}}$ [159].

The exact t test for Pearson's correlation, which is distributed as a Student's t with $n - |\mathbf{Z}| - 2$ degrees of freedom [159] is defined by:

$$\mathrm{t}(X, Y \mid \mathbf{Z}) = \rho_{XY|\mathbf{Z}} \sqrt{\frac{n-2}{1-\rho^2_{XY|\mathbf{Z}}}}, \tag{2.10}$$

On the other hand, Fisher's Z test [159] is calculated using the following formula:

$$FisherZ(X, Y \mid \mathbf{Z}) = \frac{\sqrt{n - |\mathbf{Z}| - 3}}{2} \log \frac{1 + \rho_{XY|\mathbf{Z}}}{1 - \rho_{XY|\mathbf{Z}}}. \tag{2.11}$$

2.7 Regression

Regression is a statistical process to estimate the functional relationship between a dependent variable Y (outcome, or response variable) and one or more independent variables $X_1, X_2, \ldots, X_n$ (predictors, covariates, or features). In this dissertation we mainly focus on the concept of **linear regression**, where the goal is to find a linear function that describes the output Y in terms of the values of the predictors such that the sum of squared errors is minimized. When there are more than one predictor variables, it is referred to as **multiple linear regression**. A regression model is expressed as follows.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_n + \epsilon, \tag{2.12}$$

where $\beta_i \in \mathbb{R}$ is the regression coefficient of X_i, $i = 1, 2, \ldots, n$, and ϵ is the error term.

The best known way of estimating the regression coefficients is by a process known as the method of **least-squares**. It makes the following assumptions:

- **Weak exogeneity**: Assumes that the predictor variables are error-free, and can be treated as fixed values, rather than random variables.
- **Linearity**: Assumes the mean of the outcome variable is a linear combination of the regression coefficients and the predictors.
- **Constant variance (homoscedasticity)**: Assumes that the variance of the errors of the response variable is the same for all of its values. In practice, this assumption doesn't usually hold.
- **Independence of errors**: Assumes the errors of the outcome are uncorrelated with each other.

- **Lack of perfect multicollinearity in the predictors**: Assumes that there is no single covariate that can be linearly predicted from the others with a substantial degree of accuracy.

The method of least-squares calculates the regression coefficients by minimizing the sum of the squares of the residuals for each point. The residual is the difference between an observed value and the fitted value provided by the model.

The least-squares solutions of $Ax = b$ are the solutions of $A^TAx = A^Tb$, and the regression coefficients can be found by solving $\hat{x} = (A^TA)^{-1}A^Tb$, where $\hat{x}$ are the estimated regression coefficients, and A is the $m \times n$ matrix whose columns correspond to the n predictor variables and the m rows correspond to the samples or realizations. Finally, the column vector b contains the values of the outcome variable for the m realizations.

2.8 Time Series Correlations

The **Crosscorrelation** (sliding dot product) is the correlation between different time series. It is a measure of similarity of two series as a function of the displacement of one relative to the other. The **Autocorrelation** (serial correlation) is the correlation between a signal and itself at different time points (lags). It is thus the crosscorrelation of a time series with itself. The **Partial correlation** is the correlation between the same or different time series with the effect of lower order correlations removed. It controls for other random variables, by removing the effect of confounding variables, and it is calculated as follows:

$$\rho_{XY \cdot Z} = \frac{\rho_{XY} - \rho_{XZ}\rho_{ZY}}{\sqrt{1 - \rho_{XZ}^2}\sqrt{1 - \rho_{ZY}^2}}. \tag{2.13}$$

2.9 Graphical Models

An **undirected graph** is an ordered pair $G = (V, E)$, where $V = \{v_1, v_2, \ldots, v_n\}$ is the collection of elements called vertices or nodes, and $E \subseteq \{(v_i, v_j) \mid (v_i, v_j) \in V^2 \text{ and } v_i \neq v_j)\}$

is a set of unordered pairs of vertices, called edges (or links, or arcs) connecting the vertices. Two nodes are called *adjacent* in a graph G if there is an edge between. A *path* p from X to Y in G is a sequence of distinct nodes $(X = Z_1, \ldots, Z_m = Y)$ such that Z_i and Z_{i+1} are adjacent in G for all $i \in \{1, \ldots, m-1\}$. X and Y are called the *endpoints* of the path p. If there is a path from X to Y, then X and Y are said to be *connected*. A path where the two endpoints are the same vertex is called a *cycle*.

When the edges correspond to ordered pairs of vertices, the resulting graph is a **directed graph**. A path is called a *directed path* from Z_1 to Z_m if $Z_i \rightarrow Z_{i+1}$ for all $i \in \{1, \ldots, m-1\}$ [129]. If there is a directed path from X to Y, then Y is said to be *reachable* from X. If there is a directed edge e_z from X_i to X_j, we call X_i a parent of X_j, and X_j the child of X_i. Furthermore, we define $Pa^G(X_i) \subseteq V$ as the set of parents of X_i in G. The number of incoming edges to a vertex is called the indegree (number of parents), and the number of edges leaving the vertex is its outdegree (number of children).

A **directed acyclic graph** (DAG) is a directed graph with no *directed cycles*, i.e., no collection of edges $(e_1, e_2, \ldots, e_n)$ in the DAG with a vertex sequence $(v_1, v_2, \ldots, v_n, v_1)$.

A related notion is that of a *probabilistic graphical model* (PGM), which is a DAG that provides a convenient graphical representation of the structure of the joint probability distribution it represents. Nodes represent random variables, while directed edges capture conditional dependence relationships between the random variables. There are two main types of PGM; **Directed Graphical Models**, otherwise known as Bayesian Networks (see Section 3.1), and **Undirected Graphical Models** or Markov Random Fields [90].

A path p from nodes A to B is **d-connected** with respect to a set of nodes $\boldsymbol{S}$ if and only if the following two conditions are satisfied:

1. For every **chain** $X \rightarrow Z \rightarrow Y$ or **fork** $X \leftarrow Z \rightarrow Y$ on path p, the middle node $Z \notin \boldsymbol{S}$.
2. For every **collider** $X \rightarrow Z \leftarrow Y$ on path p, either the middle node Z or one of its descendants is in set $\boldsymbol{S}$.

A and B are **d-separated** with respect to $\boldsymbol{S}$ if they are not d-connected with respect to $\boldsymbol{S}$. This is a way of linking conditional independence with graphical representation [125]. If $X \rightarrow Z$ and $Y \rightarrow Z$ are edges in graph G, and X and Y are not adjacent, then the triple (X, Z, Y) is called a *V-structure*. The node Z is referred to as a *collider* in the V-structure, and it is also a collider on any path p that uses the edges of the V-structure. A path p with no collider is called an *unblocked path*.

Next we define the concept of the **Y-structure**. Let $X \rightarrow Z \leftarrow Y$ be a V-structure. If there is a node W such that there is an edge from Z to W, and no edges from X or Y to W, then the nodes (X, Y, Z, W) form a Y-structure.

CHAPTER 3

BACKGROUND AND REVIEW

3.1 Bayesian Networks

A **Bayesian Network** (BN) [126, 35] is a PGM that represents the joint distribution of a set of interdependent random variables. BNs allow to efficiently represent the joint probability distribution of all the variables as the products of conditional probability distributions [156]. Formally, a BN is represented by a DAG, $G = (V, E)$, where V represents the the set of n random variables $\{X_1, X_2, \ldots, X_n\}$, and E represents the dependencies between those variables. In addition to its graphical structure, the BN also contains a collection of conditional probability tables for each node, since each random variable X_i has an associated probability distribution given its parents. The modifications that allow a BN to model longitudinal data and make temporal inferences are called Dynamic Bayesian Networks (DBNs), which are explained in detail in Chapter 4.

Under the first-order **Markov assumption**, X_i only depends on $Pa^G(X_i)$, and is either marginally or conditionally independent of all other variables. Thanks to this Markov property, we can decompose the global (joint) probability distribution $P(\mathbf{X})$ as the product of all local conditional probabilities.

We say that a joint probability distribution factorizes with respect to the DAG G if:

$$P(X_1, X_2, \ldots, X_n) = P(X_1 \mid Pa^G(X_1)) \cdot P(X_2 \mid Pa^G(X_2)) \cdot, \ldots, \cdot P(X_n \mid Pa^G(X_n)), \tag{3.1}$$

which holds only under the Markov assumption. Then, we can compactly express the global probability distribution of all variables of the BN as:

$$P(\mathbf{X}) = \prod_{i=1}^{n} P(X_i \mid Pa^G(X_i)). \tag{3.2}$$

The explicit representation of G in Eq. 3.2 provides many advantages in terms of manageability and computation cost. Take for example a simple case where each of the n variables are binary; a joint distribution would require the specification of a table of $2^n - 1$ entries, which are the probabilities of the 2^n different combinations of assignments of $X_1, X_2, \ldots, X_n$. Thus the mere specification of the joint distribution becomes intractable as n grows. The compact representation of the joint distribution as the product of the conditional probabilities, conditioned on the parents, causes great savings in terms of computation and memory.

The last concept that is important for the understanding of BNs is that of the **Markov blanket** of a node $MB(X_i)$; the set containing the node's parents, children, and other parents of its children. An important property is that any node is conditionally independent of every other node in G given its Markov blanket. Formally, $X_i \perp\!\!\!\perp \mathbf{X} \setminus MB(X_i) \mid MB(X_i)$.

The BN formalism that we will be using through out this dissertation is that of **conditional Gaussian Bayesian network** (CGBN), or mixed BN, where the nodes may represent a mix of continuous and discrete random variables [65, 36, 105]. The main assumption is that nodes representing continuous random variables follow Gaussian distributions that can be expressed as a linear combination of the parent nodes with continuous distributions and with parameters conditioned upon the values of any discrete parents. In this formalism, discrete nodes can have continuous parents, but the discrete nodes cannot have continuous parents, although the joint distribution can still be captured by the network. Formally, in a CGBN, the set of nodes $\mathbf{V}$ is partitioned into a set of discrete nodes $\mathbf{\Delta}$, and a set of continuous nodes, $\mathbf{\Gamma}$. Associated with each $\Gamma_i \in \mathbf{\Gamma}$ of the continuous nodes are conditional Gaussian (CG) regressions, one for each configuration in the state space of its discrete parents $Pa^G(\Gamma_i)$.

3.2 Bayesian Networks Learning

Learning a BN has two parts; **structure learning**, which infers all the statistical connections between the nodes, and **parameters learning,** which assigns a strength or statistical measure to those connections. Cooper [33] showed that learning a BN is an NP-hard problem, therefor efforts should be placed in learning an approximate solution instead of the optimal BN for a given problem. Moreover, Dagum and Luby [40] showed that approximating probabilistic inference is also NP-hard, so we should find ways to efficiently restrict the search space to infer a BN in reasonable time.

3.2.1 Constraint-based Structure-learning Methods

Constraint-based algorithms stem from the **inductive causation** (*IC*) algorithm [188] and the *SGS* algorithm [170] depicted in Algorithm 1. These procedures assume that graphical separation and probabilistic independence imply each other (faithfulness assumption, Section 3.4.1). Then they apply a series of conditional independence tests to learn the structure of the network and a series of rules to learn the orientation of the edges (the causal direction).

Algorithm 1: SGS algorithm [170]

Input: V, vertex set; Conditional independence information or conditional independence test
Output: A directed or partially directed acyclic graph
Form the complete undirected graph H on the vertex set V.
For each pair of vertices A and B, if there exists a subset $\mathbf{S}$ of $V \setminus \{A, B\}$ such that A and B are d-separated given $\mathbf{S}$, remove the edge between A and B from H.
Let K be the undirected graph resulting from Step 2. For each triple of vertices (A, B, C) such that the pair (A, B) and the pair (B, C) are each adjacent in K, but the pair A and C are not adjacent in K, orient $A - B - C$ as $A \rightarrow B \leftarrow C$ if and only if there is no subset $\mathbf{S}$ of $\{B\} \cup V \setminus \{A, C\}$ that d-separates A and C.
while *more edges can be oriented* **do**
    If $A \rightarrow B$, B and C are adjacent, A and C are not adjacent, and there is no arrowhead at B, then orient $B - C$ as $B \rightarrow C$. If there is a directed path from A to B, and an edge between A and B, then orient $A - B$ as $A \rightarrow B$.
end

3.2.2 Hybrid Structure-learning Methods

These algorithms combine constraint-based and score-based algorithms (described below in Section 3.2.3) with the hope of learning better networks more efficiently. Two of the most famous hybrid learning algorithms are *Sparse Candidate* [54] and *Max-Min Hill-Climbing* [180]. Both these algorithms have mainly two steps: *restrict* and *maximize*. In the restrict step, the candidate set for the parents of each node is reduced from the whole node set to a small subset. The maximize phase, on the other hand, optimizes the score among the available candidate networks to find the optimal DAG in the space restricted in the maximize step [158].

3.2.3 Score-based Structure-learning Methods

The score-based approach considers structure learning as a search problem; it uses a *score* to evaluate how well the network fits the data, and then searches over the space of DAGs for a structure that optimizes the score. The score for a BN structure G with data D can be broadly described as:

$$Score(G : D) = LL(G : D) - \phi(|D|)||G||, \tag{3.3}$$

where $LL(G : D)$ is the log-likelihood of the data given the network, and is defined as $\log P(D \mid \Theta, G)$, Θ denotes the set parameters of the model, and $\phi(\cdot)$ represents the penalty term. Given that every edge added will improve the log-likelihood by at least a small quantity, we would end up with a fully connected network without the penalty term. The penalty term increases with the number of parameters (edges) in G. The two most commonly used scoring functions are the Bayesian Information Criteria (BIC) [155], where the penalty term used is $\phi(t) = 1$ and the Akaike Information Criteria (AIC) [2], where $\phi(t) = \log(t)/2$.

Bayesian Dirichlet (BD) represents yet another family of score functions [34]. The probability of the data given the network structure is given by:

$$P(D \mid G) = \int P(D \mid G, \Theta) P(\Theta \mid G)\, d\Theta, \tag{3.4}$$

where Θ is the set of parameters of the model. Note that as in Bayesian approach, there is considerably flexibility in assuming the prior probability, $P(\Theta \mid G)$. When we assume that the prior probability follows a Dirichlet distribution, we have:

$$P(D \mid \Theta) = \prod_{X_i}^{N} \prod_{Pa^G(X_i)} \left[\frac{\Gamma(\sum_j N'_{X_i, Pa^G(X_i), j})}{\Gamma(\sum_j N'_{X_i, Pa^G(X_i), j} + N_{X_i, Pa^G(X_i), j})} \prod_j \frac{\Gamma(N'_{X_i, Pa^G(X_i), j} + N_{X_i, Pa^G(X_i), j})}{\Gamma(N'_{X_i, Pa^G(X_i), j})} \right], \tag{3.5}$$

where $N_{i, Pa^G(X_i), j}$ is the count of variable i taking value j with parent configuration $Pa^G(X_i)$, N' represents the counts in the prior (equivalent sample size), and $\Gamma()$ is the standard Gamma function. With a prior for G, $P(\Theta)$ (a uniform one, for example), the BD score can be defined as: $\log P(D|\Theta) + \log P(\Theta)$.

We do not need to add an explicit penalty term, since the BD score implicitly penalizes complicated structures by integrating over the parameter space. Since specifying all the hyperparameters, $N'_{i, Pa^G(X_i), j}$, is non-trivial, the **BD equivalent** score (BDe) was developed [66] assuming likelihood equivalence ($P(\Theta \mid G) = P(\Theta \mid G')$) and that $P(G) > 0$ for any complete DAG G. Buntine [21] proposed a particular variation of BDe called **Bayesian Dirichlet equivalent uniform** (BDeu). This metric is the only score-equivalent BD score [24], that is, the only BD score that takes the same value for DAGs in the same equivalence class [157]. The downside is that in practice, BDeu is highly sensitive to the equivalent sample size N' chosen.

After defining a metric, we need to use a search algorithm to explore the search space of possible structures. Common algorithms for BN learning are **greedy search** and **local search**. However, since the graph space is highly "non-convex", both approaches can get

stuck at local optima. The greedy search, for example, the K3 algorithm [18], restricts the parents of each variable to the variables whose nodes come after it in the topological ordering. Then it adds the edge that increases a score by the largest amount, repeating the process until it converges. Local search algorithms start with an empty graph and either add, remove, or reverse the edges that contribute the most to the score, until convergence [41]. Of note is the Chow-Liu algorithm [27], that uses the simple maximum likelihood score, because it restricts the search space to tree-like structures, which limits overfitting. Finally, the integer linear programming (ILP) approach transforms the graph structure, scoring, and constraints into a linear programming (LP) problem, and then uses state-of-the-art LP solvers for the learning [11].

3.2.4 Parameters Learning

After the structure of the BN has been learned, parameter learning is the process of estimating all the **conditional probability tables** (CPT) associated with each node in the BN using the training data.

We assume that there are no missing data. However, algorithms for incomplete data, such as Expectation-Maximization (EM) [44], the Robust Bayesian Estimate (RBE) [133], the Monte-Carlo method [109], or the Gaussian Approximation [15], are also available. For complete data the **Bayesian Estimation**, and the **Maximum Likelihood Estimation** (MLE) are the best known ones.

Bayesian Estimation

For the discrete case, it assumes Θ is a random variable with a prior probability distribution $P(\Theta)$. The goal of the method is to calculate its posterior probability $P(\Theta \mid D)$, given the complete data D. Let $n_i = |\Omega_{X_i}|$ be the number of states of X_i, $\Theta_i = P(X = X_i)$, $\Theta_{i.k} = \{\Theta_{i,j,k} \mid j = 1, \ldots, n_i\}$, and $\Theta_{i,j,k} = P(X_i = j \mid Pa^G(X_i) = k)$. It makes the assumptions of local, and global independence, which together make for the parameter independence

assumption:

$$P(\Theta) = \prod_{i,k} P(\Theta_{i.k}), \tag{3.6}$$

where $P(\Theta_{i.k})$ is the Dirichlet distribution, and $P(\Theta)$ is the product Dirichlet distribution.

Maximum Likelihood Estimation

Likelihood is the standard approach to evaluate the quality of the parameters, Θ [77, 131], and it is given by:

$$L(\Theta : D) = P(D \mid \Theta) = \prod_{j=1}^{n} P(D_j \mid \Theta) = \prod_{i=1}^{r} L_i(\Theta_i : D),$$

where m is the number of samples, n the number of subjects, and r is the number of possible states. If $P(x_i \mid Pa^G(i))$ satisfies a polynomial distribution, the likelihood function can be decomposed as follows:

$$L_i(\Theta_i : D) = \prod_{j}^{X_i} \prod_{k}^{Pa^G(X_i)} \Theta_{i,j,k}, \tag{3.7}$$

where j represents the value X_i can take, and k iterates over each parent of X_i. If $N(\alpha)$ is the number of cases that satisfy condition α, then MLE can obtain the estimated parameters of Θ as:

$$\Theta_{i,j,k} = \frac{N(X_i = j, Pa^G(Xi) = k)}{N(Pa^G(Xi) = k)}. \tag{3.8}$$

3.3 Inference with Bayesian Networks

BNs perform three main inference tasks: **probabilistic** inference, **exact** inference, and **approximate** inference. However, none of these methods will be used in this thesis, since we adapt the less-Bayesian method of linear regression. Every edge of our BN is annotated with a regression coefficient λ_i, which can be used to easily infer the value of the child node, given the values of all the continuous parents. For the case of discrete parents, a

set of coefficients will be learned for each node configuration, and the inference will then proceed as outlined above.

3.4 Causality

Causality is the study of **cause** and **effect**. It is the influence that one event or process (cause) exerts on another event or process (effect). The cause is responsible (at least partly) for the effect, and the effect is (at least partly) dependent on the cause. More interesting for this thesis is the concept of causal inference; the process of extracting these cause/effect interactions between variables from training data.

3.4.1 Assumptions

Causal discovery from observational data depends on the following five assumptions[46, 187, 146]:

- **Causal Sufficiency**: Implies that there are no unmeasured common causes, i.e., hidden confounders. Formally, a set of variables **S** used is said to be *causally sufficient* or *causally complete*, if every common cause of any two or more variables in **S** are also in **S**.
- **Causal Markov Condition**: Establishes that once we know the parents of a variable X_i, then all other variables of G become independent of X_i. Formally, X_i only depends on $Pa^G(X_i)$, and is either marginally or conditionally independent of all other variables (see Section 3.1).
- **Faithfulness (Stability)**: Implies that only the variables that are d-separated in G will be independent (i.e., all others will be dependent), and vice-versa. This means that a data set D does not contain any "accidental" independence relationships that are not a consequence of the underlying causal model that generated them.
- **i.i.d.**: Assumes that all random variables are independent and identically distributed.

- **Consistency**: Implies that all causal relationships remain consistent for all subset of samples.

3.4.2 Inferring Causal Networks

Causal inference algorithms learn a causal BN where the edges represent not only statistical associations, but causal ones. They are usually learned by constraint-learning methods, and almost all stem from the SGS or IC algorithms. These methods evolved into the PC algorithm [170], which tries to minimize the number d-separation tests, for a growing set of neighbours. After the skeleton of the network is learned, it applies a series of rules to orient the edges, based on the V-structure concept. However, V-structures are not sufficient for discovering that X or Y causes Z. Unless we make the causal sufficiency assumption, the most that could be made is an acausal discovery [102]. This is not the case for Y-structures, where there is an extra arc from Z to W, which forms an un-confounded causal relationship [102]. The **PC-stable** [31] algorithm addresses the order-dependency problem of PC, i.e., its output is independent of the order of the input. However, it still produces too many spurious relationships in the presence of hidden confounders. The Fast Causal Inference (**FCI**) algorithm [31] applies two iterations of a modified version of PC-stable, and extra edge-orientation rules. It has a better behaviour in the presence of these hidden confounders, but it is computationally infeasible for large graphs. The **RFCI** algorithm [32] was developed to address this issue.

Time series data facilitates the learning of these networks, because causal effects can only go forward in time, helping in orienting the causal edges. Also, the stationarity and Markov assumptions further simplifies the learning of temporal causal graphs. **tsFCI** [47] uses a sliding window approach to be able to apply FCI without radical modifications, and orients the edges based on the nature of time. This method eventually evolved into **SVAR-GFCI** [101], which assumes the data-generation process is a structural vector autoregression (SVAR) with latent components. It is a hybrid method that combines a greedy

approach to select causal models with incrementally improved score, in addition to other improvements in the causal discovery and orientation rules. Other notable longitudinal causal discovery algorithm is **Tigramite** (PCMCI) [147], which applies PC-stable and then runs a momentary conditional independence (MCI) stage to further test for conditional independence, and computes p-values and adjusts for false discovery.

3.5 Spline Interpolation

Splines are piece-wise polynomials with boundary continuity and smoothness constraints used especially for interpolation problems. They are widely used because of their simplicity, accuracy, and ease of evaluation for a multitude of applications [10, 48, 75]. They can be used to smooth time series, avoiding problems such as overfitting, deal with irregular sampling rate, missing timepoints, and more. In this dissertation we use cubic splines, since they are the lowest degree polynomials that allow for an inflection point. A cubic spline can be expressed as follows:

$$s(t) = \sum_{i=1}^{n} C_i P_i(t), \quad t_{min} \leq t < t_{max}, \tag{3.9}$$

where $P_i(t)$ are the spline polynomials as a function of time t, and the C_is are the coefficients [8]. We need n equations to determine these coefficients, and it also typically requires that the value of the function, and its first and second derivatives are continuous at every boundary between the cubic pieces.

These splines can be made more mathematically flexible by writing down the cubic polynomial as a function of a set of four normalized *basis* functions [143], which are called B-splines. The basis coefficients C_i can be interpreted geometrically as *control points*, with only local influence on the curve. A periodic cubic B-spline ($o = 4$) can be expressed as:

$$s(t) = \sum_{i=1}^{m} C_i b_{i,4}(t), \quad k_4 \leq t \leq k_{n+1}, \tag{3.10}$$

where o is the order of the basis polynomials, and the values k_i are called knots (values of t at which the pieces join), where $i = 1, \ldots, n + o$.

3.6 Time Series Alignment

A challenge when comparing biological time series obtained from different samples is the fact that while the overall process studied in these individuals may be similar (i.e., follows the same sequence of steps), the *rates* of change may differ based on several "hidden" factors (age, gender, co-morbidities, etc.). Thus, prior to modeling the relationships between the different taxa we first "temporally" align the data sets between individuals by warping the time scale of each sample into the scale of another representative sample referred to as the *reference*. The goal of the temporal alignment step is to determine, for each individual i, a transformation function $\tau_i(t)$ which takes as an input a reference time t and outputs the corresponding time for individual i. Once the time series for each individual is transformed using their transformation function, we can compare corresponding values for all individuals sampled at equivalent time points. This approach effectively sets the stage for accurate discovery of trends and patterns, hence, further disentangling the dynamic and temporal relationships between entities in the microbiome.

While several options exist for the type of the transformation function τ_i that can be used in the temporal alignment, most methods used to date use either a linear, piecewise linear, or a polynomial function [7, 167]. Work on the analysis of gene expression data have shown that if the number of time points is relatively small then simpler functions tend to outperform the more complicated ones [9]. Therefore, we used a first degree polynomial (linear) form: $\tau_i(t) = \frac{(t-b)}{a}$ for the alignment function for tackling the temporal alignment problem, where a and b are the parameters of the transformation function.

Formally, let $s_r^j(t)$ be the spline curve for microbial taxa j at time $t \in [t_{min}, t_{max}]$ in the reference time-series sample r, where t_{min} and t_{max} denote the starting and ending time points of s_r^j, respectively. Similarly, let $s_i^j(t')$ be the spline for individual i in the set of

samples to be warped for taxa j at time $t' \in [t'_{min},\ t'_{max}]$. Next, analogously to Bar-Joseph *et al.* [7], we define the alignment error for microbial taxa j between s_r^j and s_i^j is defined as

$$e^j(r,i) = \frac{\int_\alpha^\beta (s_i^j(\tau_i(t)) - s_r^j(t))^2 dt}{\beta - \alpha}, \tag{3.11}$$

where $\alpha = \max\{t_{min}, \tau_i^{-1}(t'_{min})\}$ and $\beta = \min\{t_{max}, \tau_i^{-1}(t'_{max})\}$ correspond to the starting and ending time points of the alignment interval. Observe that by smoothing the curves, it is possible to estimate the values at any intermediate time point in the alignment interval $[\alpha, \beta]$.

3.7 Bioinformatics Concepts

Metagenomics is the science of the study of the genomes in a microbial community and constitutes the first step to investigating and understanding the microbiome [1], and its main purpose is the inference of the taxonomic profile of microbial community. The two main technologies used for this include *Whole Metagenome Sequencing* (WMS) and *16S rRNA* gene sequencing (MTS for metataxonomic sequencing). The two approaches capture different parts of the genomes of the microbes present in the sample: WMS captures DNA fragments from anywhere on the microbe's genome, while WTS captures and amplifies specific parts of the 16S rRNA gene. Consequently, WMS datasets provide a higher resolution of detection than WTS, but are more expensive and generate more data [135]. Standard analyses for sequencing datasets usually begin by aligning sequencing reads to a microbial reference database [177, 194], and producing abundance counts [178]. The quality of the analysis depends on the quality of the database of reference genomes. An alternative metagenomics analysis pipeline could perform *de novo* assembly of the reads into contigs. This method is limited by the ability of the algorithms to perform the assembly correctly.

Metatranscriptomics is the study of the genes expressed by the microbes in a microbial community. Unlike WMS, which captures DNA material in the sample, this approach captures the entire mRNA produced by transcriptional processes and that are present in the sample. With the use of functional annotations of expressed genes, it is possible to infer the functional profile of a community under specific conditions [110]. As with WMS, a metatranscriptomics analysis pipeline can either map the reads to a reference microbial genome or transcriptome database, or perform *de novo* assembly of the reads into contigs. As with WMS, the former is limited by the quality of the reference databases, while the latter is limited by the quality of the assembly algorithms.

Metabolomics is the study of the metabolites present in a sample [51]. The metabolome is considered the most direct indicator of the health status of an environment, and provides information about the metabolic activities in the microbiome. This is of interest because it complements the information provided by other omics, and sheds light on the result of microbial interactions.

To summarize, metagenomics answers the question "what is the microbial composition of a community?", metatrascriptomics answer the question "what genes are expressed by the microbial community?", and metabolomics answers "what byproducts are produced by the microbial community?".

3.8 Prior work in interaction inference

Microbes in the microbiome are not isolated, and interact with each other in a way that resembles a social network [50]. These interactions can be either positive (beneficial), or negative (harmful). The interactions shape the composition and function of the microbiome throughout time, thereby influencing the host's health [140] [50].

Early work on identifying which microbial taxa interacted with each other relied on the fact that two taxa cannot interact with each other consistently unless they have a strong pattern of co-occurrence (or co-avoidance) [50]. Thus strong correlations (positive or neg-

ative) between the vectors of abundance values of a pair of taxa was assumed to be a proxy for the strength of the relationship between the taxa. Using correlation techniques it is possible to learn a network where the nodes correspond to the entities of the microbiome, and the edges represent the strength of their co-occurrence (and by proxy, the potential strength of their relationships). Despite being useful in finding useful patterns in the microbial relationships in a microbiome, they have significant limitations in deciphering indirect relationships, and complex non-linear interactions such as between three or more species, and are unable to detect some ecological relationships [197]. An important problem is the lack of methods to detect false positives, i.e., pairs of co-occurring taxa that do not interact. Furthermore, because of the nature of correlations, these inferred relationships are symmetric and not directional. In conclusion, these limitations highlight the old adage that "correlation does not imply causation".

An alternative is to use the generalized Lotka-Volterra (gLV) equations, which are able to describe the time-dependent population dynamics and predict ecological relationships between members of different biological species. They are based on non-linear differential equations, and have been widely used in the literature [173, 162, 52, 103, 112, 93, 56, 78]. However, gLV-based systems are best applied to understand transient behaviors, while most of the data sets we consider are for systems that have reached homeostasis, with relatively minor interest in understanding the dynamics prior to reaching equilibrium.

Another solution is to use Bayesian techniques. BNs suggest possible directional interactions between the entities in a microbiome and can address the major limitation of correlation analysis. Thus, Bayesian techniques provide a powerful approach in revealing complex associations within microbes. They include implicit parameter estimation techniques for inferring complex networks from noisy data, but have not been widely used in the microbiome context [107, 96].

CHAPTER 4

DYNAMIC BAYESIAN NETWORKS

4.1 Dynamic Bayesian network models

Dynamic Bayesian Networks (DBNs) are a flavor of BNs suited for the representation of temporal connections between variables, and conducting time-varying probabilistic inference and causal analysis under system uncertainty, because its edges represent lagged dependencies. They were developed to unify models such as Kalman filters, autoregressive–moving-average models (ARMA), or hidden Markov models (HHMs) into a general probabilistic model and inference mechanism [38, 39], and are conceptually similar to Probabilistic Boolean Networks (PBN) [87]. DBNs can model all the above types of correlation and capture even more complex relationships.

Up to now, various DBN models based on different probabilistic representations have been proposed in the literature [191]:

- Discrete models [121, 204]
- Multivariate autoregressive process (VAR) and structural VARs (SVARs), especially popular in economics [165, 122].
- State Space (SSM) or Hidden Markov Models [13, 130, 134, 200],
- Nonparametric additive regression models [74, 84, 174], and
- Low-order independence models [89].

In this dissertation, we are going to focus on a version of DBNs called Two-Timeslice BN (2TBN), which relates variables to each other over adjacent time steps. The variables can be connected by intra edges (within the same time slice) or inter edges (between time slices). Any variable X_i^t can be calculated from the internal regressors, the current time point t and the previous time point $t-1$. This is possible because of the following two assumptions:

- **Stationarity** or time invariance: Assumes that the data was generated by a stochastic stationary process. This means that that the mean, variance and autocorrelation structure do not change over time. Because of this, if the DBN has an edge between X_i and X_j ($X_i \not\perp\!\!\!\perp X_j$) in any time slice, that edge is also present for all of the time slices. Formally,

$$P(X_i^{t+1} \mid X_j^t) \perp\!\!\!\perp t \quad \forall i,\ j \in \mathbb{R},\ t \in [0, T]. \tag{4.1}$$

  There has been some effort in developing DBNs that relax this assumption, and are called non-stationary dynamic Bayesian networks (nsDBNs), but they are not going to be used in this work [142, 61, 60].

- **First order Markov assumption** or forgetting rule. Under this assumption, all variables of the next timepoint are independent of all variables from the previous timepoint, given the current timepoint. Formally,

$$X_i^{t+1} \perp\!\!\!\perp X_i^{t-1} \mid X_i^t \quad \forall\ 0 \leq i < n. \tag{4.2}$$

4.2 Model construction

We use a "two-stage" DBN model (2TBN) in which only two slices are modeled and learned at a time. Typically, analysis using DBNs is divided into two components: learning the network structure and parameters and inference on the network. The former can be further sub-divided into (i) structure learning which involves inferring from data the causal connections between nodes (i.e., learning the intra and inter edges) while avoiding overfitting the model, and (ii) parameter learning which involves learning the parameters of each intra and inter edge in a specific network structure. There are only a limited number of open software packages that support both learning and inference with DBNs [106, 198] in the presence of discrete and continuous variables. We used the freely available CGBayesNets

package [106, 108] for learning the network structure and performing inference for Conditional Gaussian Bayesian models [88]. While useful, CGBayesNets does not support several aspects of DBN learning including the use of intra edges, searching for a parent candidate set in the absence of prior information and more. We have thus extended the structure learning capabilities of CGBayesNets to include intra edges while learning network structures and implemented well-known network scoring functions for penalizing models based on the number of parameters such as Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) [128].

4.2.1 Learning DBN model parameters

Let Θ denote the set of parameters for the DBN and G denote a specific network structure over discrete and continuous variables in the microbiome study. As in McGeachie *et al.* [108], we can decompose the joint distribution as

$$D(\Delta)C(\Psi|\Delta) = \prod_{X \in \Delta} P(X \mid Pa^G(X)) \prod_{Y \in \Psi} F(Y \mid Pa^G(Y)), \tag{4.3}$$

where D denotes a set of conditional probability distributions over discrete variables Δ, C denotes a set of linear Gaussian conditional densities over continuous variables Ψ, and $Pa^G(X)$ denotes the set of parents of the node representing variable X in G. Since we are dealing with both, continuous and discrete variables in the DBN, we will often use the term "continuous node" (or "discrete node") to mean that the variable represented by the node is continuous (or discrete, respectively). In our method, continuous variables (i.e., microbial taxa compositions) are modeled using a Gaussian with the mean that is set based on a regression model over the set of continuous parents as follows

$$F(Y \mid U_1, \cdots, U_k) \sim N(\lambda_0 + \sum_{i=1}^{k} \lambda_i \times U_i, \sigma^2), \tag{4.4}$$

where $U_1, \cdots, U_k$ are continuous parents of Y; λ_0 is the intercept; $\lambda_1, \cdots, \lambda_k$ are the corresponding regression coefficients for $U_1, \cdots, U_k$; and σ^2 is the standard deviation. For example, the conditional linear Gaussian density function for variable $T_{4\_t_{i+1}}$ in Fig. 5.3e is modeled by

$$F(T_{4\_t_{i+1}} \mid T_{4\_t_i}, C_{3\_t_i}, T_{2\_t_{i+1}}) = N(\lambda_0 + \lambda_1 \times T_{4\_t_i} + \lambda_2 \times C_{3\_t_i} + \lambda_3 \times T_{2\_t_{i+1}}, \sigma^2), \tag{4.5}$$

where $\lambda_1, \lambda_2, \lambda_3$ and σ^2 are the DBN model parameters. We note that if Y has discrete parents then we need to compute coefficients $L = \{\lambda_i\}_{i=0}^{k}$ and standard deviation σ^2 for the configuration of each discrete parent. In general, given a longitudinal data set D and known structure G, we can directly infer the parameters Θ by maximizing the likelihood of the data given by our regression model.

4.2.2 Learning DBN structure

Learning the DBN structure can be expressed as finding the structure and parameters that optimizes the following likelihood expression:

$$\max_{\Theta, G} P(D \mid \Theta, G) P(\Theta, G) = \max_{G} P(D, \Theta \mid G) P(G), \tag{4.6}$$

where $P(D \mid \Theta, G)$ is the likelihood of the data given the model. Intuitively, the likelihood increases as the number of valid parents $Pa^G(\cdot)$ increases, thus, making it challenging to infer the most accurate model for data set D. Therefore, the goal is to effectively search over possible structures while using a function that penalizes overly complicated structures and protects from overfitting.

Here, we maximize $P(D, \Theta \mid G)$ for a given structure G using maximum likelihood estimation (MLE) coupled with BIC score instead of the Bayesian Dirichlet equivalent sample-size uniform (BDeu) metric used in CGBayesNets [106, 108]. The BDeu score

requires prior knowledge (i.e., equivalent sample size priors) which are typically arbitrarily set to 1; however, multiple studies have shown the sensitivity of BDeu to these parameters [163, 171], as well as the harm in using improper prior distributions [119]. In contrast, BIC score does not depend on the prior over the parameters, and is thus an ideal approach for scenarios where prior information is not available or difficult to obtain. The modified log likelihood score using BIC is give as follows:

$$BIC(G, D) = \log P(D \mid \Theta, G) - \frac{|\Theta|}{2} \log |D|, \tag{4.7}$$

where $|\Theta|$ is the number of DBN model parameters in structure G, and $|D|$ is the number of observations in D. We also tried the AIC score given as follows: $\text{AIC}(G, D) = \log P(D \mid \Theta, G) - |\Theta|$. However, preliminary results have shown that the BIC score consistently outperformed the AIC score (data not shown).

Next, in order to maximize the full log-likelihood term we implemented a greedy hill-climbing algorithm. We initialize the structure by first connecting each taxa node at the previous time point (for example $T_{1\_t_i}$ in Fig. 5.3e) to the corresponding taxa node at the next time point ($T_{1\_t_{i+1}}$ in Fig. 5.3e). We call this setting the *baseline* model since it ignores dependencies between the taxa and only tries to infer taxon abundance levels based on its levels at previous time points. For each node in the DBN, we considered adding edges leading into it as follows. We ignored all parents that were disallowed by the specified constraints. We also ignored all parents for which the edges if added would form cycles. For each remaining candidate, we computed the increase of the log-likelihood score that resulted by adding edge from the parent to the node, and greedily picked the edge that would lead to the largest increase until the bound on the number of parents was reached. Note that we imposed an upper bound limit on the maximum number of possible parents ($maxParents \in \{1, 3, 5\}$) for each microbial taxon abundance node X (i.e., $|Pa^G(X)| \leq maxParents$).

4.2.3 Constraining the DBN structure

An important innovation in the work described in Chapters 5 and 6 lies in how we constrain the DBN structure to conform to our proposed framework of relationships and desired direction of interactions. These constraints (in the form of a matrix received as an input to the function) only allow edges between certain types of nodes, highly reducing the complexity of searching over possible structures and preventing over-fitting. Note that these constraints can be readily changed during execution of the software by adding more data types or different restrictions in the input file containing the adjacency matrix. Specifically, we allowed intra edges from environmental and host transcriptomics variables to microbial taxa (abundance) nodes, from taxa nodes to gene (expression) nodes and from gene nodes to metabolites (concentration) nodes. All other interactions within a time point (for example, direct gene to taxa) were disallowed. We also allowed inter edges from metabolites to taxa nodes in the next time point, and *self-loops* from any node X_i^t to X_i^{t+1} for each i and time t, except for environmental or host transcriptomics variables for which no incoming edges were allowed (host genes were only measured at a single time point so no incoming temporal edges were allowed for them). These restrictions reflect our understanding of the basic ways the different entities interact with each other, i.e., environmental and host gene expression variables are independent variables, taxa express genes, which are involved in metabolic pathways; finally, the metabolites impact the growth of taxa (in the next time slice).

4.3 Summary of DBN contributions

The following is a summary of the modifications made to the standard DBN and its learning capabilities based on the CGBayesNets package [106, 108]:

1. Added self-loop initialization that connects variable X_i^t to X_i^{t+1} for each i and time t.

2. Added support for BIC and AIC global measures, as opposed to the local measures used in the original implementation.
3. Added regression coefficient and bootstrap score labelling into the edges of the output network.
4. Re-implemented MLE capabilities.
5. Added parallel support for cross-validation and bootstrap learning.
6. Allowed for a multi-omic restriction framework to be inputed to the function, with dynamic learning constraints based on the node types.
7. Added support for a prior network structure, that starts the learning process with a skeleton structure of edges that do not have to be learned before the greedy hill-climbing process starts.

CHAPTER 5

LONGITUDINAL MICROBIAL NETWORK INFERENCE

Several studies have focused on the microbiota living in environmental niches including human body sites. In many of these studies researchers collect longitudinal data with the goal of understanding not just the composition of the microbiome but also their evolution over time. However, analysis of such data is challenging and very few methods have been developed to reconstruct dynamic models from time series microbiome data. Even in existing studies with limited time points, they do not focus on the interactions between the different taxa or the evolution of the interactions.

In this chapter we present a computational pipeline called PRIMAL that enables the integration of data across individuals for the reconstruction of such models. Our pipeline starts by aligning the data collected for all individuals. The aligned profiles are then used to learn a dynamic Bayesian network which represents causal relationships between taxa and clinical variables. Testing our methods on three longitudinal microbiome data sets we show that our pipeline improve upon prior methods developed for this task. We also discuss the biological insights provided by the models which include several known and novel interactions. The extended CGBayesNets package is freely available under the MIT Open Source license agreement. The source code and documentation can be downloaded from `http://biorg.cis.fiu.edu/primal/`.

In this chapter we propose a computational pipeline for analyzing longitudinal microbiome data. Our results provide evidence that microbiome alignments coupled with dynamic Bayesian networks improve predictive performance over previous methods and enhance our ability to infer biological relationships within the microbiome and between taxa and clinical factors.

5.1 Background

Multiple efforts have attempted to study the microbiota living in environmental niches including human body sites. These microbial communities can play beneficial as well as harmful roles in their hosts and environments. For instance, microbes living in the human gut perform numerous vital functions for homeostasis ranging from harvesting essential nutrients to regulating and maintaining the immune system. Alternatively, a compositional imbalance known as dysbiosis can lead to a wide range of human diseases [26], and is linked to environmental problems such as harmful algal blooms [3].

While many studies profile several different types of microbial taxa, it is not easy in most cases to uncover the complex interactions within the microbiome and between taxa and clinical factors (e.g., gender, age, ethnicity). Approaches based on co-occurrence patterns of microbial taxa are useful in implicating interactions between taxa [50, 197], but only provide circumstantial evidence. Approaches based on causal inferencing have been tried using metagenomics data [152], but these methods have not ventured into analyzing longitudinal microbiome data yet, with the isolated attempt from TIME [6] which is based on Granger causality . Microbiomes are inherently dynamic, thus, in order to fully reconstruct these interactions we need to obtain and analyze longitudinal data [57]. Examples include characterizing temporal variation of the gut microbial communities from pre-term infants during the first weeks of life, and understanding responses of the vaginal microbiota to biological events such as menses. Even when such longitudinal data is collected, the ability to extract an accurate set of interactions from the data is still a major challenge.

To address this challenge we need computational time-series tools that can handle data sets that may exhibit missing or noisy data and non-uniform sampling. Furthermore, a critical issue which naturally arises when dealing with longitudinal biological data is that of temporal rate variations. Given longitudinal samples from different individuals (for example, gut microbiome), we cannot expect that the rates at which interactions take place is

exactly the same between these individuals. Issues including age, gender, external exposure, etc. may lead to faster or slower rates of change between individuals. Thus, to analyze longitudinal data across individuals we need to normalize for these variable rates by first temporally aligning the microbial data. Using the aligned profiles we can next employ other methods to construct a model for the process being studied.

Most current approaches for analyzing longitudinal microbiome data focus on changes in outcomes over time [55, 86]. The main drawback of this approach is that individual microbiome entities are treated as independent outcomes, hence, potential relationships between these entities are ignored. An alternative approach involves the use of dynamical systems such as the generalized Lotka-Volterra (gLV) models [28, 104, 172, 179, 58]. While gLV and other dynamical systems can help in studying the stability of temporal bacterial communities, they are not well-suited for temporally sparse and non-uniform high-dimensional microbiome time series data (e.g., limited frequency and number of samples), as well as noisy data [57, 58]. Additionally, most of these methods eliminate any taxa whose relative abundance profile exhibits a zero entry (i.e., not present in a measurable amount at one or more of the measured time points. Finally, probabilistic graphical models (e.g., hidden Markov models, Kalman filters and dynamic Bayesian networks) are machine learning tools which can effectively model dynamic processes, as well as discover causal interactions [108].

In this work we first adapt statistical spline estimation and dynamic time-warping techniques for aligning time-series microbial data so that they can be integrated across individuals. We use the aligned data to learn a Dynamic Bayesian Network (DBN), where nodes represent microbial taxa, clinical conditions, or demographic factors and edges represent causal relationships between these entities. We evaluate our model by using multiple data sets comprised of the microbiota living in niches in the human body including the gastrointestinal tract, the urogenital tract and the oral cavity. We show that models for these systems can accurately predict changes in taxa and that they greatly improve upon models

constructed by prior methods. Finally, we characterize the biological relationships in the reconstructed microbial communities and discuss known and novel interactions discovered by these models.

5.2 Methods

5.2.1 Data sets

We collected multiple public longitudinal microbiome data sets for testing our method. Table 5.1 summarizes each longitudinal microbiome data set used in this study, including the complete list of clinical features available.

Data set	n_i	n_s	n_t	Sampling	Clinical attributes
Infant gut	58	922	29	Every day or two	Day of life Gestational age at birth Post-conceptional age Gender (female or male) Mode of birth (C-section or vaginal delivery) Room type (single or multi-patient) Human milk used (% of enteral volume provided by human milk) Days of antibiotics (% of days of life on antibiotic)
Vaginal	32	937	330	Twice a week	Day period (days since menses started) Nugent category (low, intermediate , high) Age group ($\leq 30, > 30$ and ≤ 40, or > 40) Race (white, black, hispanic, or other) Tampon used (yes or no) Vaginal douching (yes or no) Sexual activity (yes or no)
Oral cavity	18	374	1,420	Every week during gestation	Gestational day of delivery Ethnicity (hispanic or non-hispanic)

Table 5.1: **Summary of longitudinal microbiome data sets**. For each data set, we show the total number of individuals n_i, number of time series samples n_s, number of microbial taxa reported n_t, original sampling rate and list of clinical attributes available.

Infant gut microbiome This data set was collected by La Rosa *et al.* [86]. They sequenced gut microbiome from 58 pre-term infants in neonatal intensive care unit (NICU). The data was collected during the first 12 weeks of life (until discharged from NICU or deceased) sampled every day or two on average. Following analysis 29 microbial taxa were reported across the 922 total infant gut microbiome measurements. In addition to the taxa information, this data set includes clinical and demographic information for example, gestational

age at birth, post-conceptional age when sample was obtained, mode of delivery (C-section or vaginal), antibiotic use (percentage of days of life on antibiotic), and more (see Table 5.1 for complete list of clinical features available).

Vaginal microbiome The vaginal microbiota data set was collected by Gajer *et al.* [55]. They studied 32 reproductive-age healthy women over a 16-week period. This longitudinal data set is comprised of 937 self-collected vaginal swabs and vaginal smears sampled two times a week. Analysis identified 330 bacterial taxa in the samples. The data also contains clinical and demographic attributes on the non-pregnant women such as Nugent score [118], menses duration, tampon usage, vaginal douching, sexual activity, race and age. To test the alignment methods we further sub-divided the microbial composition profiles of each subject by menstrual periods. This resulted in 119 time-series samples, an average of 3-4 menstrual cycles per woman. Figure 5.1a shows four sub-samples derived from an individual sample over the 16-week period along with corresponding menses information.

Oral cavity microbiome The oral cavity data was downloaded from the case-control study conducted by DiGiulio *et al.* [45] comprised of 40 pregnant women, 11 of whom delivered pre-term. Overall they collected $3,767$ samples and identified a total of $1,420$ microbial taxa. Data was collected weekly during gestation and monthly after delivery from four body sites: vagina, distal gut, saliva, and tooth/gum. In addition to bacterial taxonomic composition, these data sets report clinical and demographic attributes which include gestational status, gestational or postpartum day when sample was collected, race and ethnicity. In this paper, we solely focus on the tooth/gum samples during gestation from Caucasian women in the control group to reduce potential confounding factors. This restricted set contains 374 temporal samples from 18 pregnant women.

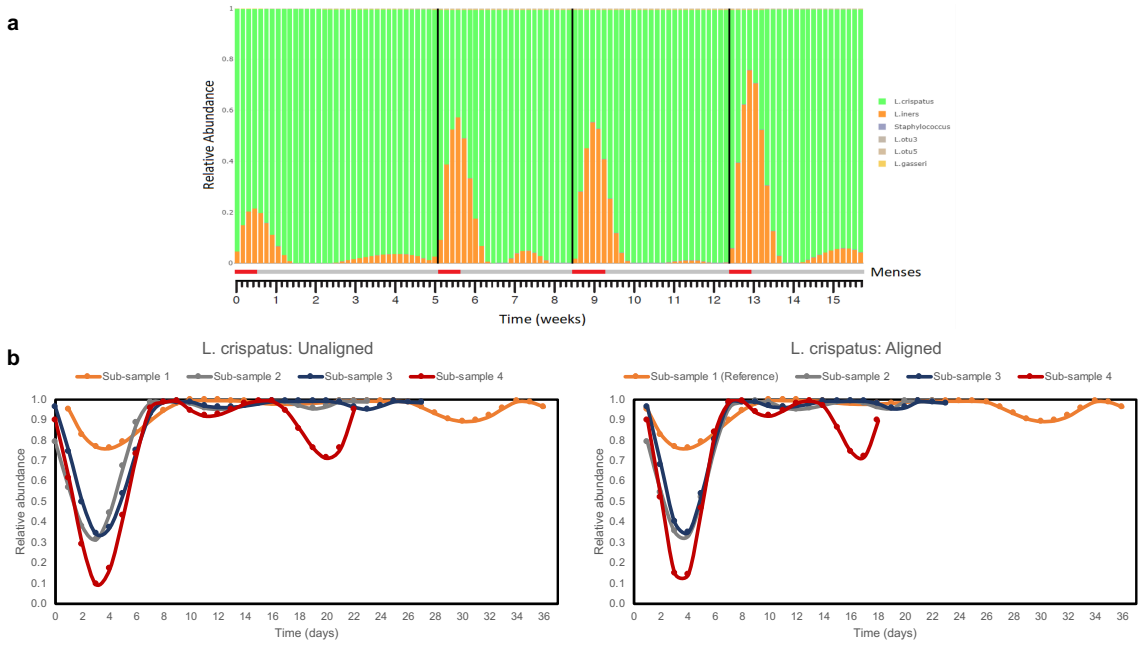

Figure 5.1: **Representative vaginal microbiome sample for subject** 28 **over the** 16**-week period**. **a)** Relative abundance profile of six vaginal taxa for subject 48 over 16 weeks annotated with menses information. The vertical black lines correspond to the division of sub-samples based on menstrual periods (i.e., 4 sub-samples). Note the interpolated shift in dominance during menses between *L. crispatus* and *L. iners*. **b)** Temporal alignment between the sub-samples from subject 28 time-series data for taxa *L. crispatus* using the first menstrual period sub-sample as reference (shown in orange). Figure also shows abundance profile of *L. crispatus* for each sub-sample before (left) and after (right) alignment.

5.2.2 Data pre-processing

Since alignment relies on continuous (polynomial) functions while the data is sampled at discrete intervals, the first step is to represent the sample data using continuous curves as shown by the transition from Fig. 5.3a to Fig. 5.3b. Following prior work [9], we use B-splines for fitting continuous curves to microbial composition time-series data, thus, enabling principled estimation of unobserved time points and interpolation at uniform intervals. To avoid overfitting we removed any sample that had less than nine measured time points. The resulting pre-processed data is comprised of 48 individual samples of the infant gut, 116 sub-samples of the vaginal microbiota and 15 pregnant women samples of the oral microbiome. We next estimated a cubic B-spline from the observed abundance profile for all taxa in remaining samples using *splrep* and *BSpline* from the Python func-

tion *scipy.interpolate*. In particular, *splrep* is used to find the B-spline representation (i.e., vector of knots, B-spline coefficients, and degree of the spline) of the observed abundance profile for each taxa, whereas *BSpline* is used to evaluate the value of the smoothing polynomial and its derivatives. Figure 5.2 shows the original and cubic spline of a representative microbial taxa from a randomly selected individual sample across each data set.

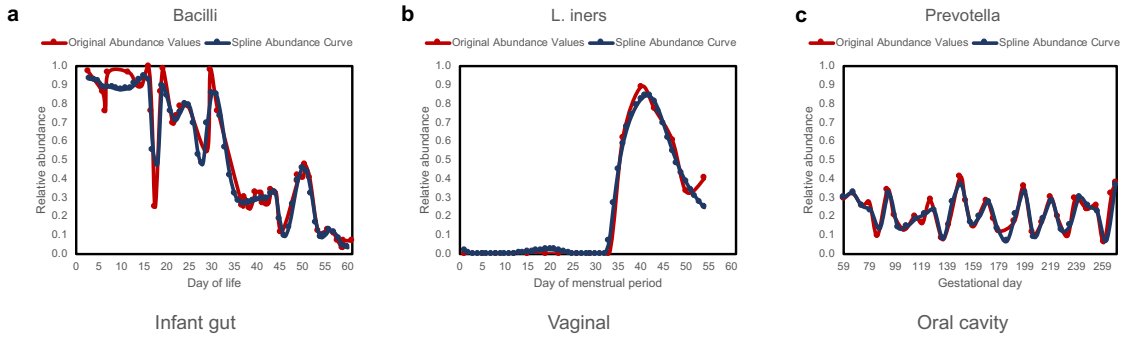

Figure 5.2: **Original and cubic spline of the abundance profile of a representative microbial taxa for each data set**. Figure shows the original abundance values vs. the cubic B-spline curve for a representative taxa profile from a randomly selected individual sample across each data set. **a)** *Bacilli* from the infant gut microbiome. **b)** *L. iners* from the vaginal microbiome. **c)** *Prevotella* from the oral cavity microbiome.

5.2.3 Aligning microbial taxon

To discuss the alignment algorithm we first assume that a reference sample, to which all other samples would be aligned, is available. In the next section we discuss how to choose such reference.

Formally, let $s_r^j(t)$ be the spline curve for microbial taxa j at time $t \in [t_{min}, t_{max}]$ in the reference time-series sample r, where t_{min} and t_{max} denote the starting and ending time points of s_r^j, respectively. Similarly, let $s_i^j(t')$ be the spline for individual i in the set of samples to be warped for taxa j at time $t' \in [t'_{min}, t'_{max}]$. Next, analogously to Bar-Joseph *et al.* [7], the alignment error for microbial taxa j between s_r^j and s_i^j is defined as

$$e^j(r,i) = \frac{\int_\alpha^\beta (s_i^j(\tau_i(t)) - s_r^j(t))^2 dt}{\beta - \alpha}, \tag{5.1}$$

where $\alpha = \max\{t_{min}, \tau_i^{-1}(t'_{min})\}$ and $\beta = \min\{t_{max}, \tau_i^{-1}(t'_{max})\}$ correspond to the starting and ending time points of the alignment interval. Observe that by smoothing the curves, it is possible to estimate the values at any intermediate time point in the alignment interval $[\alpha, \beta]$. Finally, we define the microbiome alignment error for a microbial taxon of interest S between individual samples r and i as follows

$$E_M(r, i) = \sum_{j \in S} e^j(r, i). \tag{5.2}$$

Given a reference r and microbial taxon S, the alignment algorithm task is to find parameters a and b that minimize E_M for each individual sample i in the data set subject to the constraints: $a > 0$, $\alpha < \beta$ and $\frac{(\beta-\alpha)}{(t_{max}-t_{min})} \geq \epsilon$. The latter constraint enforces that the overlap between aligned interval $[\alpha, \beta]$ and reference interval $[t_{min}, t_{max}]$ is at least ϵ, otherwise trivial solutions (for example, no overlap leading to 0 error) would be selected. Here we used $\epsilon = 0.3$ though results remain the same with larger values of ϵ. Fig. 5.3c illustrates an aligned set of four samples where reference sample r is shown in orange. Alternatively, Figure 5.1b shows the temporal alignment between the sub-samples of the vaginal microbiome sample shown in Figure 5.1 for the taxon *L. crispatus* using the first menstrual period sub-sample as reference (shown in orange).

5.2.4 Selecting a reference sample

Finding an optimal reference that jointly minimizes the error for all samples (E_M) is akin to solving a multiple alignment problem. Optimal solutions for such problems still require a runtime that is exponential in the number of samples [7] and so a heuristic approach was used instead. For this, we first find the best pairwise alignments via a grid-search parameter sweep between $a \in (0, 4]$ with increments of 0.01 and $b \in [-50, 50]$ with increments of 0.5 in the linear alignment function τ_i previously described. It is important to note that this restricted search space for parameters a and b may lead to some sample pairs

(r, i) without a temporal alignment because overlap constraint is not met. Additionally, we filtered out any microbial taxa $j \in S$ for which the mean abundance in either s_r^j or s_i^j was less than 0.1%, or had zero variance over the originally sampled time points. Lastly, an optimal reference for each data set is determined by generating all possible pairwise alignments between samples. To select the best reference r^* we employed the following criteria: (1) at least 90% of the individual samples are aligned to r^*, and (2) the alignment error E_M is minimized. We note that if no candidate reference meets these criteria, a commonly used heuristic for selecting r^* picks the sample with the longest interval or highest number of measured time points.

Abnormal or noisy samples filtering As a post-processing step, we implemented a simple procedure which takes as input the resulting individual-wise alignments to identify and filter out abnormal and noisy samples. Given an aligned microbiome data set, we (1) computed the mean μ and standard deviation δ of the alignment error E_M across all aligned individual samples, and (2) removed all samples from an individual where $E_M > \mu+(2\times\delta)$. Fig. 5.3d shows the filtered set for the aligned taxa in the previous step (Fig. 5.3c). This analysis can both, help to identify outliers and to improve the ability to accurately reconstruct models for interactions between taxa as shown in the Results section.

Taxon selection from alignment As previously described, the microbiome alignment error E_M for a pairwise alignment is restricted to the set of microbial taxa S that contributed to the alignment. However, this set of microbes can vary for different pairwise alignments even with the same reference. Therefore, we focused on the subset of taxa that contributed to at least half of the pairwise alignments for the selected reference. Table 5.2 lists alignment information for each data set such as reference sample, number of aligned samples and selected taxa.

Alignment simulation experiments Since temporal alignment using splines does not guarantee convergence to a global minimum [7], we performed simulation studies to investigate the susceptibility to the non-uniqueness and local optima of the splines-based

Data set	Reference sample	n_r	Selected taxa
Infant gut	Subject 48	47	*Actinobacteria* *Alphaproteobacteria* *Bacilli* *Bacteroidia* *Betaproteobacteria* *Clostridia* *Cyanobacteria* *Epsilonproteobacteria* *Erysipelotrichi* *Flavobacteria* *Fusobacteria* *Gammaproteobacteria* *Holophagae* *Unclassified*
Vaginal	Subject 26, Menses 1	112	*Aerococcus* *Anaerococcus* *Ureaplasma* *Parvimonas* *L. iners* *Finegoldia* *Staphylococcus* *Porphyromonas* *Atopobium* *Gardnerella* *L. crispatus* *Peptostreptococcus* *Sneathia* *Streptococcus* *Prevotella* *Peptoniphilus* *Incertae_Sedis_XI.1* *L. gasseri* *Corynebacterium* *Dialister* *L. otu5* *L. otu3* *Ruminococcaceae.3*
Oral cavity	Subject T18	14	*H. parainfluenzae* *Streptococcus* *Gemellaceae* *Porphyromonas* *Streptococcus* *Leptotrichia* *P. nanceiensis* *V. dispar* *Granulicatella* *Veillonella* *Streptococcus* *Neisseria* *N. subflava* *Fusobacterium* *P. melaninogenica* *Haemophilus* *Prevotella* *Porphyromonas* *Granulicatella*

Table 5.2: **Summary of alignment information**. For each data set, we show reference sample, number of aligned samples n_r and selected taxa.

heuristic approach described at the beginning of this section. In particular, we first used the originally measured time points and observed abundance profile from three taxa of a representative individual sample in the gut data set as the reference sample. We then simulated 10 different individual samples as follows: for each individual sample, we manually warped the time points with randomly selected parameters a (scaling) and b (translation) such that $a \in (0, 4]$ and $b \in [0, 50]$. We next added distinct percentage of Gaussian noise selected from $\{0, 5, 10, 15, 20, 25\}$ to the warped time points. To further test the robustness of splines, we also added Gaussian noise to the observed abundance profile of each taxa. Finally, we conducted three types of simulation experiments: (1) simulated noise-free warped time points for each individual sample but with noisy abundance profile, (2) simulated noise-free abundance profile but with noisy warped time points, and (3) noisy simulated warped time points with noisy abundance profiles.

From each simulation experiment, we aligned all simulated individual samples to the reference sample. We then computed and reported the mean absolute error (MAE) between the observed alignment parameters (i.e, a and b), as well as alignment error E_M on the aligned simulated data.

5.2.5 Dynamic Bayesian network models

Bayesian Networks (BNs) are a type of probabilistic graphical model consisting of a directed acyclic graph (see Section 3.1). In a BN model, the nodes correspond to random variables and the directed edges correspond to potential conditional dependencies between them. The absence of an edge connecting two variables indicates independence or conditional independence between them. Conditional independence allows for a compact, factorized representation of the joint probability distribution [148]. *Dynamic Bayesian networks* (DBNs) are BNs better suited for modeling relationships over temporal data (see Section 4.1). Instead of building different models across time steps, DBNs allow for a "generic slice" that shows transitions from a previous time point to the next time point, thus rep-

resenting a generic temporal transition that can occur at any time during the computation. The incorporation of conditional dependence and independence is similar to that in BNs. DBNs have been widely used to model longitudinal data across many scientific domains, including speech [116, 205], biological [42, 62, 108], or economic sequences [141, 196].

More formally, a DBN is a directed acyclic graph where, at each *time slice* (or time instance), nodes correspond to random variables of interest (e.g., taxa, post-conceptional age, or Nugent score) and directed edges correspond to their conditional dependencies in the graph. These time slices are not modeled separately. Instead a DBN contains edges connecting time slices known as *inter edges* that are repeated for each time point modeled as depicted in Fig. 5.3e. In summary, the model learns the transition probability from one time point to the next as a stationary conditional probability. DBNs are considered generative models, therefore, ideal for modeling the compositional interactions and dynamics of the microbiota given the first time point.

5.2.6 Model construction

Using the aligned time series for the abundance of taxa, we next attempted to learn graphical models that provide information about the dependence of the abundance of taxa on the abundance of other taxa and clinical or demographic variables. Here, we use a "two-stage" DBN model in which only two slices are modeled and learned at a time. Throughout this dissertation, we will refer to the previous and current time points as t_i and t_{i+1} for variable t, respectively. Fig. 5.3e illustrates a skeleton of the general structure of a two-stage DBN in the context of a longitudinal microbiome study. In this example, for each time slice, the nodes correspond to random variables of observed quantities for different microbial taxa (T_1, T_2, T_3, T_4) or clinical factors (C_1, C_2, C_3) shown as circles and diamonds, respectively. These variables can be connected by intra edges (dotted lines) or inter edges (solid lines). In this DBN model, the abundance of a particular microbe in the current

time slice is determined by parameters from both intra and inter edges, thus, modeling the complex interactions and dynamics between the entities in the microbial community.

Typically, analysis using DBNs is divided into two components: learning the network structure and parameters and inference on the network (see Section 4.2 for more information). The former can be further sub-divided into (i) structure learning which involves inferring from data the causal connections between nodes (i.e., learning the intra and inter edges) while avoiding overfitting the model, and (ii) parameter learning which involves learning the parameters of each intra and inter edge in a specific network structure. There are only a limited number of open software packages that support both learning and inference with DBNs [106, 198] in the presence of discrete and continuous variables. Here we used the freely available CGBayesNets package [106, 108] for learning the network structure and performing inference for Conditional Gaussian Bayesian models [88]. While useful, CGBayesNets does not support several aspects of DBN learning including the use of intra edges, searching for a parent candidate set in the absence of prior information and more. We have thus extended the structure learning capabilities of CGBayesNets to include intra edges while learning network structures and implemented well-known network scoring functions for penalizing models based on the number of parameters such as Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) [128] instead of the BDeu score. Additionally, we imposed an upper bound limit on the maximum number of possible parents ($maxParents \in \{1, 3, 5\}$) for each bacterial node X (i.e., $|Pa^G(X)| \leq maxParents$). For more information about the mathematical formulation and a more detailed description of our contributions, please refer to Chapter 4.

5.2.7 Inferring biological relationships

Microbial ecosystems are complex, often displaying a stunning diversity and a wide variety of relationships between community members. These biological relationships can be broadly divided into two categories: *beneficial*: (including mutualism, commensalism

and obligate), or *harmful* (including competition, amensalism and parasitism). Although the longitudinal data sets considered in this study do not provide enough information to further sub-categorize each biological relationship (e.g., mutualism vs. commensalism), we use the learned DBN model from each microbiome data set and inspect each interaction as a means for inferring simple to increasingly complex relationships. For example, consider variable $T_{4\_t_i}$ in Fig. 5.3e. Given that t_i and t_{i+1} represent the previous time point and the current time point (respectively), the possible inference in this case is as follows: Edges from $T_{4\_t_i}$ and $C_{3\_t_i}$ (inter edges), and from $T_{2\_t_{i+1}}$ (intra edge) suggest the existence of a temporal relationship in which the abundance of taxa T_4 at a previous time instant and abundance of taxa T_2 at the current time instant, as well as condition C_3 from the previous time instant impact the abundance of T_4 at the current time. We previously stated that $F(T_{4\_t_{i+1}} \,|\, T_{4\_t_i}, C_{3\_t_i}, T_{2\_t_{i+1}})$ is modeled by $N(\lambda_0 + \lambda_1 \times T_{4\_t_i} + \lambda_2 \times C_{3\_t_i} + \lambda_3 \times T_{2\_t_{i+1}}, \sigma^2)$. Therefore, inspecting the regression coefficients $\lambda_1, \lambda_2, \lambda_3$ immediately suggests whether the impact is positive or negative. In this example, the regression coefficients λ_1, λ_2 are positive ($\lambda_1, \lambda_2 > 0$) while coefficient λ_3 is negative ($\lambda_3 < 0$), thus, variables $T_{4\_t_i}$ and $C_{3\_t_i}$ exhibit positive relationships with microbial taxa $T_{4\_t_{i+1}}$ shown as green edges in Fig. 5.3e, whereas taxa $T_{2\_t_i}$ exhibits a negative interaction with $T_{4\_t_{i+1}}$ shown as a red edge (Fig. 5.3e). This simple analytic approach enables us to annotate each biological relationship with directional information.

5.2.8 Network visualization

All the bootstrap networks[1] shown are visualized using *Cytoscape* [161] version 3.6.0, using Attribute Circle Layout with Organic Edge Router. An in-house script is used to generate a custom style XML file for each network, encoding multiple properties of the underlying graph. Among these properties, the regression coefficients corresponding to edge thickness were normalized as follows: Let Y be a microbial taxa node with continuous

[1]For each data set, we ran 500 bootstrap realizations and only reported edges with bootstrap support of at least 50% in the consensus DBN.

taxa parents $U_1, \cdots, U_k$ modeled by

$$F(Y \mid U_1, \cdots, U_k) \sim N(\lambda_0 + \sum_{i=1}^{k} \lambda_i \times U_i, \sigma^2), \tag{5.3}$$

where $\lambda_1, \cdots, \lambda_k$ are the corresponding regression coefficients for $U_1, \cdots, U_k$ as previously described in this section. The normalized regression coefficients $\{\lambda_i^N\}_{i=1}^k$ are defined as

$$\lambda_i^N = \frac{\lambda_i \times \bar{U}_i}{\sum_{j=1}^{k} \left|\lambda_j \times \bar{U}_j\right|}, \tag{5.4}$$

where $\bar{U}_i$ is the mean abundance of taxa U_i across all samples.

5.3 Results

Fig. 5.3 presents a schematic diagram illustrating the whole computational pipeline we developed for aligning and learning DBNs for microbiome and clinical data. We start by estimating a cubic spline from the observed abundance profile of each taxa (Fig. 5.3b), which enable principled estimation of unobserved time points and interpolation at uniform intervals. Next, we determine an alignment which allows us to directly compare temporal data across individuals (Fig. 5.3c), as well as filter out abnormal and noisy samples (Fig. 5.3d). We thenuse the aligned data to learn causal dynamic models that provide information about interactions between taxa, their impact, and the impact of clinical variables on taxa levels over time (Fig. 5.3e). Let nodes (T_1, T_2, T_3, T_4) represent microbial taxa and (C_1, C_2, C_3) represent clinical factors shown as circles and diamonds, respectively. Figure shows two consecutive time slices t_i and t_{i+1}, where dotted lines connect nodes from the same time slice referred to as *intra edges*, and solid lines connect nodes between time slices referred to as *inter edges*. Biological relationships are inferred from edge parameters in the learned DBN which can be positive (green) or negative (red). Finally, Fig. 5.3f represents the original and predicted relative abundance across four gut taxa. The prediction perfor-

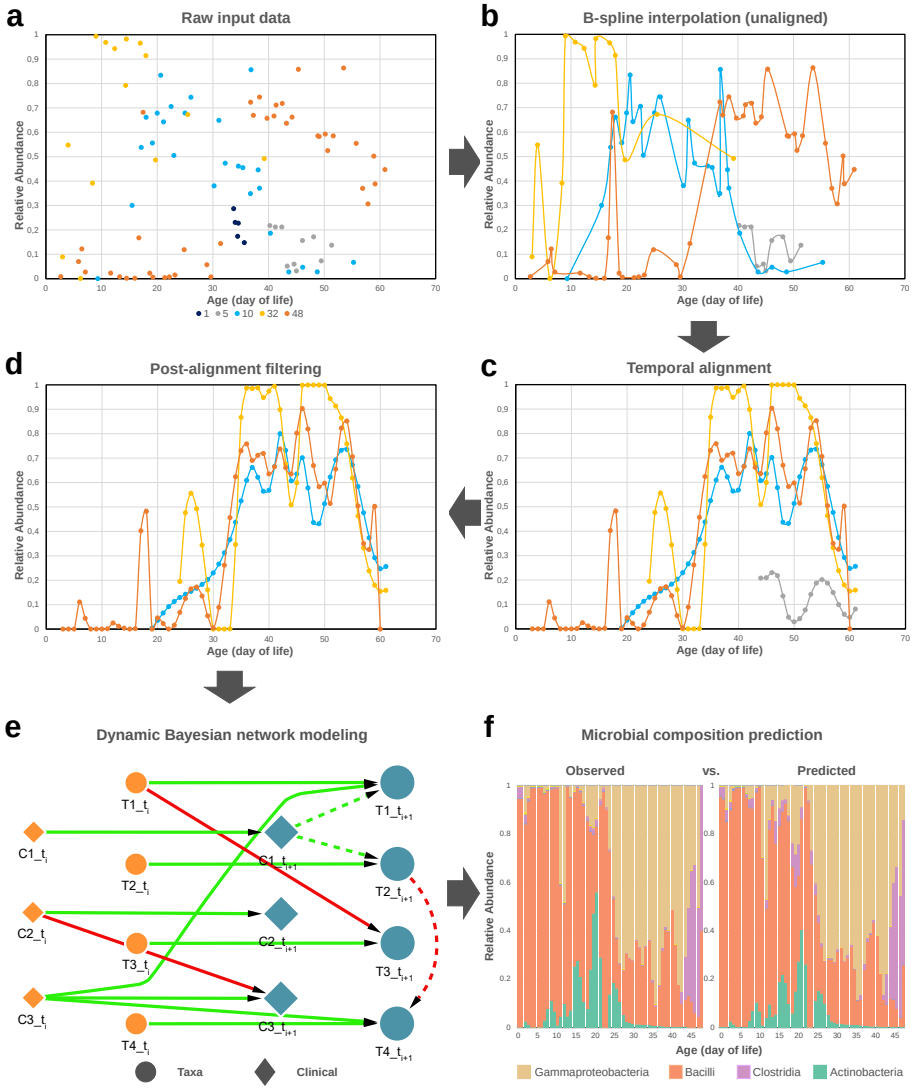

Figure 5.3: **Schematic diagram illustrating the whole computational pipeline proposed in this work**. Figure shows microbial taxa *Gammaproteobacteria* at each step in the pipeline from a set of five representative individual samples (subjects 1, 5, 10, 32 and 48) of the gut data set. **a)** Input is raw relative abundance values for each sample measured at (potentially) non-uniform intervals even within the same subject. **b)** Cubic B-spline curve for each individual sample. Sample corresponding to subject 1 (dark blue) contains less than pre-defined threshold for measured time points, thus, removed from further analysis. **c)** Temporal alignment of each individual sample against a selected reference sample (subject 48 shown in orange). **d)** Post-alignment filtering of samples with alignment error higher than a pre-defined threshold. Sample corresponding to subject 5 (grey) discarded. **e)** Learning a dynamic Bayesian network structure and parameters. **f)** Original and predicted relative abundance across four gut taxa for subject 48 at sampling rate of 1 day.

mance is evaluated by average mean absolute error (MAE) between original and predicted abundance values (MAE = 0.011).

We applied our methods to study longitudinal data sets from three human microbiome niches: infant gut, vagina and oral cavity (see Methods for full descriptions). In addition to the differences in the taxa they profile, these data sets vary in the number of subjects profiled (ranging from 15 to 48), in the number of time points they collected, the overall number of samples and time series that were studied, etc. Thus, they provide a good set to test the generality of our methods and their usefulness in different microbiome studies.

5.3.1 Temporal alignments

Below, we discuss in detail the improved accuracy of the learned dynamic models due to use of *temporal alignments*. However, even before using them for our models, we wanted to verify our splines-based heuristic alignment approach, as well as test whether the alignment results agree with biological knowledge.

Simulation experiments: To investigate whether our splines-based greedy alignment approach is able to identify good solutions, we performed several simulation experiments (described in Methods). In summary, we simulated data for 10 individual samples and aligned them against a reference sample. We next computed the alignment accuracy (MAE) between the observed and expected alignment parameters (i.e., a and b), and alignment error E_M on the simulated data. These results are shown in Figure 5.4: Figure 5.4, where the average error for alignment parameter a ranges between 0.030-0.035 at 5% noise up to 0.24-0.35 at 25% noise across all simulation experiments. Alternatively, the average error for alignment parameter b ranges between 0.25-0.30 at 5% noise up to 4.5-6.2 at 25% noise across all three experiments. Finally, the alignment error E_M is at most 7% at 25% noise which indicates large agreement between the aligned samples. Overall, these simu-

lation results provide evidence that the proposed greedy search method is able to find good alignments, thus, supporting our prior assumptions as well as the use of B-splines.

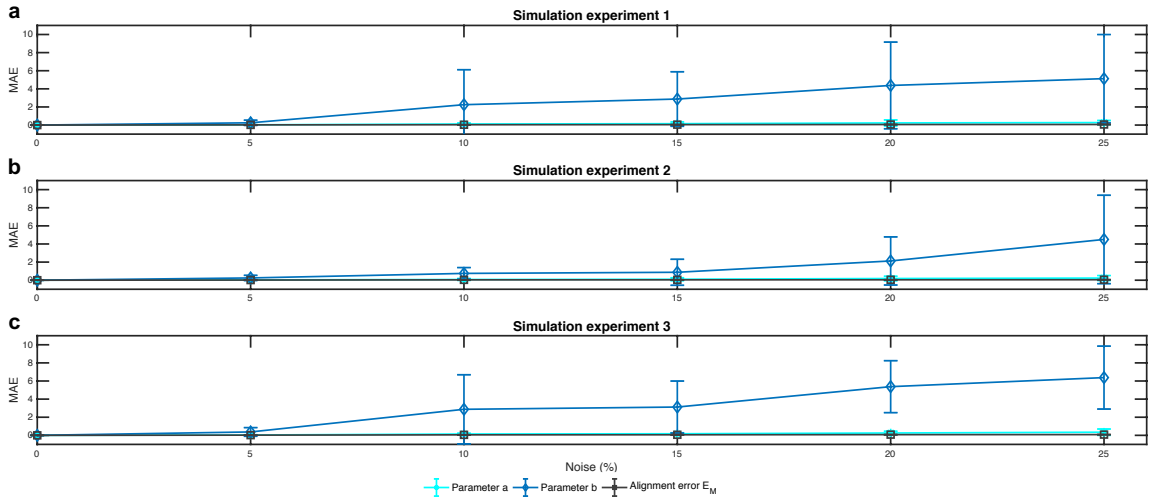

Figure 5.4: **Temporal alignment accuracy on simulated data**. Figure shows MAE alongside standard deviation for alignment parameters a and b, as well as alignment error E_M using our heuristic alignment approach as a function of percentage of Gaussian noise. **a)** Alignment performance on simulation experiment 1. **b)** Alignment performance on simulation experiment 2. **c)** Alignment performance on simulation experiment 3.

Infant gut alignments capture gestational age at birth: To test whether the alignment results agree with biological knowledge, we used the infant gut data. Infant gut microbiota goes through a patterned shift in dominance between three bacterial populations (*Bacilli* to *Gammaproteobacteria* to *Clostridia*) in the weeks immediately following birth. La Rosa *et al.* [86] reported that the rate of change is dependent on maturation of the infant highlighting the importance of post-conceptional age as opposed to day of life when analyzing bacterial composition dynamics in preterm infants. We found that our alignment method is able to capture this rate of change without explicitly using gestational or post-conceptional age.

Fig. 5.5 shows the relationship between alignment parameters a and b (from the transformation function $\tau_i(t) = \frac{(t-b)}{a}$ described in Methods) and the gestational age at birth for each infant in the gut microbiome data set. Each aligned infant sample is represented by a blue circle where the x-axis shows $\frac{-b}{a}$ and y-axis shows the gestational age at birth. As

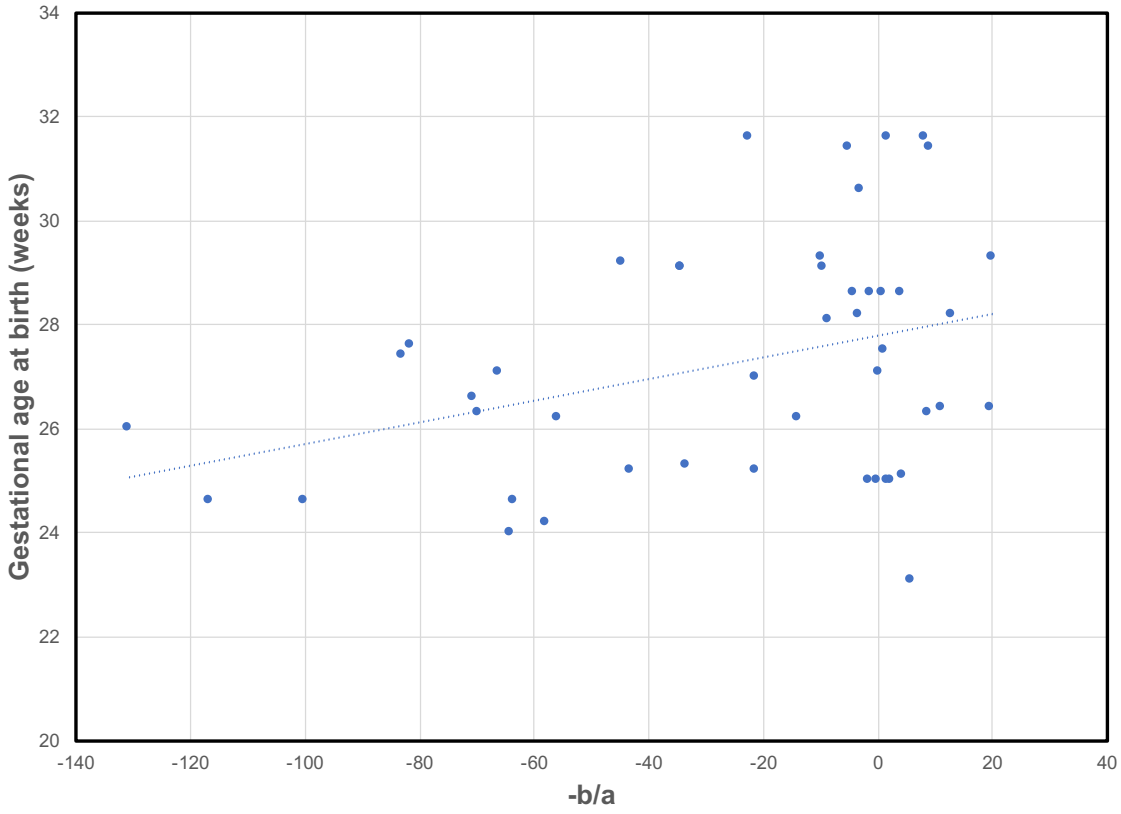

Figure 5.5: **Relationship between alignment parameters and gestational age at birth**. Figure shows the relationship between alignment parameters a and b and gestational age at birth (measured in weeks) for the aligned infant gut microbiome data set. Each blue dot represent an aligned infant sample i where x-axis shows $\frac{-b}{a}$ from transformation function $\tau_i(t) = \frac{(t-b)}{a}$ and y-axis shows the gestational age at birth of infant i. Pearson correlation coefficient = 0.35.

can be seen, the alignment parameters are reasonably well correlated with gestational age at birth (Pearson's correlation coefficient = 0.35) indicating that this method can indeed be used to infer differences in rates between individuals.

5.3.2 Resulting dynamic Bayesian network models

We next applied the full pipeline to learn DBNs from the three microbiome data sets under study. In particular, we use longitudinal data sets from three human microbiome niches: infant gut, vaginal and oral cavity as described in Methods. In this section, we highlight the overall characteristics of the learned DBN for each aligned and filtered microbiome data set (Fig. 5.6 and Figure 5.7a). By contrast, we also show the learned DBN for each

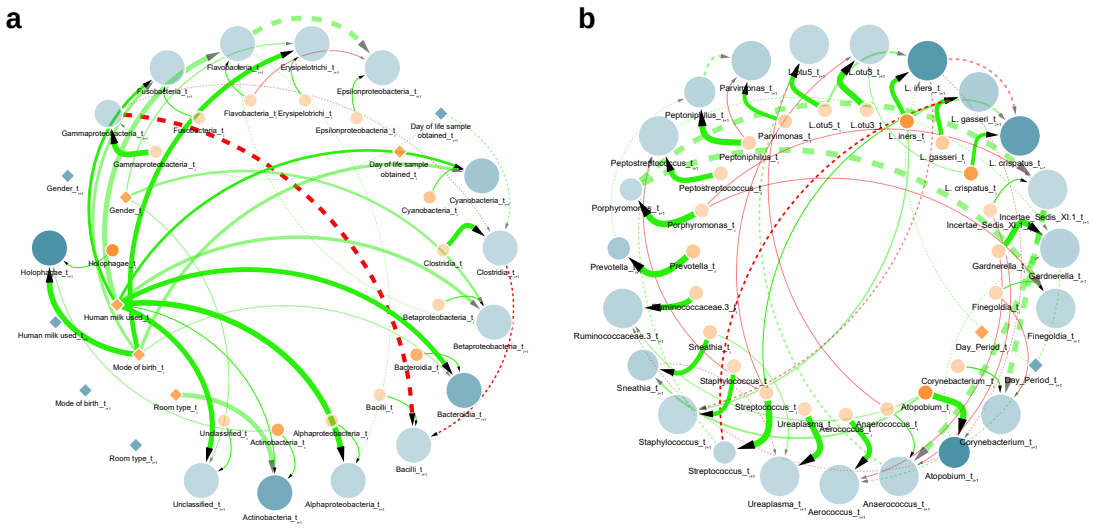

Figure 5.6: **Learned dynamic Bayesian network for infant gut and vaginal microbiomes derived from aligned samples**. Figure shows two consecutive time slices t_i (orange) and t_{i+1} (blue), where nodes are either microbial taxa (circles) or clinical/demographic factors (diamonds). Nodes size is proportional to in-degree whereas taxa nodes transparency indicates mean abundance. Additionally, dotted lines denote *intra edges* (i.e., directed links between nodes in same time slice) whereas solid lines denote *inter edges* (i.e., directed links between nodes in different time slices). Edge color indicates positive (green) or negative (red) temporal influence and edge transparency indicates strength of bootstrap support. Edge thickness indicates statistical influence of regression coefficient as described in Network visualization. **a)** Learned DBN for the aligned infant gut microbiome data at a sampling rate of 3 days and *maxParents* = 3. **b)** Learned DBN for the aligned vaginal microbiome data at a sampling rate of 3 days and *maxParents* = 3.

unaligned and filtered microbiome data set in Figure 5.7b and Figure 5.8. In all these figures the nodes represent taxa and clinical (or demographic) variables and the directed edges represent temporal relationships between them. Several triangles were also observed in the networks. In some of the triangles, directed edges to a given node were linked from both time slices of another variable. We will refer to these as *directed triangles*.

Infant gut microbiome data: The learned DBN model for the infant gut microbiome data set smoothed, aligned and sampled every 3 days with *maxParents* = 3 was computed. It contains 19 nodes per time slice (14 microbial taxa, 4 clinical and 1 demographic variable nodes) and 39 directed edges (31 inter edges and 8 intra edges) with no directed triangles

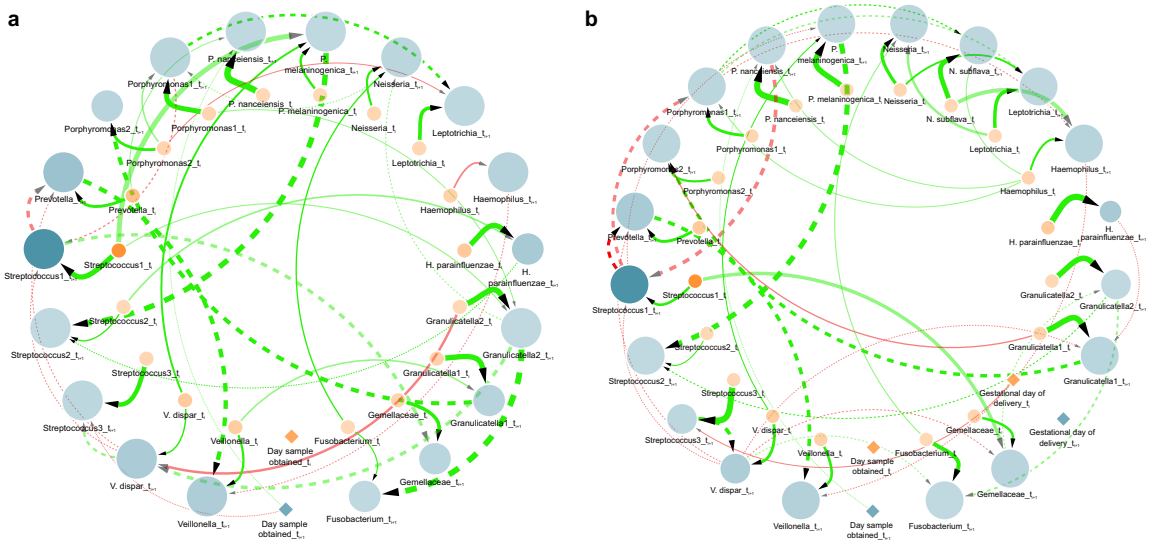

Figure 5.7: **Learned dynamic Bayesian network of the oral microbiome derived from unaligned and aligned tooth/gum samples**. Figure shows two consecutive time slices t_i (orange) and t_{i+1} (blue), where nodes are either microbial taxa (circles) or clinical factors (diamonds). Nodes size is proportional to in-degree whereas taxa nodes transparency indicates mean abundance. Additionally, dotted lines denote *intra edges* whereas solid lines denote *inter edges*. Edges color indicates positive (green) or negative (red) temporal influence and edge transparency indicates strength of bootstrap value. Edge thickness indicates statistical influence of regression coefficient as described in Network visualization. **a)** Learned DBN for the aligned oral microbiome data at a sampling rate of 7 days and *maxParents* = 3. **b)** Learned DBN for the unaligned oral microbiome data at a sampling rate of 7 days and *maxParents* = 3.

as shown in Fig. 5.6a. Since we only learn temporal conditional dependence (i.e., incoming edges) for taxa nodes at time slice $i + 1$, the maximum number of possible edges is $14 \times$ *maxParents* $= 42$, thus, most of the taxa nodes (11 out of 14) have reached the maximum number of parents allowed (i.e., *maxParents* $= 3$). Additionally, the majority of these temporal relationships are between microbial taxa. In particular, the model includes several interactions between the key colonizers of the premature infant gut: *Bacilli*, *Clostridia* and *Gammaproteobacteria*. Furthermore, the only negative interactions learned by the model comprise these microbes which are directly involved in the progression of the infant gut microbiota. Also, the nodes for gestational age at birth and post-conceptional age at birth are not shown because they are isolated from the rest of the network, without any single

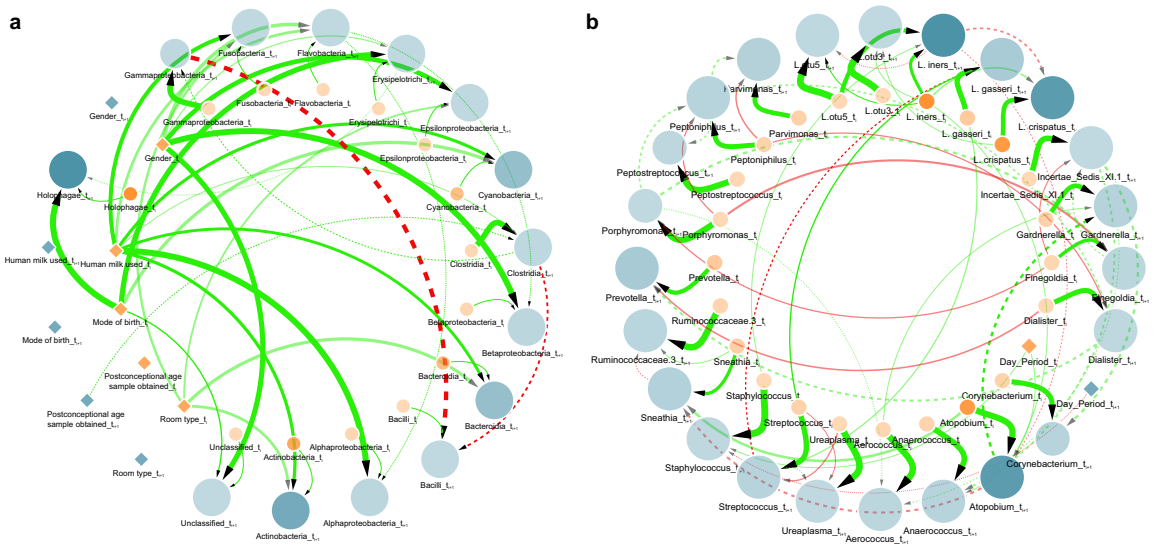

Figure 5.8: **Learned dynamic Bayesian network for gut and vaginal microbiomes derived from unaligned samples**. Figure shows two consecutive time slices t_i (orange) and t_{i+1} (blue), where nodes are either microbial taxa (circles) or clinical/demographic factors (diamonds). Nodes size is proportional to in-degree whereas taxa nodes transparency indicates mean abundance. Additionally, dotted lines denote *intra edges* (i.e., directed links between nodes in same time slice) whereas solid lines denote *inter edges* (i.e., directed links between nodes in different time slices). Edge color indicates positive (green) or negative (red) temporal influence and edge transparency indicates strength of bootstrap support. Edge thickness indicates statistical influence of regression coefficient as described in Network visualization. **a)** Learned DBN for the unaligned infant gut microbiome data at a sampling rate of 3 days and *maxParents* = 3. **b)** Learned DBN for the unaligned vaginal microbiome data at a sampling rate of 3 days and *maxParents* = 3.

edge. Overall, these trends strongly suggest that the DBN is capturing biologically relevant interactions between taxa.

Vaginal microbiome data: As with the gut microbiome data set, we learned a DBN model for the vaginal microbiome data at a sampling rate of 3 days and *maxParents* = 3 (Fig. 5.6b). The resulting DBN is comprised of 24 nodes per time instance (23 taxa and 1 clinical) and 58 edges (40 inter edges and 18 intra edges). Additionally, 12 directed triangles involving taxa nodes were observed. In preliminary analyses, additional clinical and demographic attributes (e.g., Nugent category, race and age group) resulted in networks with these variables connected to all taxa nodes, thus, removed from further analysis. Specifically, we estimated the degree of overfitting of these variables by learning and

testing DBN models with and without them. This resulted in the DBN shown in Fig. 5.6b which exhibited lowest generalization error. In this case, the maximum number of potential edges between bacterial nodes is $24 \times maxParents = 72$; however, only 16 out of 24 taxa nodes reached the threshold on the maximum number of parents. Among all the 58 edges, only one interaction *Day_Period* t_{i+1} to *L. iners* t_{i+1} involves a clinical node whereas the remaining 57 edges (including 15 negative interactions) captured temporal relationships among microbial taxa. This mixture of positive and negative interactions between taxa provides evidence of the DBNs ability to capture the complex relationships and temporal dynamics of the vaginal microbiota.

Oral microbiome data: We learned a DBN with the longitudinal tooth/gum microbiome data set with a sampling rate of 7 days and $maxParents = 3$. Figure 5.7a shows the learned DBN which contains 20 nodes for each time slice (19 taxa and 1 clinical) and 52 edges (33 inter edges and 19 intra edges) out of 57 possible edges. In addition 2 directed triangles were observed involving taxa nodes. Here, the DBN model includes multiple positive and negative interactions among early colonizers (e.g., *Veillonella* and *H. parainfluenzae*) and late colonizers (e.g., *Porphyromonas*) of the oral microbiota which are supported by previous experimental studies [85].

5.3.3 Comparisons to prior methods

To evaluate the accuracy of our pipeline and to compare them to models reconstructed by prior methods published in the literature [94, 108], we used a per-subject cross-validation with the goal of predicting microbial taxon abundances using the learned models. In each iteration, the longitudinal microbial abundance profile of a single subject was selected as the test set, and the remaining profiles were used for building the network and learning model parameters. Next, starting from the second time point, we used the learned model to predict an abundance value for every taxa in the test set at each time point using the

previous and current time points. Predicted values were normalized to represent relative abundance of each taxa across the microbial community of interest. Finally, we measured the average predictive accuracy by computing the MAE for the selected taxon in the network. We repeated this process (learning the models and predicting based on them) for several different sampling rates, which ranged from 1 up to 28 days depending on the data set. The original and predicted microbial abundance profiles can be compared as shown in Fig. 5.3f. The average MAE for predictions on the three data sets are summarized in Table 5.3. Furthermore, Fig. 4 and Figure 5.9 show violin and bar plots of the MAE distributions for ten different methods on each data set, respectively. Along with two of our DBNs (one with and one without alignments), four methods with and four without alignments were compared. These are further described below.

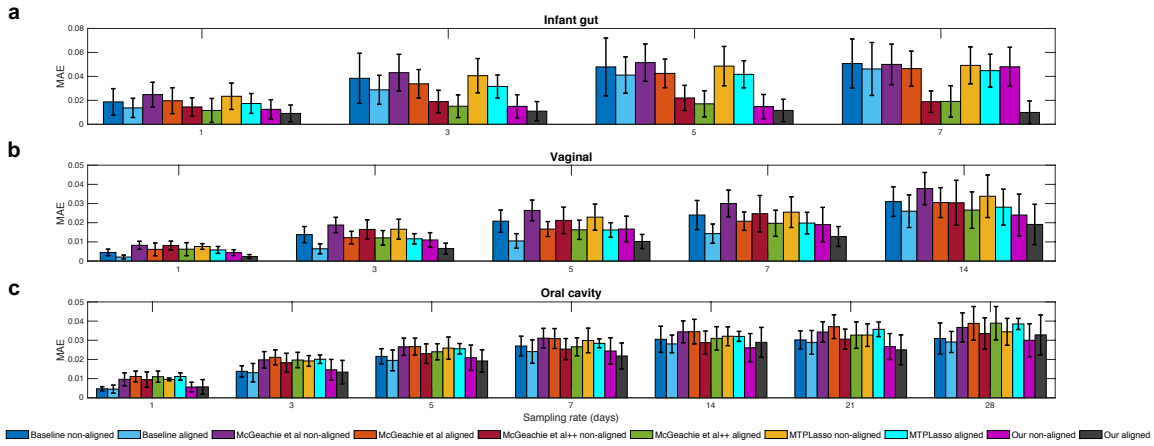

Figure 5.9: **Comparison of average predictive accuracy and standard deviation between methods on the filtered data sets**. Figure shows the average MAE and standard deviation of our proposed DBN models against a baseline method and previously published approaches as a function of sampling rates. Additionally, each method is run on the unaligned and aligned data sets. **a)** Performance results for infant gut microbiome data. **b)** Performance results for vaginal microbiome data. **c)** Performance results for oral cavity microbiome data.

First, we compared the DBN strategy to a naive (baseline) approach. This baseline approach makes the trivial prediction that the abundance value for each taxa A at any given point is exactly equal to the abundance measured at the previous time point. Given that measured abundances are continuous variables, this turns out to be an extremely compet-

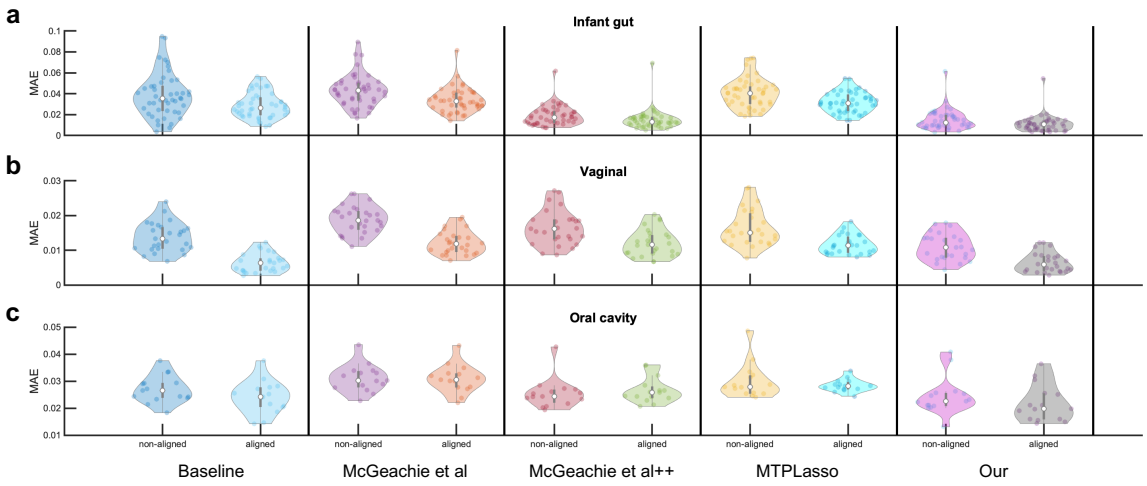

Figure 5.10: **Comparison of average predictive accuracy between methods on the filtered data sets**. Figure shows violin plots of the MAE distributions of our proposed DBN models against a baseline method and previously published approaches for a sampling rate that most closely resembles the originally measured time points. Additionally, each method is run on the non-aligned and aligned data sets. **a)** Performance results for infant gut microbiome data for sampling rate of 3 days. **b)** Performance results for vaginal microbiome data for sampling rate of 3 days. **c)** Performance results for oral cavity microbiome data for sampling rate of 7 days.

itive method and performs better than most prior methods for the data sets we tested on. Next, we compared our DBNs to three other methods suggested for modeling interactions among taxa: (a) McGeachie *et al.* [108] developed a different DBN model where network learning is estimated from the BDeu scoring metric [106] (instead of MLE), (b) McGeachie *et al.*++ an in-house implementation that extends the method of McGeachie *et al.* to allow for intra edges during structure learning, and (c) MTPLasso [94] that models time-series microbial data using a gLV model. In all cases, we used the default parameters as provided in the original publications.

As can be seen by Table 5.3 and Figure 5.9 our method outperforms the baseline and previous methods for the infant gut data. It also performs favorably when compared to baseline on the other two data sets. Temporal alignments improved the predictive performance over unaligned samples across gut and vaginal microbiomes by about 1-4 percentage points. In particular, a two-tailed t-test indicates significant (denoted by *) performance improvements for most sampling rates (*infant gut*: p-value = 0.043* for 1 day,

p-value = 0.034* for 3 days, p-value = 0.109 for 5 days, and p-value < 1.00E-05* for 7 days; *vaginal*: p-value < 1.00E-06* for 1 day, p-value < 1.00E-05* for 3 days, p-value = 5.50E-05* for 5 days, p-value = 3.10E-03* for 7 days, and p-value = 0.097 for 14 days). On the other hand, alignments did not show significant predictive performance improvements on the oral data set and is consistent with previous analysis on the same data set [45]. Surprisingly, the simple baseline approach outperforms all previously published methods: McGeachie *et al.* [108] and MTPLasso [94] across the three data sets. Finally, Fig. 5.10 shows violin plots of the MAE results for each data set across a sampling rate that most closely resembles the originally measured time points.

Data set	*sr*	**Baseline**		**McGeachie *et al.***		**McGeachie *et al.*++**		**MTPLasso**		**Our**	
	(days)	non-aligned	aligned	non-aligned	aligned	non-aligned	aligned	non-aligned	aligned	non-aligned	aligned
Infant gut	1	1.87 ± 1.11	1.37 ± 0.80	2.48 ± 1.03	1.97 ± 1.08	1.45 ± 0.76	1.16 ± 1.00	2.34 ± 1.10	1.74 ± 0.83	1.25 ± 0.81	**0.91 ± 0.70**
	3	3.84 ± 2.09	2.88 ± 1.21	4.31 ± 1.53	3.38 ± 1.19	1.90 ± 0.94	1.51 ± 0.95	4.06 ± 1.43	3.16 ± 0.96	1.50 ± 0.97	**1.10 ± 0.80**
	5	4.79 ± 2.41	4.11 ± 1.52	5.15 ± 1.56	4.25 ± 1.19	2.20 ± 1.05	1.71 ± 1.09	4.86 ± 1.64	4.17 ± 1.13	1.48 ± 1.01	**1.16 ± 0.93**
	7	5.07 ± 2.05	4.62 ± 2.21	5.00 ± 1.69	4.65 ± 1.46	1.89 ± 0.90	1.91 ± 1.31	4.92 ± 1.54	4.48 ± 1.37	4.80 ± 1.63	**0.99 ± 0.96**
Vaginal	1	0.45 ± 0.17	**0.21 ± 0.10**	0.82 ± 0.21	0.61 ± 0.33	0.81 ± 0.23	0.62 ± 0.34	0.76 ± 0.15	0.58 ± 0.18	0.44 ± 0.15	0.24 ± 0.10
	3	1.38 ± 0.42	**0.64 ± 0.26**	1.88 ± 0.40	1.22 ± 0.33	1.65 ± 0.50	1.21 ± 0.38	1.66 ± 0.52	1.16 ± 0.27	1.10 ± 0.38	0.65 ± 0.28
	5	2.08 ± 0.58	1.05 ± 0.38	2.64 ± 0.54	1.67 ± 0.39	2.12 ± 0.69	1.63 ± 0.50	2.29 ± 0.68	1.62 ± 0.38	1.67 ± 0.67	**1.02 ± 0.37**
	7	2.40 ± 0.76	1.43 ± 0.50	3.00 ± 0.70	2.08 ± 0.48	2.47 ± 0.95	1.97 ± 0.68	2.55 ± 0.80	1.98 ± 0.56	1.90 ± 0.90	**1.28 ± 0.52**
	14	3.10 ± 0.77	2.60 ± 0.85	3.78 ± 0.84	3.05 ± 0.78	3.04 ± 1.17	2.66 ± 0.95	3.38 ± 1.11	2.81 ± 0.94	2.40 ± 1.09	**1.91 ± 1.05**
Oral cavity	1	0.47 ± 0.11	**0.45 ± 0.21**	0.96 ± 0.34	1.11 ± 0.28	0.95 ± 0.40	1.10 ± 0.29	0.96 ± 0.07	1.11 ± 0.19	0.56 ± 0.25	0.57 ± 0.38
	3	1.38 ± 0.29	**1.31 ± 0.48**	1.98 ± 0.43	2.11 ± 0.39	1.83 ± 0.49	1.97 ± 0.40	1.92 ± 0.28	2.01 ± 0.22	1.46 ± 0.54	1.34 ± 0.61
	5	2.16 ± 0.40	1.95 ± 0.54	2.66 ± 0.45	2.67 ± 0.44	2.30 ± 0.51	2.40 ± 0.42	2.59 ± 0.58	2.56 ± 0.27	2.09 ± 0.66	**1.92 ± 0.58**
	7	2.70 ± 0.51	2.41 ± 0.61	3.11 ± 0.51	3.09 ± 0.52	2.54 ± 0.55	2.66 ± 0.47	2.99 ± 0.64	2.84 ± 0.23	2.44 ± 0.69	**2.18 ± 0.68**
	14	3.05 ± 0.68	2.81 ± 0.46	3.44 ± 0.57	3.45 ± 0.64	2.88 ± 0.60	3.10 ± 0.61	3.21 ± 0.49	3.20 ± 0.26	**2.61 ± 0.74**	2.89 ± 0.78
	21	3.02 ± 0.47	2.89 ± 0.62	3.43 ± 0.53	3.71 ± 0.62	3.06 ± 0.52	3.27 ± 0.69	3.27 ± 0.59	3.57 ± 0.38	2.67 ± 0.68	**2.50 ± 0.78**
	28	3.09 ± 0.81	**2.91 ± 0.55**	3.67 ± 0.76	3.88 ± 0.88	3.35 ± 0.82	3.89 ± 0.87	3.44 ± 0.70	3.85 ± 0.29	3.00 ± 0.86	3.28 ± 1.04

Table 5.3: **Summary of average predictive accuracy and standard deviation between methods on the filtered data sets**. For each data set, we list the average MAE and standard deviation (presented as percentage) of our proposed DBN models against a baseline method and previously published approaches across different sampling rates. Additionally, each method is run on the non-aligned and aligned data sets. The highest predictive accuracy for each sampling rate is shown in boldface.

5.3.4 Anomaly detection using alignment

When analyzing large cohorts of microbiome data, it is important to implement a strategy to remove outliers as these can affect our ability to generalize from the collected data. As discussed in Methods, we can use our alignment error E_M score to identify such subjects and remove them prior to modeling. In the context of the gut data set, this resulted in the identification of two infant samples: Subjects 5 and 55 (highlighted in red within Figure 5.11a) which are likely processing errors, contaminated samples, or just natural anoma-

lies. Sample 55 has been previously identified as a likely abruption event by McGeachie *et al.* [108] using a different approach. Similarly, Figure 5.11b shows the distribution of alignment errors E_M for the vaginal microbiome data. In this case, we remove 6 sub-samples from 4 different women (highlighted in red). We note that there were no outliers identified in the oral cavity microbiome data set. When learning DBNs following the filtering we obtain even better models. Figure 5.12 compares the average MAE results of our proposed DBN model between the unfiltered and filtered samples for the gut and vaginal data sets. As can be seen, a large performance improvement is observed for the gut data while a slight improvement is observed for the vaginal data when removing the outliers. These results suggest that even though the method uses less data to learn the models, the models that it does learn are more accurate.

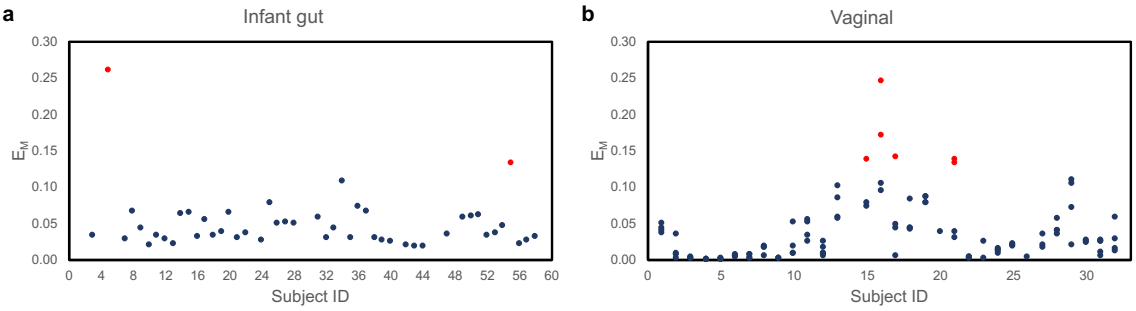

Figure 5.11: **Distribution of microbiome alignment error E_M for infant gut and vaginal data sets**. **a)** E_M scores for 47 infant gut samples aligned against a common reference gut sample. **b)** E_M scores for 112 vaginal microbiome sub-samples aligned against an optimal reference sub-sample. In both panels, the scores highlighted in red represent samples with E_M at least two standard deviations away from the mean of the distribution of microbiome alignment errors, thus, identified as outliers and removed.

5.4 Discussion

5.4.1 The power of temporal alignments

We developed a pipeline for the analysis of longitudinal microbiome data and applied it to three data sets profiling different human body parts. To evaluate the reconstructed networks we used them to predict changes in taxa abundance over time. Interestingly, ours

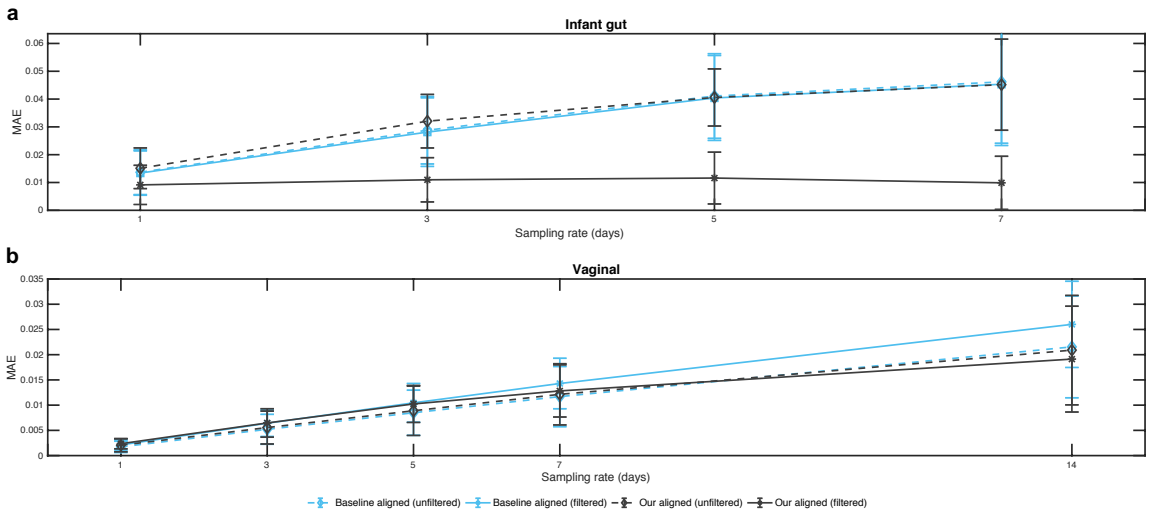

Figure 5.12: **Effect of outliers on average predictive accuracy from aligned data sets**. Figure shows the average MAE for our proposed DBN model and baseline method as a function of sampling rates before (labeled as *unfiltered*) and after (labeled as *filtered*) removal of outliers. **a)** Performance results for infant gut microbiome data. **b)** Performance results for vaginal microbiome data.

is the first method to improve upon a naive baseline (Figure 5.9). While this does not fully validate the accuracy of the models, it does mean that the additional interactions determined by our method contribute to the ability to infer future changes and so at least some are likely true.

As part of our pipeline we perform temporal alignment. While ground truth for alignments is usually hard to determine, in one of the data sets we analyzed we could compare the alignment results to external information to test its usefulness. In the context of the infant gut data, it has been shown that using day of life as the independent variable hinders the identification of associations between bacterial composition and day of sampling. Therefore, previous work have re-analyzed the premature gut microbiota with post-conceptional age, uncovering biologically relevant relationships [86]. By using alignment we were able to correct for this difference without the need to rely on the external age information. In addition to the results presented in Fig. 5.5, the learned DBN in Fig. 5.6a does not show any relationships to post-conceptional age or gestational age at birth indicating that our alignment was able to successfully compensate for. By contrast, the learned DBN from

unaligned samples in Figure 5.8: Figure 5.8 shows relationships to post-conceptional age. While for this data such correction could have been made using post-conceptional age, in other cases the reason for the rate change may not be obvious and without alignment it would be hard to account for such hidden effects.

5.4.2 Uncovering biological relationships

We next discuss in more detail the learned DBN models.

Infant gut: As mentioned in Results, the only negative relationships identified supports the known colonization order, that is, a shift in dominance from *Bacilli* to *Gammaproteobacteria* to *Clostridia*) [86], as the infant goes through the first several weeks of life. These edges show incoming negative relationships to *Bacilli* from *Gammaproteobacteria* and *Clostridia*. In particular, an increase in the abundance of the parents is associated with a decrease in the abundance of the child. The negative edge from *Gammaproteobacteria* to *Clostridia* agrees with previous findings where *Clostridia*'s abundance is found to increase at a gradual rate until it peaks at post-conceptional age between 33 and 36 weeks whereas *Gammaproteobacteria* decreases as infants age [86, 108]. It is important to note that this negative edge from *Gammaproteobacteria* to *Clostridia* is not found in the learned DBN from unaligned samples (Figure 5.8a). This relationship is also confirmed by the edges from *Day of life* to *Gammaproteobacteria* and *Clostridia* (Fig. 5.6b). Moreover, the DBN model indicates a relationship between breastfeeding and *Actinobacteria*, *Bacteroidia*, and *Alphaproteobacteria*. These bacteria are known to be present in breast milk which is known to heavily influence and shape the infant gut microbiome [79].

Vaginal: It has been established that microbial composition can change dramatically during the menses cycle and later return to a 'stable' state before the next menstrual period [67, 136]. Previous studies have identified a subset of individuals in this data set as exhibiting a microbial composition dominated by *L. crispatus* with a notable increase of *L.*

iners around the start of each menstrual period [55, 67] (Figure 5.1a). These interactions were also captured by the learned DBN model in the form of a directed triangle involving *L. crispatus* and *L. iners* (Fig. 5.6b). The edge from the *Day Period* to *L. iners* strengthens this relationship, which is not present in the learned DBN from unaligned vaginal sub-samples (Figure 5.8b). On the other hand, subjects from another group were characterized as dominated by *L. gasseri* coupled with shifts to *Streptococcus* during menstruation [55]. These relationships were also captured by the DBN. Furthermore, while *L. iners* has a lower protective value than the other *Lactobacillus* [132], the negative edge between *L. iners* and *Atopobium* suggests a relationship related to environment protection. Also, the positive edge from *Atopobium* to *Gardnerella* is supported by the synergy observed between these two taxa in bacterial vaginosis [63]. Although many of these microbial relationships are also observed in the learned DBN from unaligned sub-samples, there are some biological relationships which cannot be found within the DBN derived without alignments. However, given our limited understanding of the interactions within the vaginal microbiome, we cannot determine whether or not these previously unseen interactions are biologically relevant. Finally, it is worth highlighting that the shifts and composition of the vaginal microbiome vary considerably between each women [136, 55].

Oral: For oral microbiomes, several *Streptococcus* species, including *S. oralis*, *S. mitis*, *S. gordonii*, and *S. sanguis* are well known as early colonizers lying close to the tooth pellicle [85]. While our learned DBNs (Figure 5.7) cannot identify specific species, it suggests interactions between some species of *Streptococcus* and other later colonizers in the oral microbiome such as *Porphyromonas* and *Prevotella*. The learned DBN derived from aligned tooth/gum samples also provided novel predictions, for example taxa *Granulicatella* is interacting with *Veilonella*. Furthermore, there are other microbial relationships uniquely observed on each DBN which are also potentially interesting.

5.4.3 Triangles in DBNs

An interesting aspect shared by all of the DBNs discussed above is the fact that they contain triangles or feed-forward loops. In particular many of these directed triangles are created from nodes representing both time slices of another variable, but with different signs (one positive and the other negative). For example, microbial taxa *L. crispatus* displays a directed triangle with another taxa *L. iners* in the vaginal DBN (Fig. 5.6b). In this triangle, positive edges from *L. iners*$\_t_i$ interact with *L. iners*$\_t_{i+1}$ and *L. crispatus*$\_t_{i+1}$ whereas a negative edge connects *L. iners*$\_t_{i+1}$ to *L. crispatus*$\_t_{i+1}$.

The triangles in the DBNs represent a relationship where the abundance of a child node cannot be solely determined from the abundance of a parent at one time slice. Instead, information from both the previous and the current time slices is needed. This can be interpreted as implying that the child node is associated with the *change* of the abundance values of the parents rather than with the absolute values which each node represents.

5.4.4 Limitation and future work

While our pipeline of alignment followed by DBN learning successfully reconstructed models for the data sets we looked at, it is important to understand the limitations of the approach. First, given the complexity of aligning a large number of individuals, our alignment method is based on a greedy algorithm, thus, it is not guaranteed to optimize our objective function, i.e., total warping. Even if the alignment procedure is successful, the DBN may not be able to reflect the correct interactions between taxa. Issues related to sampling rates can impact the accuracy of the DBN (missing important intermediate interactions). On the other hand, if not enough data is available the model can overfit and predict non-existent interactions.

Given these limitations we would attempt to improve the alignment method and its guarantees in future work. We are also interested in studying the ability of our procedure to integrate additional molecular longitudinal information including gene expression and

metabolomics data which some studies are now collecting in addition to the taxa abundance data [72]. As we show in the next chapter, our approach for integrating information across individuals in order to learn dynamic models is useful for analyzing longitudinal multi-omics data from microbiome studies.

5.5 Conclusion

In this paper, we propose a novel approach to the analysis of longitudinal microbiome data sets using dynamic Bayesian networks with the goal of eliciting temporal relationships between various taxonomic entities and other clinical factors describing the microbiome. The novelty of our approach lies in the use of temporal alignments to normalize the differences in the pace of biological processes inherent within different subjects. Additionally, the alignment algorithm can be used to filter out abrupt events or noisy samples. Our results show that microbiome alignments improve predictive performance over previous methods and enhance our ability to infer known and potentially novel biological and environmental relationships between the various entities of a microbiome and the other clinical and demographic factors that describe the microbiome.

CHAPTER 6

LONGITUDINAL MULTI-OMIC NETWORK INFERENCE

6.1 Introduction

Microbiomes are communities of microbes inhabiting an environmental niche. The study of microbial communities offers a powerful approach for inferring their impact on the host environment, and their role in specific diseases and health. *Metagenomics* involves analyzing sequenced reads from the whole metagenome in a microbial community in order to determine a detailed profile of microbial taxa [139]. More recently, additional types of biological data are being profiled in microbiome studies, including *metatranscriptomics*, which involves surveying the complete metatranscriptome of the microbial community [12], *metabolomics*, which involves profiling the entire set of small molecules (metabolites) present in the microbiome's environmental niche [182], and *host transcriptomics*, which provides information about the levels of genes expressed in the host [23].

The goal of the second phase of the Human Microbiome Project (HMP) [183], called the integrative Human Microbiome Project [71], is to generate longitudinal multi-omics data sets as a means to study the dynamics of the microbiome and the host across select diseases, including preterm births, type 2 diabetes, and irritable bowel disorders.

A major challenge in microbiome data analysis is the integration of multi-omics data sets [124]. Most multi-omic studies focus on a separate analysis of each omics data set without building a unified model [14]. There have been some attempts [202, 99, 201, 203, 92] and tools to facilitate the analysis [16, 150], but there is still much room for improvement regarding reproducibility, flexibility, and biological validity [124, 22, 184]

Deep learning approaches for integrating multi-omics [98, 111] have also been developed, but their lack of interpretability prevents these models from providing insights into the interplay of the different omics entities. Even Partial Least Squares models have been used to facilitate this integration [49], but they have their own set of limitations depending

on the underlying data generation model, and are prone to provide spurious results when applied to high-dimensional data [145].

In addition, microbiomes are inherently dynamic, and so to fully understand the complex interactions that take place within these communities, longitudinal microbiome data appears to be critical [57]. Many attempts have been made to analyze data from longitudinal studies [86, 92, 203]; however, these approaches do not attempt to study interactions between taxa. An alternative approach involves the use of dynamical systems such as the generalized Lotka-Volterra (gLV) models [172, 58], however the large set of parameters in these models diminishes their utility for probabilistic inference.

Previously, we have shown that probabilistic graphical models, specifically dynamic Bayesian networks (DBNs), can be used to study metagenomic sequence data from microbiome studies leading to models that can accurately predict future changes as well as identify interactions within the microbiome [96]. However, these prior methods were only able to analyze a single omic data set. Here we present a new Pipeline for the Analysis of Longitudinal Multi-omics data (PALM), which, in addition to modeling metagenomics interactions can also incorporate time series metatranscriptomics, metabolomics and host expression data to learn an integrated model of microbiome-host interactions.

PALM overcomes a number of challenges associated with such large scale integration. First, modeling such data leads to a sizable increase in the size of the model and the number of parameters in the DBN, which grows as the product of the number of entities in each omics data set. Additionally, such large number of nodes and parameters can lead to overfitting. PALM overcomes these challenge by restricting the set of allowable interactions (edges) between the omics entities based on sound biological assumptions and by relying on continuous representation and alignment to integrate a large set of observations when learning a specific model.

An additional challenge with modeling microbiomes is the difficulty of validating the model's predictions. To address this, PALM uses *in silico* approaches employing multi-

ple public databases (genomic sequence database and metabolic pathway database) and recently proposed software tools for the validation.

Applying PALM to Inflammatory Bowel Disease (IBD) data led to models that correctly predict microbiome levels and identifies known and novel interactions. Statistical validations indicate that PALM can accurately recover known interactions and improves upon prior approaches. We further experimentally validated a few of the high scoring metabolite-taxa interactions predicted by the model.

6.2 Materials and Methods

Below we describe the computational pipeline, PALM, developed to integrate and model the interactions between the different types of omics.

6.2.1 Data

To test PALM's proposed analysis pipeline, which combines temporal alignment with Bayesian network learning and inference for multi-omics microbiome data, we used the Inflammatory Bowel Disease (IBD) cohort from a study that included 132 individuals across five clinical centers [92]. During a period of one year, each subject was profiled (biopsies, blood draws, and stool samples) every two weeks on average. This yielded temporal profiles for metagenomes, metatranscriptomes, proteomes, metabolomes and viromes across all subjects. Additionally, for each subject, host- and microbe-targeted human RNA sequencing was yielded from biopsies collected at initial screening colonoscopy sampled at two locations (ileum and rectum). All data are fully described and available at `https://ibdmdb.org/`.

Each data set was associated with the week when it was sampled. Because only a single biopsy was obtained for each location sampled (ileum and rectum), host transcriptomics were used as static variables in the DBN.

6.2.2 Data pre-processing

The different data types were processed separately. First, the taxon, metabolite, and gene abundances were normalized to make each type separately add up to 1 for each subject. Then metabolites and genes were scaled to match the mean of the taxa because of the very distinct number of entities in each category. Additionally, RNA-seq data from ileum and rectum of each host was analyzed using DESeq2 [95] and TPM values. Note that we did not explicitly address the compositional nature of the taxa with projections like PhILR [164], since they are not a normalization scheme. These approaches transform the data into another non-compositional space, with different variables (balances) rather than taxa. As a proof-of-concept, for each body site, we selected the top 20 genes with the highest variance across all subjects. We note that this set of genes is limited as previous studies have reported over 1000 differently expressed genes for IBD individuals at these two sampled locations compared to individuals without IBD [92]. Metabolites without an HMDB correspondence were removed. Next, we filtered out metabolites for which the mean intensity was less than 0.1%, or had zero variance over the originally sampled time points. Next, we performed temporal alignment of time series data from individuals as described in Lugo-Martinez *et al.* [96]. For this, we need to represent each discrete time series using a continuous function. Here we used B-splines for fitting continuous curves to the time-series multi-omic data profiled from each subject, including the microbial composition, gene expression, and metabolic abundance. To improve the accuracy of the reconstructed profiles, we removed any sample that had less than five measured time points in any of the multi-omics measurements. This led to the removal of 70 individuals from the cohort resulting in 62 individual multi-omic time series that were used for further analysis.

6.2.3 Temporal alignments

Given longitudinal samples from different subjects, we cannot expect that the rates at which various multi-omics levels change would be exactly the same between these in-

dividuals [9]. To facilitate the analysis of such longitudinal data across subjects, we first align the time series from the microbiome samples using the microbial composition profiles (see Section 3.6). As described earlier, these alignments use a linear time transformation function to warp one time series into a common, representative sample time series used as the reference [96]. While prior alignment methods relied on taxa information, when multi-omics data is available, PALM can use other genomic information for the alignment. Specifically, here we also tested the use of gene expression and metabolite abundance profiles for determining accurate alignments of patients. As we show, by using a better omics data type the resulting DBNs can more accurately capture and predict taxa-metabolite and taxa-gene relationships.

For each omics data (i.e., taxa, genes, or metabolites) we select an optimal reference sample from the 107 time series as follows: we generated all possible pairwise alignments between them and selected the time series that resulted in the least total overall error in the alignments. We then filtered out abnormal and noisy samples from the resulting set of alignments as follows: (1) computed the mean μ and standard deviation δ of the alignment error, and (2) removed all samples from an individual where alignment error exceeded $\mu + (2 \times \delta)$, as previously described in [96]. Figure 6.1(a)-(d) shows the overall alignment process of *Bacteroides dorei*, from the taxa-based alignment perspective.

Given an individual's warped/aligned time series over a specific omic type, the other multi-omics data were incorporated as follows: the same transformation applied to the aligned sample was applied to all the complementary multi-omics time series data. The resulting set used for the modeling comprised of 60 individual-wise heterogeneous alignments (after filtering out high alignment error individuals) involving 102 microbial taxa, 72 genes, and 70 metabolites. This smaller number of attributes was used because learning a Bayesian network is NP-Hard [33, 40], and henceforth has an exponential run time with the number of features.

6.2.4 Dynamic Bayesian network models

Using the aligned time series multi-omics data, we next learned graphical models that provide information about the relationships between the different omics (taxa, genes, metabolites, host-genes) and environmental (exogenous) variables. In PALM, we extend the DBN model proposed in Lugo-Martinez *et al.* [96] to account for multi-omics microbiome data with the goal of inferring the temporal relationships between the heterogeneous entities in a microbial community. A DBN is a directed acyclic graph where, at each time slice, nodes correspond to random variables of interest (e.g., taxa abundance, gene expression, age, etc.), and directed edges correspond to their conditional dependencies in the graph. These edges are defined as either: *intra edges* connecting nodes from the same time slice, or *inter edges* connecting nodes between consecutive time slices. In our DBN model, only two slices are modeled and learned, as shown in Figure 6.1(e). In PALM, our DBN models encode five types of nodes: (i) taxon abundance, (ii) gene expression, (iii) metabolite concentration, (iv) host gene expression, and (v) sample metadata information. The first three types represent continuous variables, whereas the last two types can be either discrete or continuous. For our DBNs, we use the formalism of conditional Gaussian Bayesian networks [106] to take advantage of its ability to seamlessly integrate discrete and continuous variables in a single probabilistic framework. For more information about the mathematical formulation and our contributions, please refer to Chapter 4.

As highlighted in Figure 6.1(f), the conditional linear Gaussian density function for variable $T_1^{t_{i+1}}$ is modeled by

$$f(T_1^{t_{i+1}} \mid T_1^{t_i}, M_1^{t_i}, E_1^{t_{i+1}}, E_2^{t_{i+1}}) = N(\beta_0 + \beta_1 \times T_1^{t_i} + \beta_2 \times M_1^{t_i} + \beta_3 \times E_1^{t_{i+1}} + \beta_4 \times E_2^{t_{i+1}}, \sigma^2), \quad (6.1)$$

where $\Theta = \{\beta_1, \beta_2, \beta_3, \sigma^2\}$ are the DBN model parameters.

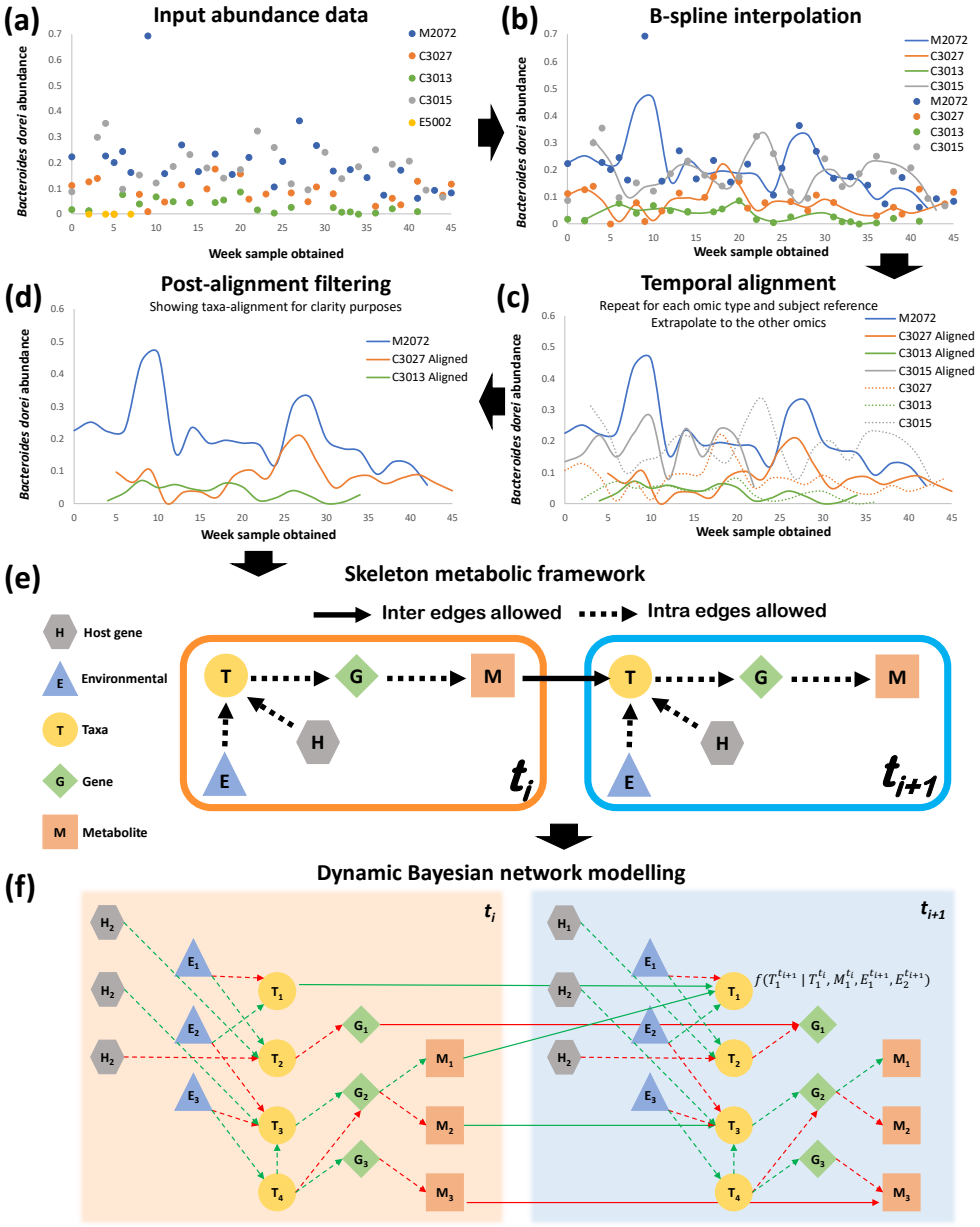

Figure 6.1: **Computational pipeline proposed in this chapter**. Figure shows microbial taxa *Bacteroides dorei* at each step in the pipeline from a set of five individual samples (subjects M2072, C3027, C3013, C3015, and E5002) of the IBD data set. **(a)** Relative abundance for each sample. **(b)** Cubic B-spline curve for each individual sample. **(c)** Temporal alignment of all taxa of each individual to correct for the different progression rates. **(d)** Post-alignment filtering of samples with a higher alignment error than a pre-defined threshold. **(e)** Biologically-inspired *Skeleton* constraints imposed on learning the DBNs computed by PALM. **(f)** Learning a two-stage DBN structure and parameters for the host genes, environmental variables, taxa, genes, and metabolites.

6.2.5 Constraining the DBN structure

An important innovation in PALM lies in the structure constraining of the network to conform to our proposed metabolic framework that ensures the desired flow of interactions.

These constraints (in the form of a matrix received as an input to the function) only allow edges between certain types of nodes, highly reducing the complexity of searching over possible structures and preventing over-fitting. Note that these constraints can be easily changed by adding more data types, or different restrictions in the input file containing the adjacency matrix. Specifically, we allowed intra edges from environmental and host transcriptomics variables to microbial taxa (abundance) nodes, from taxa nodes to gene (expression) nodes and from gene nodes to metabolites (concentration) nodes. All other interactions within a time point (for example, direct gene to taxa) were disallowed. We also allowed inter edges from metabolites to taxa nodes in the next time point, and *self-loops* from any node X_i^t to X_i^{t+1} for each i and time t, except for environmental or host transcriptomics variables for which no incoming edges were allowed (host genes were only measured at a single time point so no incoming temporal edges were allowed for them). These restrictions referred to as the *Skeleton* and depicted in 6.3(a) reflect our understanding of the basic ways the different entities interact with each other, i.e., environmental and host gene expression variables are independent variables, taxa express genes, which are involved in metabolic pathways; finally, the metabolites impact the growth of taxa (in the next time slice).

We also learned DBNs using a less constrained framework referred to as *Augmented* as shown in 6.3(b). Unlike Skeleton, the Augmented framework also allowed direct edges between taxa and metabolites to account for cases where noise or other issues related to profiling of genes can limit our ability to indirectly connect taxa and the metabolites they produce. 6.3 summarizes each framework in the form of an adjacency matrix.

Note that other constraints such as requiring that taxa could only connect to genes present in their genome were not imposed since genomics reference databases are not always complete and so they may lead to missing key interactions.

We used a greedy hill-climbing approach for structure learning where the search is initialized with a network that connects each node of interest at the previous time point to

the corresponding node at the following time point. Next, nodes are added as parents of a specific node via intra or inter edges depending on which valid edge leads to the largest increase of the log-likelihood function beyond the global penalty incurred by adding the parameters as measured by the BIC score approximation.

Every network was bootstrapped by randomly selecting with replacement as many subjects as in the data set, and learning a different network 100 times. Although we explore multiple values as the maximum number of possible parents for each node (see Figure 6.2), unless otherwise stated, the maximum number of possible parents was fixed to 3. The networks were then combined, and the regression coefficient of the edges was averaged. Each edge was also labeled with the bootstrap support (percentage of times that edge appears). Each repetition was set to run independently on a separate processor using Matlab's Parallel Computing Toolbox. Other parallel implementations include parallelizing the cross-validation computation of the inference error and each independent alignment error calculation using Python's Parallel library.

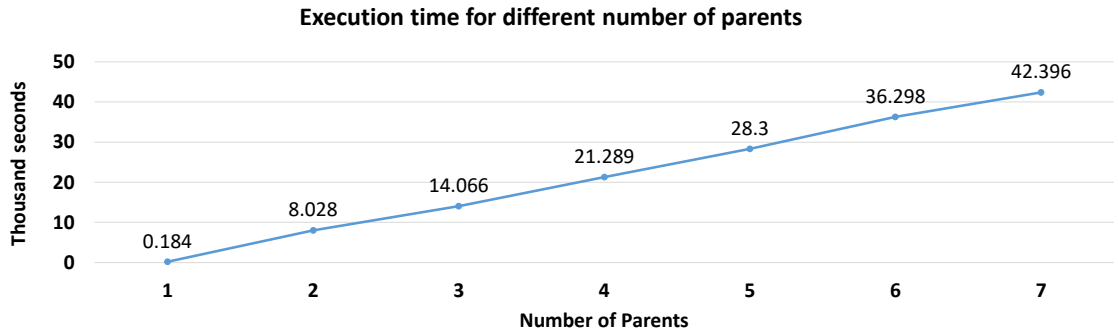

Figure 6.2: **Execution time for different maximum number of parents**. The figure shows experimentally that the execution time grows linearly with the number of parents. Note that while the times shown are for 1 repetition, the DBN figures shown are with 100.

6.2.6 Validating DBNs

A major challenge in building models of biological interactions lies in developing methods to validate them and in providing confidence measures. Since DBNs are generative models, one approach is to predict time series using previous time points and thus to achieve

cross validation [96]. Such technical validations, while informative, could be thought as of "black-box" validation, and do not shed light on the accuracy of specific edges and interactions predicted by the model that we are interested in.

We broadly discuss approaches to validate the types of edges present in the DBN (see Figure 6.1(e)), which are the parameters learned by the model, and hence closer to "white-box" validation. Edges from taxa to genes can be circumstantially validated by verifying that (a) the taxon presence is guaranteed by its non-zero abundance, (b) the taxon genome has the gene, and (c) the gene is expressed. PALM, therefore, handles this using the *in silico* validation strategies mentioned below in Section 6.2.7. Similarly, edges from genes to metabolites or taxa to metabolites could potentially be validated.

The challenge is in validating edges from metabolites to taxa, for which an *in silico* approach is unlikely to work since no such database has been compiled to the best of our knowledge. In Section 6.2.8, we propose a validation approach involving laboratory experiments.

6.2.7 *In silico* validation of DBN edges

In silico validations of DBN edges are handled by verifying the information against a database of known interactions between taxa to genes and/or taxa to metabolites. Unfortunately, no such comprehensive database exists. For example, highly curated databases such as HMDB [199], MetaCyc [81], or the findings of the large scale study of Maier et al. (2019) [100] turned out to be inadequate since the intersection of their contents with the species and metabolites in our networks was too small.

To assist in the validation of taxa-metabolite ($T \rightarrow M$) edges in our networks, we relied on the tool MIMOSA [117]. MIMOSA calculates the metabolic potential of each species, i.e., the capability of a species to produce a metabolite under the conditions of the data set. The list of all taxon-metabolite pairs from our DBNs that resulted in a positive score in MIMOSA was used as a validation database.

For taxa-gene ($T \rightarrow G$) validations, we used KEGG to build a validation database of bacterial taxa and the genes present in their genomes. To keep this database small, we only used taxa and genes present in our network. If multiple strains were available for a bacterial species, then all genes from each strain were aggregated. The one-time creation of a local validation database also speeded up our computations considerably.

To calculate the statistical significance of validated interactions compared to a null model, a Poisson-Binomial distribution test was executed. The main reason that a simple binomial test cannot be performed is the differences in the in-degree distribution between different nodes in the validation set (essential metabolites or genes would have a high probability of being connected to any given bacteria in the validation database). Because some nodes have many more validated interactions when compared to others, a uniform model for each edge does not accurately capture the null probability of selecting such an edge. This was solved with the function `ppoisbinom` from the R package `poisbinom` [120], which gives the cumulative distribution function of the probability of validating by chance at least as many interactions as the number of true positives, where each possible interaction has a different probability of being selected. The validation precision of the network was also calculated as the percentage of validated interactions from the ones predicted, even though this homogeneous metric ignores the differential significance of each interaction.

6.2.8 Laboratory validations of edges from metabolites to taxa

Wet lab experiments were carried out to validate predicted $M \rightarrow T$ interactions. Testing each such edge is not a feasible proposition. We first sorted all predicted $M \rightarrow T$ interaction based on their confidence, which we defined as the value of $|normalize(weight)| * bootstrap$. We applied this operation to the three parents Skeleton for the gene-aligned and no-alignment networks. The normalization was performed to counteract the differences of the abundance between the parent and child nodes following [96]. We then narrowed it down to edges that involved the species *P. aeruginosa* or *E. coli* because of the ready avail-

ability of these species and the expertise and facilities available to us in our laboratories. Then, we selected the top three interactions involving these two taxa. The full list of sorted interactions can be seen together with the source code and data. For positive controls we selected metabolites known to enhance growth, and as negative control we selected one metabolite that was not connected to the taxon in any of our learned networks.

The goal of the experiments was to validate a $M \rightarrow T$ edge by studying the impact of the metabolite M on the growth of taxon T. While the experimental set up does not recreate the conditions of the interaction in the microbiome, we consider this an important step in the right direction. As with the *in silico* validations, the laboratory validation confirms that the inferred interaction is a strong possibility. We selected three predicted interactions involving readily available bacteria and metabolites from the generated networks. The experiments were performed by growing relevant taxa in isolation, and adding the relevant metabolite to measure impact on growth. These metabolites were expected to positively impact the growth of the taxon because of the edge between metabolite concentration and taxon abundance.

After plotting the growth curves with the bacterium and metabolite in question, we assessed if each metabolite was enhancing/inhibiting the taxon growth using a two-tailed paired t-test when compared to growth without the metabolite.

Preliminary experiments

Three preliminary experiments were run that paved the way for the final experiment. The bacterial strains used *Escherichia coli* HB101 [19] and *Pseudomonas aeruginosa* PAO1 [68] were routinely cultured in Luria Bertani (LB 20%) broth (5 g tryptone, 10 g sodium chloride, and 5 g yeast extract per liter) or agar (LB broth with 1.5% agar) (Difco, NJ, USA). Growth curve assays were performed in media supplemented with the metabolites at 37°C. For the three preliminary experiments, we attempted to closely mimic limited nutrient environment.

1. Experiment 1, tested 0.2 mM:
    - Minimal Media (MM; gL^{-1}: $(NH_4)_2SO_4$, 2.0; K_2HPO_4, 0.5; $MgSO_4 \cdot 7H_2O$, 0.2; $FeSO_4 \cdot 7H_2O$, 0.01, pH 7.2±0.2).
    - MM + Glucose
2. Experiment 2: LB 20%, tested 0.2, 1.0 and 2 mM
3. Experiment 3: LB 20%, tested 0.2, and 1.0 mM

4-methylcatechol (4-MC, $C_7H_8O_2$) and 4-hydroxyphenylacetate (4-HPA, $C_8H_8O_3$) was used to grow *E. coli*. *P. aeruginosa* was grown in the presence of D-xylose ($C_5H_{10}O_5$), and 1-methylnicotinamide (1-MNA, $C_7H_9N_2O+$).

1. Experiment 1 (0.2 mM of metabolites)
    - The cells reached stationary phase at a very low OD; suggesting that this is not the right media to be used.
    - No effect on the exponential phase.
    - Any effect of the compound seen at the stationary phase.
2. Experiment 2 (0.2, 1 and 2 mM of metabolites)
    - The cells reached stationary phase at a higher OD.
    - No effect on the exponential phase.
    - 2 mM is lethal
3. Experiment 3 (0.2, and 1 mM of metabolites)
    - The cells reached stationary phase at a higher OD.
    - No effect on the exponential phase.
    - Effect is seen during the stationary phase

Final experiment

Because no significant difference was observed in the exponential growth rate and consequently the doubling time in all the conditions tested for both *E. coli* and *P. aeruginosa*, a new test was run in which the metabolites were added at the beginning of the stationary phase to test its effect on it.

The growth of *E. coli* was monitored hourly in the absence (control) and presence of 4-MC, 4-HPA, D-Xylose and glucose at 0.2 and 1 mM. Glucose and D-Xylose were used as enhancer positive controls. At the lower concentration (0.2 mM) compared to the control (LB 20%), 4-HPA has no effect and 4-MC, D-Xylose and glucose are enhancing (Figure 6.14(a)). The compound 4-HPA has no effect at low concentration, however at 1 mM, there is a significant enhancing effect starting at early stationary phase. At the highest concentration all metabolites produce an enhancer effect statistically significant (t-test, $p<0.05$), there is also a more pronounced enhancer effect of 4-MC and glucose compared to D-Xylose and 4-HPA (Figure 6.13).

The growth of *P. aeruginosa* PAO1 was monitored hourly in the absence (control) and presence of D-xylose, 1-MNA and succinate at 0.2 and 1 mM. Succinate was used as enhancer positive control, and 1-MNA as negative control. It is worth noting that 1-MNA does not appear in any of our networks learned, for alignment, no-alignment, Skeleton, Augmented, or any number of parents tested. D-Xylose, and succinate at 0.2 mM appears to have an enhancer effect in the stationary phase (Figure 6.14(b)), but they are not statistically significant at this concentration (Table 6.2). No effect was observed on *P. aeruginosa* growth in the presence 1-MNA. Though at 1 mM concentration, D-Xylose, and succinate produce an enhance the growth and it is statistically significant (Table 1) (t-test, $p<0.05$). The presence of 1-MNA did not have a significant effect on *P. aeruginosa* growth, it could potentially be an inhibitory compound (Figure 6.13).

6.3 Results

We developed a computational pipeline (PALM), presented in Figure 6.1, to process multi-omic microbiome data and infer their interactions. The starting point is the the relative abundance for each sample measured at potentially non-uniform intervals, even within the same subject (Figure 6.1(a)). PALM first normalizes the data and then performs Cubic B-spline interpolation using continuous curves to enable imputation of missing time points and to overcome irregular sampling (Figure 6.1(b)). Subject E5002 (yellow) does not contain enough measured time points and was removed from further analysis. The remaining smoothed curves enable estimation of unobserved time points and interpolation at specified intervals. We next temporally align the data to correct for the different progression rates of each individual (Figure 6.1(c)) against the optimal reference subject (M2072 in blue). The learned warping function is extrapolated to all omics (taxa, genes, and metabolites) of each subject. This process is then repeated, generating a different data set taking each omic as reference. as well as filter out abnormal and noisy samples (Figure 6.1(d)). Sample C3015 in grey was discarded. Alignment can be performed using either of the data types as we discuss below, and extrapolate the transformation to the other omics types, but the Figure is showing just taxa for simplicity. Our DBN learning algorithm utilizes prior knowledge to constraint the resulting model reducing overfitting and improving accuracy (Figure 6.1(e)). The biological assumption is that at the current time (t_i), the expression of host genes (hexagons) and the environmental conditions (triangles) affect the abundance of microbial taxa (circles), which impacts the expression of microbial genes (diamonds), which in turn dictates the metabolites (squares) released, and which finally impacts the abundance of taxa in the next time instant (t_{i+1}). These restrictions are flexible and can be specified as another input to the pipeline. These dynamic constraints can be customized in the form of an adjacency matrix. Using the imputed, aligned data we learn a dynamic Bayesian networks (DBN) to model interactions within and between the different data types

(Figure 6.1(f)). Figure shows two consecutive DBN time slices t_i and t_{i+1}, where dotted lines connect nodes from the same time slice referred to as *intra edges*, and solid lines connect nodes between time slices referred to as *inter edges*. Note that because both slices have the same *intra edges*, showing only the *intra edges* of slice t_{i+1} would be enough. All the other DBN figures of this manuscript will make use of this simplification. According the specified multi-omic framework, only certain types of interactions are allowed between different entities. Biological relationships are inferred from edge parameters in the learned DBN which can be positive (green) or negative (red). Finally, we validate the model predictive ability and the edges using a curated list of taxa-gene and taxa-metabolite interactions.

6.3.1 Resulting Dynamic Bayesian network models

We used the Inflammatory Bowel Disease (IBD) cohort from the iHMP study [92] that followed 132 individuals over a year. These were profiled every two weeks on average, for different omics types. The pre-processing steps included filtering, interpolation, temporal alignment, variable selection, and removal of subjects with too few time points or with noisy alignment scores (see Methods for complete details). Based on these pre-processing steps, the resulting set used to learn the model consisted of 60 individuals across 102 microbial taxa, 72 genes, and 70 metabolites. In addition, the model includes 40 host genes from each individual (measured at a single time point), and the only environmental variable considered was the week in which the sample was obtained. We used this data set to learn multi-omic dynamic Bayesian models that provide information about interactions between taxa, genes and metabolites, and the impact of environmental variables and host transcriptomics on these entities over time. We used two sets of constraints; *Skeleton* and *Augmented* depicted in Figure 6.3 and described in Section Constraining the DBN structure. The network with the complete IBD data set is presented in Figure 6.4, but for illustrative purposes the DBN learned for a subset of the data set with just the top 10 most abundant entities

of each omic type is presented in Figure 6.5. In the DBN figures, each node represents

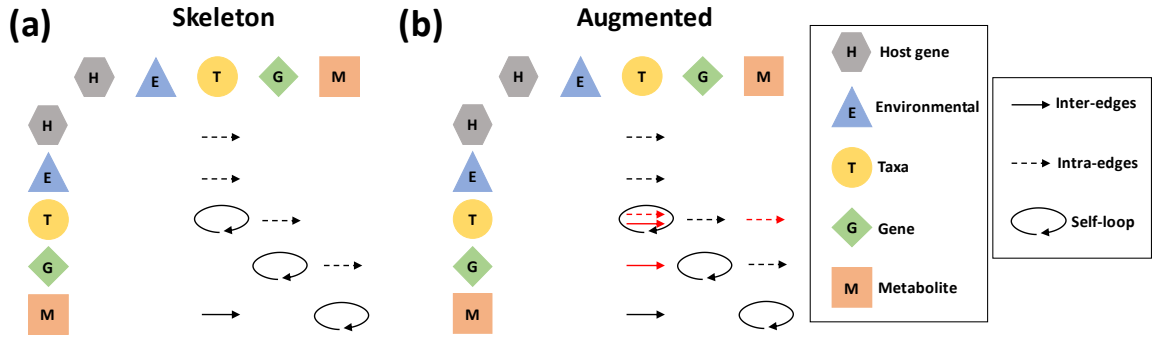

Figure 6.3: **Multi-omic frameworks used in this study.** Figure shows an adjacency matrix representation between microbiome entities for the two multi-omic frameworks used in this study: *Skeleton* (a) and *Augmented* (b). Figure highlights in red the added edge types in Augmented framework when compared to Skeleton framework.

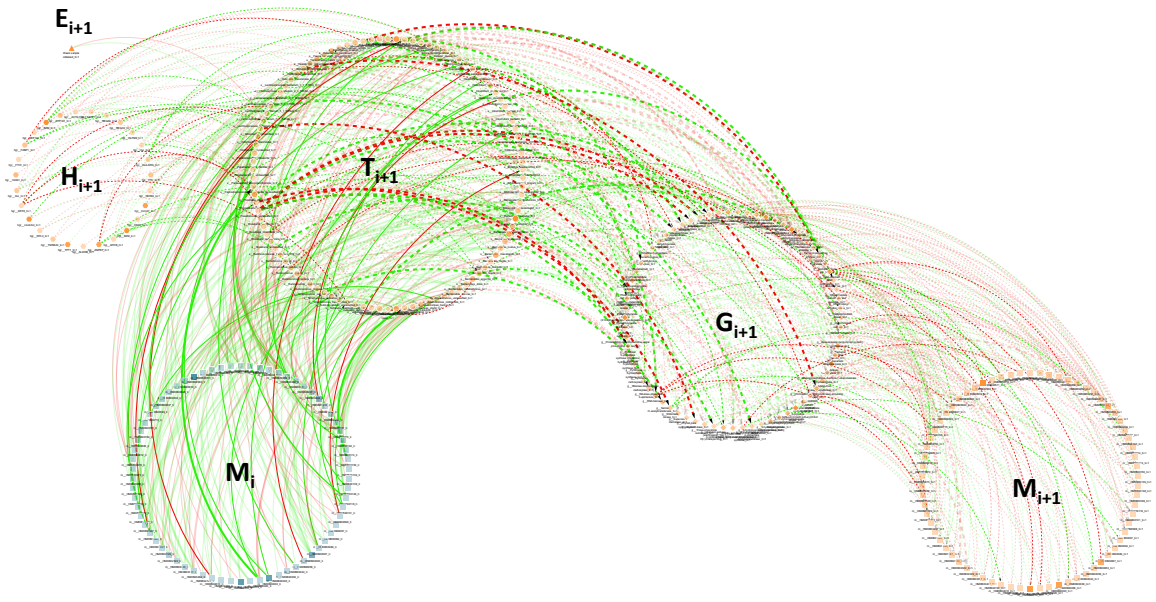

Figure 6.4: **Learned DBN by** PALM **with the Skeleton framework on the IBD data set.** Figure shows a two-stage DBN learned by PALM with Skeleton constraints and a maximum number of parents of 3. Nodes are either taxa (circles), genes (diamonds), metabolites (squares), host genes (hexagons), and environmental variables (triangles). The different node types have been grouped in different circles, their transparency is proportional to their average abundance relative to that node type. While there are two consecutive time slices t_i (blue) and t_{i+1} (orange), nodes with no neighbors and self loops were removed for simplicity. Dotted lines denote *intra edges* (i.e., directed links between nodes in same time slice), whereas solid lines denote *inter edges* (i.e., directed links between nodes in different time slices). Edge color indicates positive (green) or negative (red) temporal influence, and edge transparency indicates strength of bootstrap support. Edge thickness indicates statistical influence of regression coefficient after normalizing for parent values

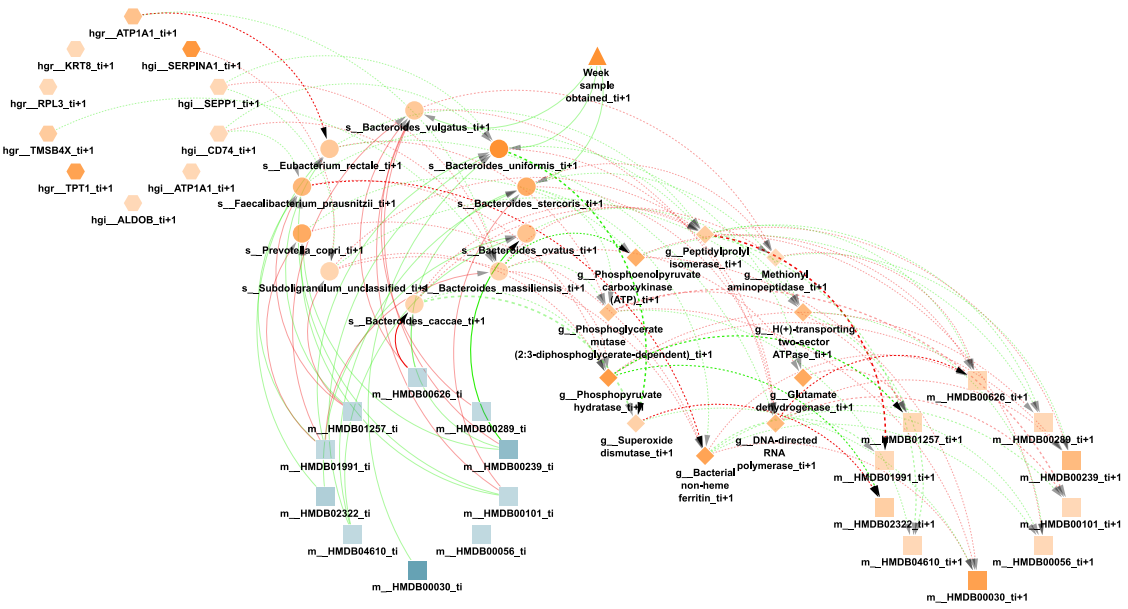

Figure 6.5: **Learned DBN by** PALM **with the Skeleton constrains on the top 10 most abundant entities of each omic type and a maximum number of parents of 3**. Nodes are either taxa (circles), genes (diamonds), metabolites (squares), host genes (hexagons), and environmental variables (triangles). The different node types have been grouped in different circles, their transparency is proportional to their average normalized abundance relative to that node type. While there are two consecutive time slices t_i (blue) and t_{i+1} (orange), nodes with no neighbours and self loops were removed for simplicity. Dotted lines denote *intra edges* (i.e., directed links between nodes in same time slice), whereas solid lines denote *inter edges* (i.e., directed links between nodes in different time slices). Edge color indicates positive (green) or negative (red) temporal influence, and edge transparency indicates strength of bootstrap support. Edge thickness indicates statistical influence of regression coefficient after normalizing for parent values, as described in [96].

either a bacterial taxon, a gene, a metabolite, or an environmental variable; directed edges represent inferred temporal relationships between these nodes. On the supporting website we also provide a Cytoscape session with an interactive version of each network, together with the original files and a list of each edge learned for every network.

Figure 6.4 shows the full network learned by PALM comprised of 284 nodes per time slice (101 microbial taxa, 72 genes, 70 metabolites, 40 host genes, and 1 environmental variable). To identify significant edges in the network we applied bootstrapping, rerunning the method 100 times with each execution using a new data set created by randomly selecting, with replacement, as many subjects as there were in the data set. We next extracted all edges from all executions, resulting in 1077 distinct directed edges (470 inter edges and

607 intra edges). Among all the 1077 edges, we observed 362 (33%) negative interactions. Interestingly, a closer look at the learned DBN revealed that 79% (193 out of 243) of nodes with potential dependencies listed at least one negative interaction. Additionally, each edge is annotated with the percentage of bootstrap iterations in which it appears. Note that while there was considerable overlap between edges learned in each iteration, since we used the union of all the networks, the number of edges in the final network is larger than the number of possible edges for a single iteration (1077 vs. 284*3 = 852). While we mainly focus on the union since it leads to more novel predictions, analysis of the intersection leads to similar statistical results. The DBN learned with the Augmented framework is shown in Figure 6.6.

6.3.2 Evaluating the learned DBN model

We first performed a technical evaluation of the learned DBN model and compared it to models constructed by other existing methods [94, 96]. The performance of each model was evaluated through leave-one-out cross-validation with the goal of predicting microbial composition using each learned model. Figure 6.7 represents the observed and predicted taxa composition for subject C3013. Additionally, we explored the effects of several different temporal alignments using taxa, genes, or metabolites. In each iteration, the whole longitudinal microbial abundance profile of a single subject was selected as the test set, and the multi-omics data from all other subjects were used for building the network and learning model parameters. Next, starting from the second time point, we used the learned model to predict an abundance value for every taxon in the test set at each time point using the previous and current time points. Finally, we normalized the predicted values in order to represent the relative abundance of each taxon and measured the average predictive accuracy by computing the mean absolute error (MAE) for the selected taxon in the network. This process of predicting microbial composition was repeated for different combinations of multi-omics training data (including metagenomics, metatranscriptomics, metabolomics

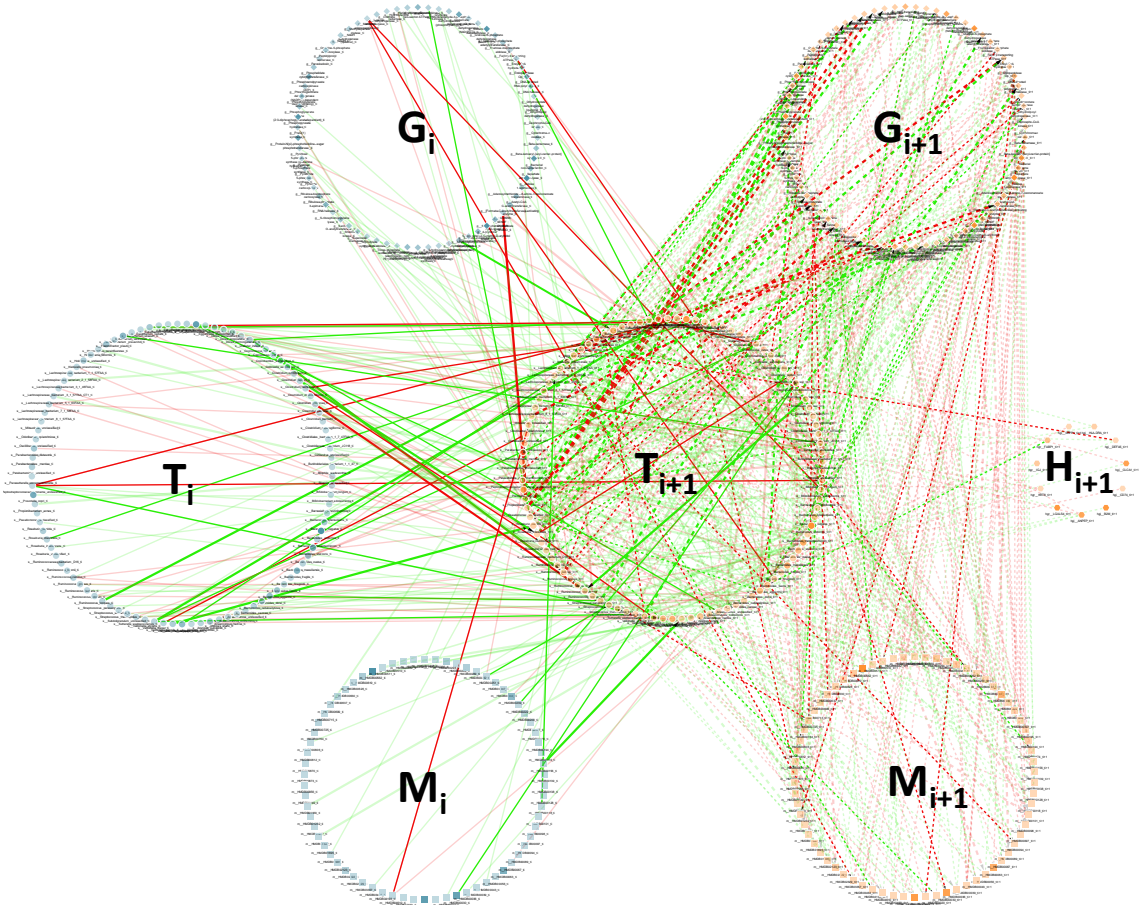

Figure 6.6: **Learned DBN by** PALM **with the Augmented framework on the IBD data set.** Figure shows a two-stage DBN learned by PALM with Augmented constraints and a maximum number of parents of 3. Nodes are either host genes (hexagons), taxa (circles), genes (diamonds), or metabolites (squares). The different node types have been grouped in different circles, their transparency is proportional to their average abundance relative to that node type, and the two time slices were separated. Dotted lines denote *intra edges*, whereas solid lines denote *inter edges*. Edge color indicates positive (green) or negative (red) temporal influence and edge transparency indicates strength of bootstrap support. Edge thickness indicates statistical influence of regression coefficient after normalizing for parent values

and host transcriptomics) on the aligned data sets, as well as unaligned data. A visual representation of the predicted trajectories for taxa- and gene-based alignment for subject C3028 is shown in Figure 6.8. The average MAE for the taxa predictions of PALM on the IBD data set for a sampling rate of two weeks using a gene-based temporal alignment is summarized in Figure 6.9. Figure 6.10 shows the average MAE of PALM across different alignments based on taxa, genes, and metabolites, respectively. We used this process to compare the multi-omics DBN strategy to the one that used only metagenomic data [96]

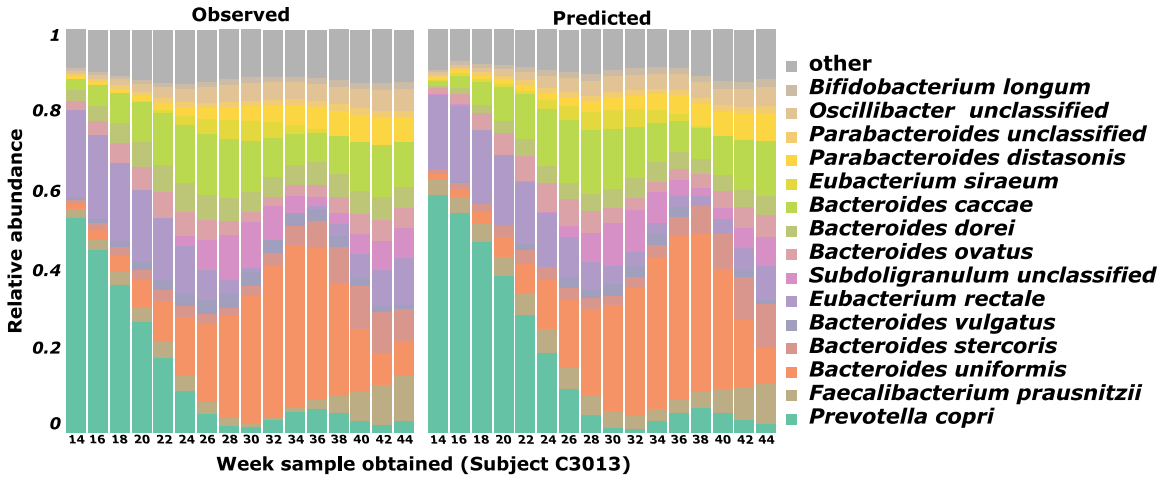

Figure 6.7: **Comparison of observed versus predicted microbial composition trajectories.** Figure shows the observed and predicted microbial composition trajectories for a representative aligned subject (C3013). Microbiota composition profile for this subject is comprised of the top 15 most abundant bacteria along with all remaining bacteria merged into the "other" category. The *y* axis corresponds to the relative abundance of each bacteria, while the *x* axis represents the original measured time point after alignment.

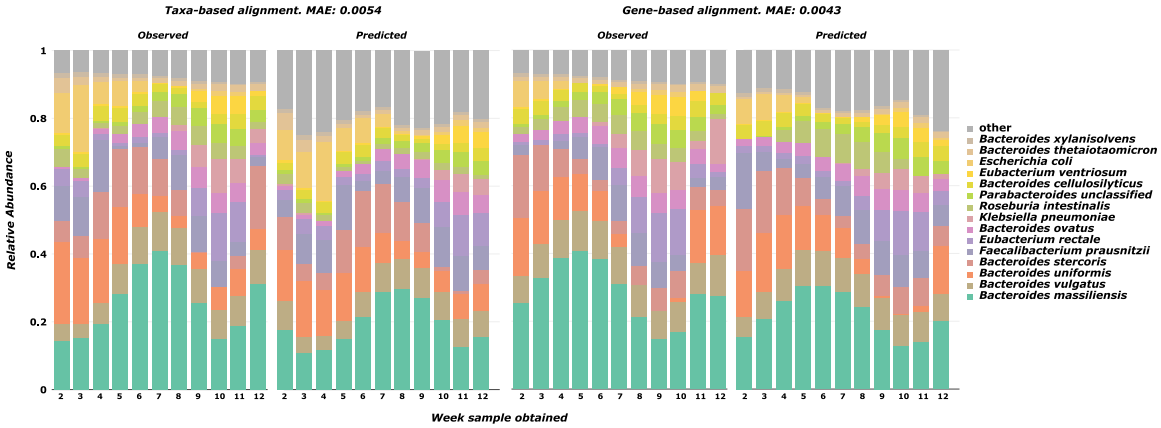

Figure 6.8: **Comparison of observed versus predicted microbial composition trajectories.** Figure shows the observed and predicted microbial composition trajectories for a representative aligned subject (C3028). Microbiota composition profile for this subject is comprised of the top 15 most abundant bacteria along with all remaining bacteria merged into the "other" category. The *y* axis corresponds to the relative abundance of each bacteria, while the *x* axis represents the original measured time point after alignment. Figure highlights the observed and predicted trajectories of this subject between taxa-based alignment (left) and gene-based alignment (right). We note that aligned interval for gene-based alignment is stretched and shifted when compared to the taxa-based alignment. For each alignment type, a DBN was learned with the Skeleton framework and a maximum number of parents of 3, and tested on the previously unseen C3028 subject. Gene-based alignment exhibits a lower prediction error (MAE=0.0043) than taxa-based alignment (MAE=0.0054). In this example, taxa-based alignment does a worse job at predicting low abundance bacteria than gene-based alignment.

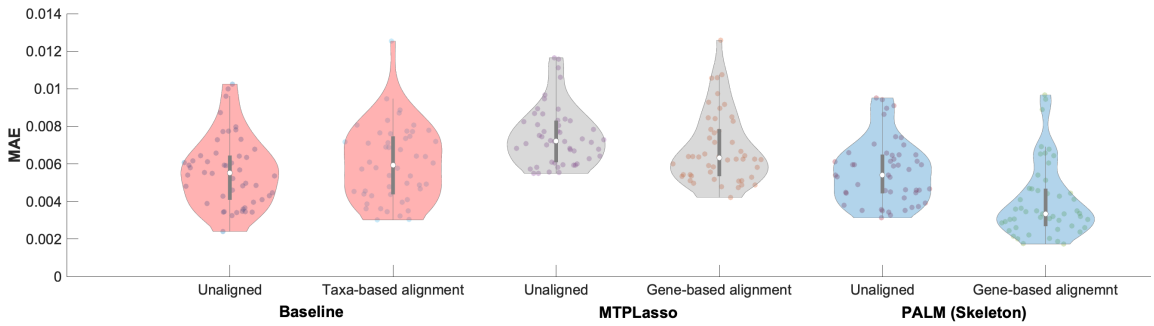

Figure 6.9: **Comparison of average predictive accuracy between methods on the IBD data.** Figure shows the MAE of our proposed DBN models against a baseline method using only metagenomic data and a previously published approach, MTPLasso, which models longitudinal multi-omics microbial data using a generalizded Lotka-Volterra (gLV) model for a sampling rate of two weeks which most closely resembles the originally measured time points. Figure also compares the performance of each method on the unaligned and aligned data sets.

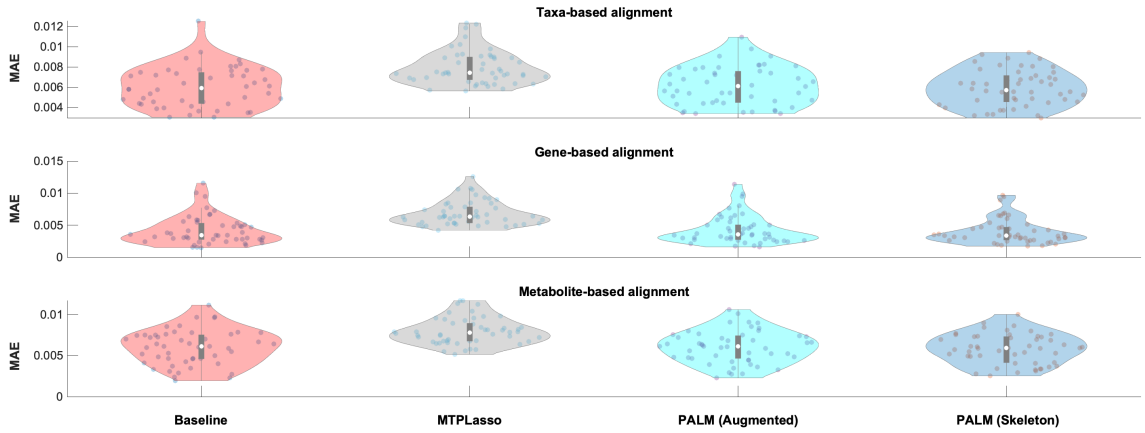

Figure 6.10: **Comparison of average predictive accuracy between methods on the IBD data sets aligned using taxa, gene and metabolite data.** Figure shows the MAE of PALM models (Augmented and Skeleton) against a baseline method and a previously published approach (MTPLasso) for a sampling rate of two weeks which most closely resembles the originally measured time points. Although baseline method uses only metagenomic data, gene- and metabolite-based alignment were generated using gene expression and metabolite intensities data, respectively.

referred to as Baseline on the unaligned and aligned IBD data, as well as MTPLasso [94] which models time-series multi-omics microbial data using a gLV model. In both cases, we used the default setup and parameters, as described in the original publications. As shown by Figure 6.9 our method outperforms Baseline and MTPLasso when using gene expression data for temporal alignment of microbiome samples. Specifically, when using

gene expression data for alignment the MAE significantly dropped to 4.01E-03 when compared to a MAE of 6.03E-03 achieved using taxa alignment as indicated by a one-tailed unpaired t-test with null hypothesis that the means are equal and alternative hypothesis that population mean of method with gene expression-based alignment is less than mean of (baseline) taxa-based alignment method (p-value = 6.71E-07). Figure also shows that gene-based alignment significantly Figure 6.10 shows that our method outperforms MT-PLasso when all microbiome entities are used in the model (taxa: 5.93E-03 vs. 7.93E-03; metabolite: 5.82E-03 vs. 7.97E-03). Moreover, figure shows that our method outperforms Baseline (taxa: 5.93E-03 vs. 6.03E-03; gene: 4.01E-03 vs. 4.19E-03; metabolite: 5.82E-03 vs. 6.01E-03). Overall, our results suggest that gene expression data is more suitable for temporal alignment of multi-omics microbiome samples. This is consistent with previous findings which reported technical noise dominates the abundance variability for nearly half of the detected taxa in gut samples [76]. Therefore, we have used gene-based alignment for the rest of the analysis discussed next.

6.3.3 Computationally validating predicted edges

We compiled a database of Taxon-Metabolite ($T \rightarrow M$) and Taxon-Gene ($T \rightarrow G$), and used that database to validate the predicted edges and score each model. A ($T \rightarrow G$) interaction was added to the database if any strain of taxon T has gene G in its genome according to KEGG. For $T \rightarrow M$ we relied on the tool MIMOSA [117], that calculates the metabolic potential of each taxon for a particular data set. See Methods 6.2.7 for complete details

Each predicted interaction was either considered "validated" if it appears in the validation database, or "not validated" if it was not found, but the parent and child nodes were part of the database. Interactions predicted between taxa and/or metabolites not included in the database were not used in this analysis. We compared the results between the DBNs learned by PALM using the Skeleton and Augmented constraints, as well as a random net-

work. To generate the random network, we used the same nodes in the multi-omic network and assigned the same number of edges as in the learned DBN by randomly selecting a parent and child from the possible interaction list (Figure 6.3). This was repeated 1000 times, averaging the metrics over all random runs. Figure 6.11(a) shows the validation

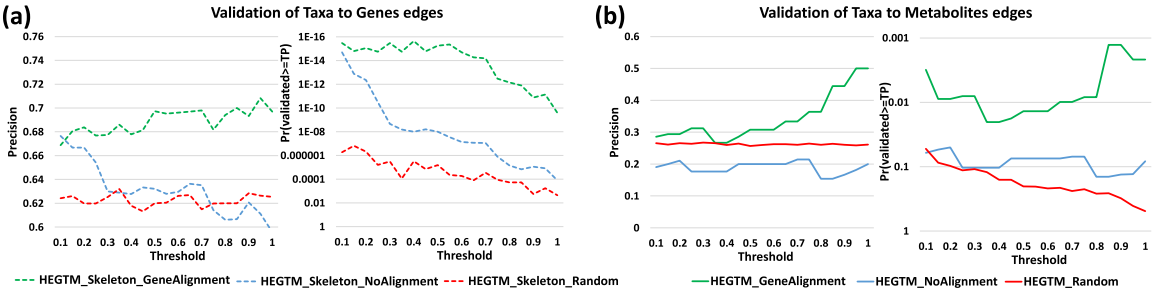

Figure 6.11: *In silico* **validation results of the predictions of** PALM **for the IBD data set**. The left part of each subfigure shows the precision (percentage of predicted edges that were validated) and the right part shows the probability of validating at least that many edges by chance (y axis in reverse logarithm scale so higher is better for both). The x represents the bootstrap value threshold that was used to select the edges included in the analysis. For example, for a threshold of 0.7, the score for edges that appear in more than 70% of the repetitions is shown. . **(a)** Validation for $T \rightarrow G$ interactions (bacterial taxon expressing a gene) **(b)** Validation for $T \rightarrow M$ interactions (bacterial taxon consuming a metabolite)

comparison for edges of the form $T \rightarrow G$. The Skeleton constraints were used to learn the networks. The DBN learned with the gene-aligned data set (green) was compared against a DBN learned with the data set that was not aligned (blue). As can be seen, the aligned data set results in networks that outperform the networks from the unaligned data and random networks, with the precision difference increasing as the threshold increases. This indicates that the bootstrap score for an edge can serve as a way to determine its likely accuracy.

Figure 6.11(b) shows the comparison for edges of the form $T \rightarrow M$. For this, we can only use the network results from the Augmented constraints since no such edges are permitted when using Skeleton. Again, we observe better performance for the networks from aligned data when compared to the networks from unaligned data and random networks, with an improvement in performance for higher bootstrap thresholds. Note that for both $T \rightarrow G$ and $T \rightarrow M$, the not aligned network does not even outperform the ran-

dom network, highlighting the importance of the alignment step. Figure 6.12 shows the computational validation results of metabolites- and gene-based alignment.

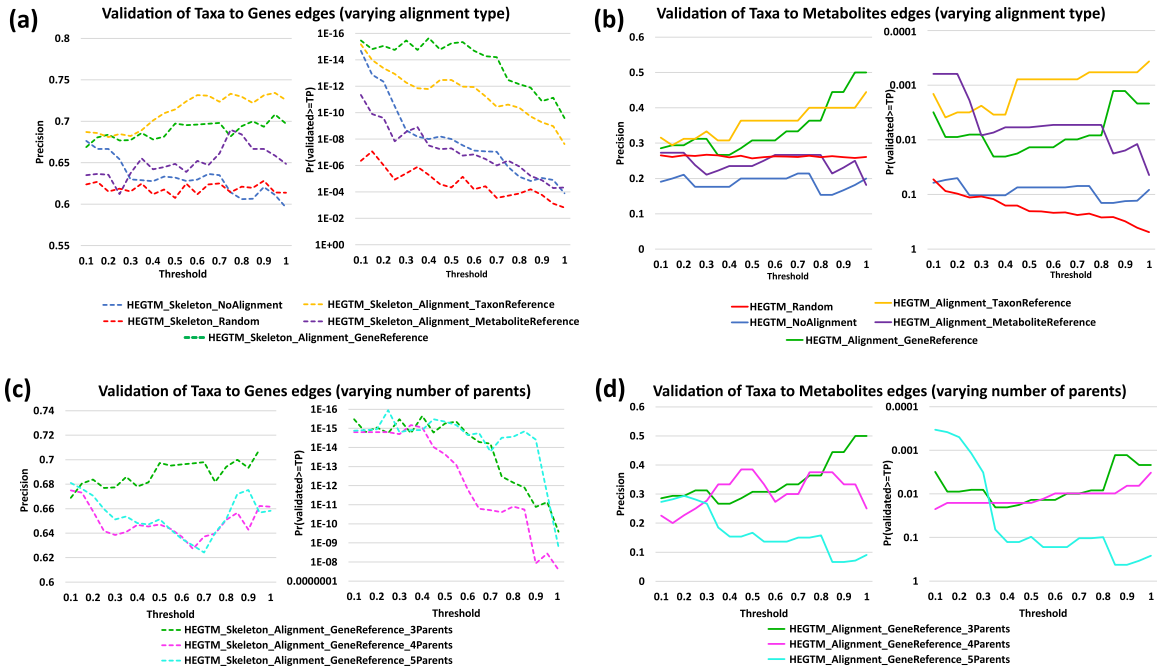

Figure 6.12: *In silico* **validation results with 100 bootstrap repetitions**. Graphs from **(a)** and **b** show the performance of different alignment types, while **(c)** and **(d)** vary the number of parents used when learning the networks. The left part of each subfigure shows the precision (percentage of predicted edges that were validated) and the right part shows the probability of validating at least that many edges by chance (y axis in reverse logarithm scale so higher is better for both). The x represents the bootstrap value threshold that was used to select the edges included in the analysis. For example, for a threshold of 0.7, the score for edges that appear in more than 70% of the repetitions is shown. The dashed lines ($T \rightarrow G$ interactions) were learned using the *Skeleton* constraints, and the solid lines ($T \rightarrow M$ interactions) were learned using the *Augmented* constraints, because the edges $T \rightarrow M$ are not allowed directly in *Skeleton*. **(a)** Validation for $T \rightarrow G$ interactions (bacterial taxon expressing a gene) varying the alignment reference used. Noalignment barely does better than the random baseline, followed closely by the metabolite-based alignment. Taxon-based alignment has a slight better precision than gene-based, but the latter has a much better probability score than the former. **(b)** Validation for $T \rightarrow M$ interactions (bacterial taxon consuming a metabolite) varying the alignment reference used. Taxon- and metabolite- based alignment have a lower precision than the random baseline, but a better probability score. **(c)** Validation for $T \rightarrow G$ interactions (bacterial taxon expressing a gene) varying the maximum number of parents allowed. Learning with 3 Parents has a much better precision than with 4 and 5, and a similar probability score. **(d)** Validation for $T \rightarrow M$ interactions (bacterial taxon consuming a metabolite) varying the maximum number of parents allowed. Learning with 3 Parents has a better precision than with 4 and 5 for small and big thresholds. Learning with 5 parents has a better probability score for low thresholds, but it seems that it is by chance, because as the thresholds becomes more stringent, it quickly fares worse, while 3 parents overtakes 4 parents by a small percentage.

6.3.4 Biological validation experiments

We performed experiments to validate a few of the interactions predicted by the DBNs. We focused on edges of the form $M \rightarrow T$, i.e., edges where a metabolite is predicted to impact the abundance of a bacterial taxon. Such edges imply that the metabolite M promotes (or represses, depending on the sign) the growth of the bacterial taxon T under appropriate growth conditions.

We first sorted all predicted $M \rightarrow T$ interactions based on their confidence (combination of normalized weight and bootstrap score). Next, we selected some of the top edges to validate taking into account the availability of the metabolites and taxa and the laboratory resources for growth experiments at our disposal. See Methods Section Laboratory validations of edges from metabolites to taxa for details on this process. Based on these considerations we focused on two common model organisms, namely *Pseudomonas aeruginosa* and *Escherichia coli* and picked from the top predictions those involving any of these two taxa for validation.

- 4-Methylcatechol (4-MC) → *Escherichia coli*
- 4-Hydroxyphenylacetate (4-HPA) → *Escherichia unclassified*
- D-Xylose → *Pseudomonas unclassified*

Standard lab strains *P. aeruginosa* PAO1 [68] and *E. coli* HB101 [19] were used in the laboratory experiments. The choice of chemicals used to verify was somewhat limited by commercial availability. A standard Luria Bertani (LB 20%) culture media was used to measure the bacterial growth curve (expressed as bacterial density, OD600) in the absence and presence of metabolites. Metabolites were added at the stationary phase when the bacteria were multiplying very slowly, mimicking a biofilm growth [113]. As positive controls, the preferred carbon sources of *E. coli* and *P. aeruginosa*, glucose, and succinate, respectively, were chosen. Figure 6.13 shows the resulting growth curves of the microbes

before and after adding the metabolites, and a control case without adding the metabolites (LB 20%). Confirming the predictions of our networks, D-xylose significantly enhanced *P. aeruginosa*, and 4-HPA and 4-MC significantly increased the *E. coli* growth. Regarding the controls, as expected D-xylose and glucose enhanced *E. coli*, and succinate enhanced *P. aeruginosa* whereas the negative control 1-Methylnicotinamide (1-MNA) did not.

The p-values for all observations can be seen in Table 6.1, where a two-tailed paired t-test was executed for the three time points with the highest difference from the baseline. For more details on the experimental settings please refer to Methods Section Laboratory validations of edges from metabolites to taxa, including Figure 6.14 and Table 6.2 for growth results at a lower concentration of 0.2 mM.

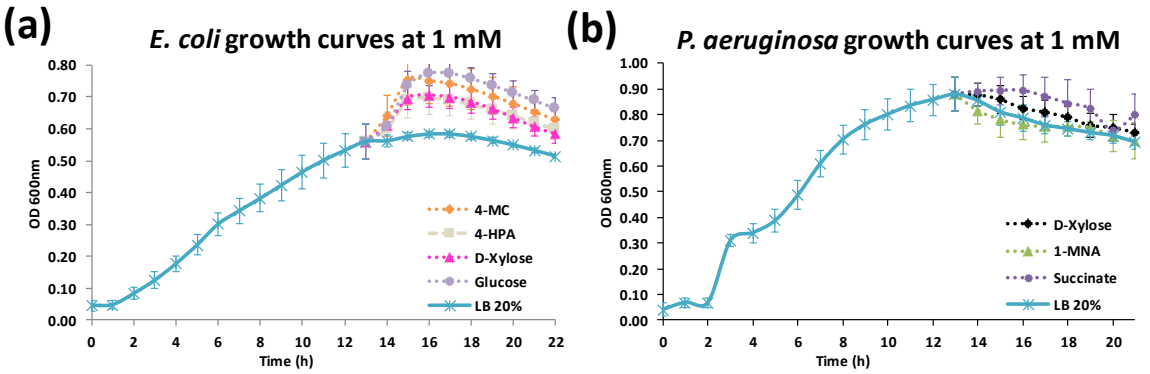

Figure 6.13: **Growth curves at 1 mM**. In this figure different metabolites were introduced at 1mM concentration at the end of the exponential phase (0-14h). Figure shows the growth curves after all data points were averaged over 10 replicates. **(a)** *E. coli*, with glucose and D-xylose as positive controls. **(b)** *P. aeruginosa*, with succinate as positive control and 1-MNA as negative control.

6.4 Discussion

Previous microbiome studies focused primarily on metagenomics sequence data. More recent data sets are much richer, notably including host and bacterial gene expression, and metabolomics data. The ability to integrate these multi-omics, longitudinal data remained a major challenge for microbiome analysis.

Escherichia coli **HB101**			
Met.	**16h**	**17h**	**18h**
LB	0.583	0.584	0.576
4-MC	0.750 (0.0002)	0.741 (0.0003)	0.724 (0.0003)
4-HPA	0.697 (0.0005)	0.692 (0.0005)	0.677 (0.0004)
D-xylose	0.704 (0.0017)	0.699 (0.0016)	0.684 (0.0019)
Glucose	0.774 (0.0005)	0.773 (0.0003)	0.758 (0.0003)

Pseudomonas aeruginosa **PAO1**			
Met.	**15h**	**16h**	**17h**
LB	0.811	0.787	0.760
D-Xylose	0.860 (0.0189)	0.825 (0.0395)	0.810 (0.0317)
1-MNA	0.782 (0.0717)	0.761 (0.1551)	0.754 (0.7547)
Succinate	0.893 (0.0716)	0.893 (0.1195)	0.870 (0.0172)

Table 6.1: **Effect of 1 mM metabolites on bacterial cell density**. Taxa density appears in black (OD600), while the p-values are inside parenthesis. Red p-values represent a significant difference compare to LB 20% ($p<0.05$). Green p-values represent a non-significant difference from LB 20%.

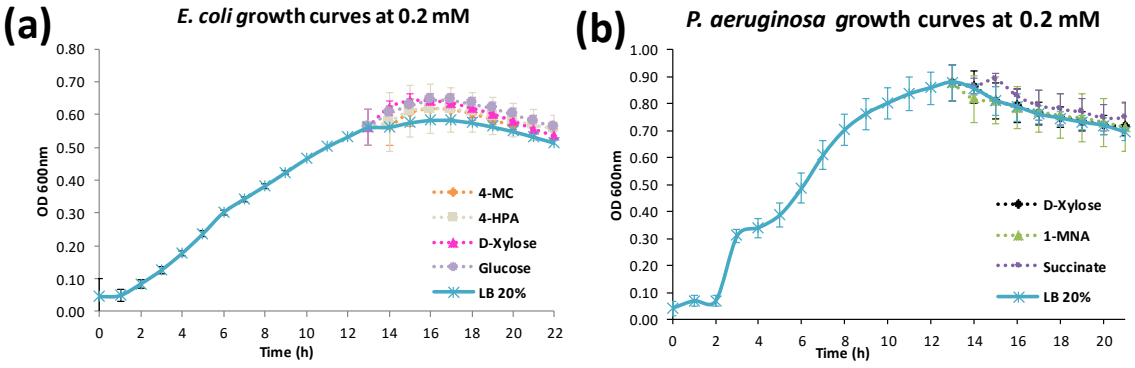

Figure 6.14: **Growth curves (0.2 mM)**. In this figure different metabolites were introduced at 0.2 mM concentration at the end of the exponential phase (up to time 14h). Figure shows the growth curves after all data points were averaged over 10 replicates. **(a)** *E. coli*, with Glucose and D-Xylose as positive controls. **(b)** *P. aeruginosa*, wich Succinate as positive control and 1-MNA as negative control.

Here we have presented PALM, a new approach based on a temporal normalization using continuous curve alignment followed by DBN modeling. Our method first represents each time series using continuous curves and then aligns them using a reference time series. Next, we sample the aligned curves uniformly and learn a DBN model that combines data from taxa, host genes, bacterial genes, and metabolites. Edges in the DBN repre-

***Escherichia coli* HB101**			
Met.	**16h**	**17h**	**18h**
LB	0.583	0.584	0.576
4-MC	0.617 (0.0108)	0.614 (0.0055)	0.602 (0.0026)
4-HPA	0.617 (0.1458)	0.616 (0.1332)	0.608 (0.1049)
D-Xylose	0.642 (0.0028)	0.635 (0.0073)	0.619 (0.0132)
Glucose	0.644 (0.0032)	0.644 (0.0036)	0.633 (0.0049)
***Pseudomonas aeruginosa* PAO1**			
Met.	**15h**	**16h**	**17h**
LB	0.811	0.787	0.760
D-Xylose	0.812 (0.9029)	0.793 (0.7561)	0.763 (0.7710)
1-MNA	0.813 (0.9117)	0.786 (0.9416)	0.772 (0.6110)
Succinate	0.889 (0.0128)	0.830 (0.6979)	0.795 (0.0530)

Table 6.2: **Metabolite effect at 0.2 mM**. Taxa density appears in black, while the p-values are inside parenthesis. Red p-values represent a significant difference compare to LB ($p<0.05$). Green p-values represent a non-significant difference from LB.

sent predicted interactions between the entities and can be used to explain changes in the microbiome over time.

Applying our methods to data from IBD patients, we show that multi-omics DBNs can successfully predict taxa abundance at future time points, thus improving on models that do not use all available data and on previous methods developed for modeling temporal taxa interactions. We curated validations for taxa-to-metabolite and taxa-to-gene interactions; interactions predicted by the learned DBNs significantly intersect these interactions. Finally, we experimentally tested and validated select predictions of metabolite $\rightarrow$ taxa relationships.

Microbiome interaction databases are critical for evaluating learned DBNs, but appear to be incomplete. More complete databases of validated interactions would help validate computational methods for this task. The laboratory validations show a viable way to validate some of the interactions. However, they could also be improved by attempting to recreate more realistic conditions for the experiments and could be enhanced to validate other omics observations as well.

Comparing DBNs constructed using different omics data allows for an important kind of inference (Figure 6.15). According to this, in the DBN built using only metagenomics

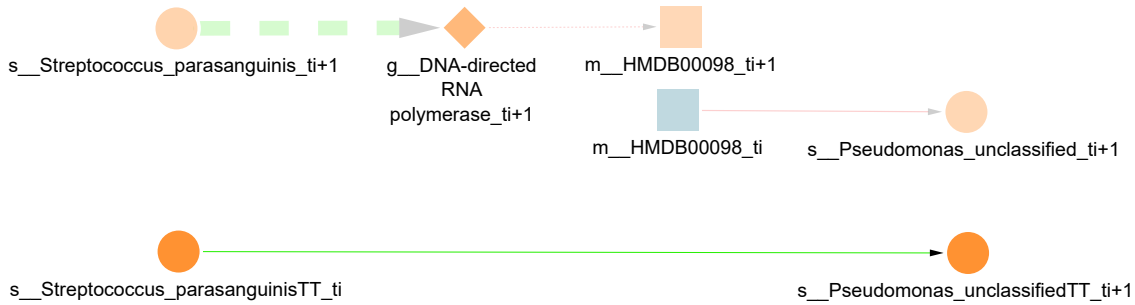

Figure 6.15: **Multi-omics inferred chain of interactions.** The edge *Streptococcus parasanguinis* → *Pseudomonas unclassified* on the bottom gets explained when added multiomic data (TT stands for Taxa-Taxa network). In the multi-omic network that interaction gets replaced by *Streptococcus parasanguinis* (T) → rna polymerase (G) → D-Xylose (M) → *Pseudomonas unclassified* (T).

data, the edge *Streptococcus parasanguinis* → *Pseudomonas unclassified* appears with a high confidence (bootstrap score of 1). In the multi-omic DBN the following chain of interactions can be found: *Streptococcus parasanguinis* (T) → rna polymerase (G) → D-Xylose (M) → *Pseudomonas unclassified* (T). It is important to note that though DBN edges may not imply causal relationships, the *in silico* validation process described in this chapter supports the above relationships. Finally D-Xylose → *Pseudomonas unclassified* was validated experimentally (Section Biological validation experiments). Thus, comparing DBNs before and after adding additional multi-omics data can "unroll" and "explain" relationships between taxa.

Our alignment, DBN methods, validation software and other scripts are implemented in either Python, R, or Matlab. The source code for PALM and the data set used will be freely available under the MIT Open Source license agreement upon publication at `http://biorg.cis.fiu.edu/palm/`. We will also include the networks learned and interactions predicted, sorted by relevance.

CHAPTER 7

LONGITUDINAL CAUSAL MULTI-OMIC NETWORK INFERENCE

In this chapter we present METALICA, a suite of novel tools and techniques to uncover significant details about the causal nature of microbial interactions.

7.1 Background

The methods described in the previous two chapters for inferring DBNs involved starting from next generation sequencing data and other omics measurement technologies. Every attempt was made to ensure that the resulting networks had biologically meaningful edges and were not a result of overfitting. However, even if an edge was directed from an entity measured at a previous time point to an entity measured at a later time point, it did not guarantee that a DBN edge represented a true and direct causal interaction, or if it was merely the result of a statistical correlation caused by an indirect causal relationship. Microbiomes are complex environments with many subtle relationships. However, causal inference relies on noisy data from error-prone technologies, and has to contend with a host of hidden confounders that may be hard or impossible to identify, let alone measure them. The jump to infer causality is a natural next step in inferring multi-omic interactions, and the lack of research in this area is striking. Most of the causal microbiome literature focuses on the causal impact of the microbiome to health or disease, but not on the causal interactions between these microorganisms [69, 97, 151, 137].

Another major challenge in building causal models of biological interactions lies in developing methods to validate them and in providing confidence measures. Validation by calculating the prediction error of the model, while informative, does not shed light on the accuracy of specific edges and interactions predicted by the model that we are interested in. We broadly discuss approaches to validate the different types of edges present in the networks, which are the parameters learned by the model. Edges from a taxon to a gene can be circumstantially validated by verifying that (a) the taxon has a non-zero abundance,

(b) the taxon's genome includes at least one copy of the gene, and (c) the gene is expressed. Similarly, edges from a gene to a metabolite or a taxon to a metabolite could potentially be validated by verifying that (a) the gene has a non-zero expression value, (b) the gene is involved in the metabolic pathway for that metabolite, and (c) the metabolite is produced and has a non-zero concentration.

7.2 Methods

Below we describe the datasets used and the network learning methods that were executed to create different causal networks. Finally, we introduce the causal network analysis methods that we developed to evaluate and compare the inferences made by the network learning algorithms.

7.2.1 Data sets and pre-processing

To test our proposed methods, we used the Inflammatory Bowel Disease (IBD) cohort from a study that included 132 individuals across five clinical centers [92]. During a period of one year, each subject was profiled (biopsies, blood draws, and stool samples) every two weeks on average. This yielded temporal profiles for metagenomes, metatranscriptomes, metaproteomes, metabolomes and viromes across all subjects. Additionally, for each subject, host- and microbe-targeted human RNA sequencing was yielded from biopsies collected at initial screening colonoscopy sampled from two sites in the gut (ileum and rectum). All data are fully described and available at `https://ibdmdb.org`.

We used the dataset generated in Chapter 6, which provides aligned and unaligned versions of metagenomics, metatranscriptomics, metabolomics, and host transcriptomics data. See Sections 6.2.1, 6.2.2, and 6.2.3 for more details on the processing that led to the data. As explained in those sections, the data were normalized and centered, the time series were smoothed, and then temporally aligned prior to inferring the Dynamic Bayesian Network models.

Since two of the methods that we used require a higher number of timepoints than the DBN method applied in Chapter 6, new datasets were generated increasing the sampling frequency. This way, instead of a sampling rate of 14 days, we duplicated the number of time points by sampling at a sampling rate of seven days. We then generated different subsets from the full dataset, each with different omic types. Let $\mathbb{T}$, $\mathbb{G}$, and $\mathbb{M}$ represent the entities in the datasets with just taxa, genes, and metabolites, respectively. We can combine them and generate different subsets. The resulting datasets are the aligned and unaligned versions of the following: {$\mathbb{T}$, $\mathbb{G}$, $\mathbb{M}$, $\mathbb{TG}$, $\mathbb{TM}$, $\mathbb{GM}$, $\mathbb{TGM}$}.

In an effort to increase the number of biologically interpretable results and to get the most significant validations of the interactions, we focused on attributes that were cataloged in KEGG. We first selected the top 50 taxa, genes, and metabolites, and then intersected that selection with the attributes supported in the databases created from KEGG and MIMOSA in Section 6.2.7. The outcome of this was the selection of 27 bacterial species, 34 genes, and 19 metabolites, in addition to one clinical variable (sampling time, represented by the week during which the sample was obtained).

7.2.2 Constraining structures

As explained in Section 4.2.3, we constrain the set of allowable edges by providing a *Skeleton* structure, which is part of the input to the DBN construction step. These constraints (in the form of a matrix received as an input to the function) only allow edges between certain types of nodes, greatly reducing the complexity of searching over possible structures and preventing over-fitting. Specifically, we allowed intra edges (i.e., edges within same time point) from taxa nodes to gene (expression) nodes and from gene nodes to metabolites (concentration) nodes. All other interactions within the same time point (for example, direct gene to taxa) were disallowed. We also allowed inter edges (i.e., edges between nodes from adjacent time points) from metabolites to taxa nodes in the next time point, and *self-loops* from any node $A_1^{t_i}$ to $A_1^{t_{i+1}}$. The restrictions in the *Skeleton* reflect our

understanding of the basic ways the different entities interact with each other, i.e., taxa express genes that they carry on their genomes; these, in turn, are involved in metabolic pathways for metabolites that they are able to produce; the metabolites impact the growth of taxa (in the next time slice).

We also used a less constrained framework referred to as the *Augmented* skeleton. Unlike the original Skeleton, the Augmented framework also allowed direct edges between taxa and metabolites to account for cases where noise or other issues related to profiling of genes can limit our ability to indirectly connect taxa and the metabolites they produce. All other edges from the skeleton were maintained. Figure 6.3 summarizes each framework in the form of an adjacency matrix.

7.2.3 Dynamic Bayesian Networks

With edges that represent lagged dependencies, DBNs are a type of BNs suited for representing temporal connections, conducting time-varying probabilistic inference, and performing causal analysis under uncertainty. In this dissertation, we focus on a version of DBNs called Two-Timeslice BN (2TBN), which relates variables to each other over adjacent time steps. Any variable X_i^t can be calculated from the internal regressors, the current time point t and the previous time point $t-1$. For more detailed discussions, refer to Chapter 4 and Chapters 5 and 6.

DBNs were learned for all subsets of datasets from Section 7.2.1 (i.e., {$\mathbb{T}$, $\mathbb{G}$, $\mathbb{M}$, $\mathbb{TG}$, $\mathbb{TM}$, $\mathbb{GM}$, $\mathbb{TGM}$}), for several different number of allowable parents ($\{3,4,5,6\}$), for aligned and unaligned datasets, and for the Skeleton and Augmented constraint frameworks. A total of 100 networks were learned by subsampling subjects with replacement from each dataset (100 bootstrap repetitions). The networks were then combined, averaging the regression coefficient of the edges. Each edge was also labeled with the bootstrap support (percentage of times that edge appears). Each repetition was set to run independently on a separate processor using Matlab's Parallel Computing Toolbox. The following

two methods do not learn based on a global score such as Likelihood, but rather on conditional independence tests.

7.2.4 Causal Networks using the TETRAD Suite

The tsGFCI (SVAR-GFCI) [101] algorithm is implemented in the TETRAD package [153, 30], for which PyCausal [25] is the wrapper that was used in this dissertation. The tsGFCI algorithm is a version of tsFCI [47] and GFCI, while tsFCI is, in turn, the evolution of FCI [31], which in turn is a modification of PC-stable, which was designed by changing PC, which an adaptation of the SGS algorithm [170] (see Section 3.4.2 for a more detailed description).

Algorithm tsFCI (SVAR-FCI) is based on the FCI algorithm, with the following modifications: it uses the direction of time to orient interactions, and it enforces repeating structures for both adjacencies and orientations based on the stationarity assumption. Since the hybrid score-based GFCI is usually more accurate in finite samples than FCI, similar modifications were made in the development of tsGFCI; It uses a greedy initial adjacency search enforcing time order and repeating structures, and scores the structures using BIC [155].

Different networks were learned with N bootstrapping repetitions for each significance threshold, $\alpha \in \{0.0001, 0.001, 0.01, 0.1\}$, for the `PositiveCorr` CI test (one of seven possible tests available), for the `FisherZScore` network score (one of eight possible choices available), and for each combination of omics datasets. For our experiments, we used $N = 10$.

7.2.5 Causality with Tigramite

Tigramite [147] implements the PCMCI algorithm, which has two stages:

1. **Condition selection via PC_1**: It obtains an estimate of a superset of the parents $Pa(X_i)^G$ for all variables from $\mathcal{U} = \{X_1, X_2 \ldots, X_n\}$, with a modified version of the

PC-stable algorithm, adapted for time series. For every variable, first the preliminary parents are initialized to all possible parents. Then, for a growing conditioning set size, test for all variables $X_i^{t-\tau}$ from $Pa(X_j^t)$ if the null hypothesis,

$$\mathbf{H_0} : X_i^{t-\tau} \not\perp\!\!\!\perp X_i^t \mid \boldsymbol{S}, \text{ for any } \boldsymbol{S} \subseteq Pa(X_j^t) \text{ with } \mid \boldsymbol{S} \mid = p, \tag{7.1}$$

can be rejected at a significance threshold α and removes the current parent from the set. We iterate over all possible sets $\boldsymbol{S} \subseteq Pa(X_j^t) \setminus \{X_i^{t-\tau}\}$ with cardinality p, up to a maximum number of combinations q_{max}.

2. **Causal discovery stage**: Using the parents from the previous stage, we apply the MCI algorithm, which tests all pairs of variables, and a set of time delays $\tau \in \{1, \ldots, \tau_{max}\}$, and establishes the edge $X_i^{t-\tau} \to X_j^{\tau}$ if and only if

$$X_i^{t-\tau} \not\perp\!\!\!\perp X_j^t \mid Pa(X_j^t) \setminus \{X_i^{t-\tau}\}, Pa_{p_x}(X_i^{t-\tau}), \tag{7.2}$$

where $Pa_p(X_i^{t-\tau})$ denotes the p strongest parents.

Since Tigramite assumes that all the data points belong to one subject, bootstrap cannot be implemented in the usual way of subsampling subjects with replacement. Instead, a different network was learned for each subject, and the resulting networks were then combined. The percentage of times that a given edge appears in all the different networks was annotated in the edge, together with the averaged cross-link strength. Different networks were learned for each significance threshold in $\alpha \in \{0.0001, 0.001, 0.01, 0.1\}$, for each CI test available (`GPDC, CMIknn, ParCorr`) [147] and for each omics dataset.

The following sections introduce a series of causal network analysis techniques, which will be applied to the networks learned with the methods introduced in Sections 7.2.3 – 7.2.5 using DBNs, TETRAD and Tigramite.

7.2.6 Unrolling

Typical algorithms for network learning and analysis do not differentiate between direct and indirect interactions, and fail to elucidate the actual reason why two entities are causally related to each other. An important challenge in microbiome analysis is to determine why and how two taxa are interacting with each other given multi-ommics data. We introduce the term **unrolling** as the process of determining the sequential steps by which two omic entities potentially interact with each other. This is done by learning different independent networks using different subsets of omics. By learning the networks with the $\mathbb{T}$ and the $\mathbb{TM}$ datasets, we can explain away interactions between microbial taxa as suggested by the former using the interactions between microbial taxa through metabolic intermediaries as suggested by the latter.

To make this more formal, we will let $G_{\mathbb{X}}$, $V_{\mathbb{X}}$, and $E_{\mathbb{X}}$ represent the graph, the vertex set and the edge set, respectively, of the network learned using dataset $\mathbb{X}$. Now, an explanation by unrolling occurs if the following three conditions are true:

1. There is an edge between T_i and T_j in $G_{\mathbb{T}}$, for some $T_i, T_j \in V_{\mathbb{T}}, i \neq j$.
2. There is *no* edge between T_i and T_j in the network $G_{\mathbb{TM}}$.
3. The edges from T_i to M_x and from M_x to T_j exist in $G_{\mathbb{TM}}$ for some metabolite in $V_{\mathbb{TM}}$.

If the above three conditions are met, we infer that the interaction between the taxa T_i and T_j is not direct, but is happening through an intermediary metabolite M_x, which is produced by T_i and consumed by T_j.

This process can be replicated by unrolling the edges of the network inferred from $\mathbb{T}$ with the one inferred from $\mathbb{TG}$ to discover the genes that are likely driving the interaction between the same pair of taxa. Finally, the network from $\mathbb{TG}$ ($G_{\mathbb{TG}}$) or $\mathbb{TM}$ ($G_{\mathbb{TM}}$) can be unrolled using $G_{\mathbb{TGM}}$ to find fully unrolled chains of the form $T_i \rightarrow G_y \rightarrow M_x \rightarrow T_j$ in $G_{\mathbb{TGM}}$ with the capability to simultaneously explain the edges $T_i \rightarrow T_j$ in $G_{\mathbb{T}}$, the chain $T_i \rightarrow M_x \rightarrow T_j$ in $G_{\mathbb{TM}}$, and the chain $T_i \rightarrow G_y \rightarrow T_j$ in $G_{\mathbb{TG}}$.

This step-wise unrolling is necessary to discover unrollings, where the network learned from $\mathbb{T}$ was unrolled in a network learned from some subset of $\{\mathbb{TG}, \mathbb{TM}, \mathbb{TGM}\}$. The number of the networks from $\{\mathbb{TG}, \mathbb{TM}, \mathbb{TGM}\}$ that support the unrolling provide a degree of confidence for that unrolling. Furthermore, the bootstrap score for each of the edges involved in the process is reported, together with an overall score that is computed as the product of the individual bootstrap scores of the two replacement edges.

7.2.7 De-confounding

Most current causal inference techniques rely on the *causal sufficiency* assumption, which assumes that there are no hidden confounders (for any pair of variables) in the data. These are variables that are either (a) unknown, (b) known but not measured, or (c) measured but not used in the analysis, but affect both the cause and the effect of at least one predicted interaction. Predictions of interactions with hidden confounders could be incorrect. The predicted interaction may be enhanced or diminished when the hidden confounder is not used in the analysis. It is also possible that the predicted interaction may introduce spurious edges when the hidden confounder is not used in the analysis.

In general, the causal sufficiency assumption may be "too strong" and may be impossible to verify, even with the availability of richer data sets that include multi-omics data, thus making this assumption a key obstacle to performing accurate causal inference [4]. Going beyond the multi-omic domain, causal sufficiency is an assumption that does not strictly hold in most observational datasets, since it is difficult or impossible to include all possible explanatory variables in a study.

A recent paper by Wang and Blei [193] attempts to perform **de-confounding**, which is the process of removing the effect of *all* confounders. They introduce the concept of "substitute confounders", which attempts to account for the effect of all hidden confounders in order to arrive at unbiased estimates of causal effects. Note that one of the limitations of their method is that the de-confounded interactions are not identified, which is something

that would be of interest. Furthermore, there may not be a one-to-one correspondence between the substitute confounder and some real confounder, meaning that one substitute confounder may be an approximation for a combination of several hidden confounders.

In this chapter we take on a different approach for the task of de-confounding interactions and is inspired by the unrolling approach of Section 7.2.6. We iteratively learn independent networks with different subsets of data with the hope that by adding a new omics layer we would be able to identify some of the problematic variables and interactions. As before, we let $G_{\mathbb{X}}, V_{\mathbb{X}}$, and $E_{\mathbb{X}}$ represent the graph, the vertex set and the edge set, respectively, of the network learned using dataset $\mathbb{X}$. By learning a network with the $\mathbb{T}$ and $\mathbb{TM}$ datasets, we can de-confound interactions if the following three conditions are satisfied:

1. There is an edge between T_i and T_j in $G_{\mathbb{T}}$, for some $T_i, T_j \in V_{\mathbb{T}}, i \neq j$.
2. There is *no* edge between T_i and T_j in the network $G_{\mathbb{TM}}$.
3. The edges from M_x to T_i and from M_x to T_j exist in $G_{\mathbb{TM}}$ for some metabolite $M_x \in V_{\mathbb{TM}}$.

Using this method, if the above conditions are satisfied for a pair of taxa, T_i and T_j, we can deduce that the directed edge (T_i, T_j) in $G_{\mathbb{T}}$ and the inferred interaction between the two taxa were spurious, and that the metabolite M_x was the responsible confounder. We can also infer that the metabolite impacts the abundance of both taxa, T_i and T_j. One possible scenario is that the metabolite, M_x, could be an essential metabolite for both taxa, and its absence from the analysis could make their abundance appear correlated.

As before, this process can be repeated by de-confounding $G_{\mathbb{T}}$ with edges from $G_{\mathbb{TG}}$ to discover genes/proteins that could nullify a causal connection between the taxa. In general, the networks learned using the $\mathbb{T}, \mathbb{G}$, and/or $\mathbb{M}\}$ datasets can be de-confounded by networks learned using one or more of the datasets from $\{\mathbb{TG}, \mathbb{TM}, \mathbb{GM}, \mathbb{TGM}\}$. Similarly, networks learned using one of $\mathbb{TG}, \mathbb{TM}$, or $\mathbb{GM}\}$ datasets can be de-confounded by networks learned

using TGM. This could lead to chains of de-confoundings when an interaction that led to the de-confounding a relationship is itself later de-confounded.

As before, for each de-confounding discovery, we report (a) the confounded edge, (b) the de-confounder, (c) the bootstrap score for the edges involved in the discovery, (d) the overall score of the discovery computed as the product of the individual bootstrap scores of the two replacement edges, and (e) the two datasets that were used to discover the specific de-confounding.

7.2.8 Validation

Validations of DBN edges that are directed from taxon to gene, taxon to metabolite, or gene to metabolite are handled by verifying the information against existing databases of genomes, genes, and metabolic pathways (as outlined in Section 6.2.7). To assist in the validation of taxa-metabolite ($T \rightarrow M$) edges in our networks, we relied on the tool MIMOSA [117], which calculates the capability of a species to produce a metabolite under the conditions of the data set. For taxa-gene ($T \rightarrow G$) validations, we used the KEGG database to build a validation database of bacterial taxa and the genes present in their genomes. The one-time creation of a local validation database also speeded up our computations considerably. We evaluated our results using the precision metric. We calculated the statistical significance of validated interactions by comparing the validation statistics to a null model using a Poisson-Binomial distribution test.

7.3 Results

A large number of networks were learned with the different data subsets, modified methods, and parameter settings as mentioned in Sections 7.2.3, 7.2.4, 7.2.5 respectively for DBN, TETRAD, and Tigramite. We implemented unrolling and de-confounding, and applied them to all the learned networks. Results are presented below.

7.3.1 Resulting networks

Figure 7.1 shows the DBNs learned from the $\mathbb{T}$, $\mathbb{TM}$, $\mathbb{TG}$ datasets, and $\mathbb{TGM}$ without alignment. Self loops were hidden to avoid unnecessary clutter. The remarkable information gain obtained by using additional omics data sets is readily observable in Figure 7.1 d), with a more complete picture of the state of the whole system. The one non-omics variable (week of sample obtained), which we generically refer to as a "clincal variable" did not have a strong enough effect in $\mathbb{TG}$, but it did have an effect in the other networks.

7.3.2 Tool analysis

In addition to analyzing the networks, we also explored the effect of the different network parameters. The heatmap of Figure 7.2 shows the percentage of unrolling taking place in the networks learned by PyCausal (TETRAD). The first three columns represent the percentage of taxon to taxon interactions in the network learned with $\mathbb{T}$ that got unrolled with the networks learned with $\mathbb{TGM}$, $\mathbb{TG}$, and $\mathbb{TM}$ respectively. It is obvious that as the alpha parameter decreases, the percentage of unrolling cases drastically decreases. The smaller the alpha, the easier for two variables is to be dependent, so the more edges the network has. This means that for a bigger alpha, the average confidence on each edge should be higher, since it is more difficult for it to get learned by chance. This is in accordance with the higher percentage of unrolling, indicating that the edges with higher support get unrolled more frequently, adding support for the unrolling process. Interestingly, there is a clear reversal of the pattern for the overall bootstrap score (last column) for no-aligment, where the smaller the alpha, the larger the overall score is, which would seem to contradict our intuition. Interestingly, aligning the dataset seems to fix this problem, which would support the necessity of alignment as a pre-processing step.

Figure 7.3 shows the average percentage of unrolling taking place in the different methods, averaged over all parameters. The first three columns represent the percentage of taxon to taxon interactions in the network learned with $\mathbb{T}$ that got unrolled with the networks

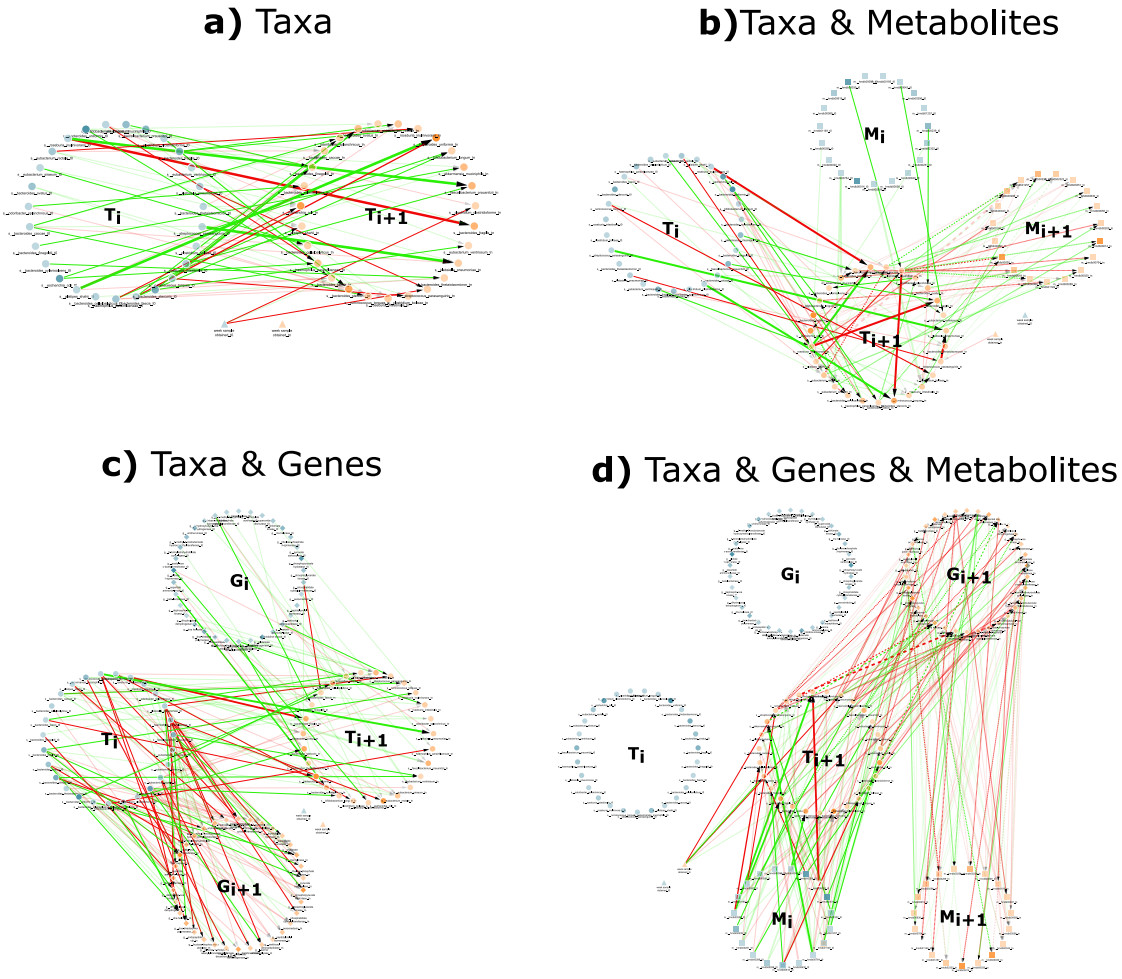

Figure 7.1: The **two-time-slice DBN networks for four different multi-omic subsets, hiding self-edges**. Each network was learned with a maximum number of parents of 3, and has two versions of each node organized in large circles, one representing the variable for the current time point (blue) and the other for the next time point (orange). Taxa nodes are represented as filled circles, metabolites as filled squares, genes as filled diamonds, and clinical variables as filled triangles. Red (green) edges represent negative (positive resp.) regression coefficients. Edge width is proportional to the regression coefficient and edge opacity to the bootstrap score. Finally, node opacity is proportional to abundance. a) DBN learned with just taxa abundance (𝕋). The dataset included abundance of 27 bacteria and a clinical variable indicating the week the sample was obtained and resulted in a network with 95 edges. b) DBN learned with taxa and metabolites (𝕋𝕄). A set of 19 metabolites were added to the previous dataset, and 164 edges were learned in this network. c) DBN learned with the taxa and genes dataset (𝕋𝔾). A set of 34 genes were added to the taxa dataset, and a network with 230 edges was learned. d) DBN learned with the 27 taxa, 34 genes, and 19 metabolites (𝕋𝔾𝕄), resulting in a total of 311 edges.

	%UnrolledTGMT	%UnrolledTGT	%UnrolledTMT	overallScore
PyCausal_NoAlignment_a0.01	0.76984127	0.658730159	0.666666667	0.018728972
PyCausal_NoAlignment_a0.001	0.274725275	0.604395604	0.450549451	0.023861538
PyCausal_NoAlignment_a0.0001	0.057971014	0.391304348	0.333333333	0.060310204
PyCausal_Alignment_a0.01	0.723684211	0.710526316	0.157894737	0.039487963
PyCausal_Alignment_a0.001	0.288461538	0.75	0.230769231	0.054606452
PyCausal_Alignment_a0.0001	0.116666667	0.416666667	0.35	0.047136

Figure 7.2: **PyCausal** (TETRAD) **network unrolling analysis for alignment and no-alignment as the alpha parameter varies.** The heatmap contains information for the percentage of unrolling happening in each of the parameter configurations, together with the overall bootstrap score.

learned with TGM, TG, and TM respectively. Tigramite unrolls a larger percentage with TG and TM than the other two methods, but falls short when unrolling with TGM, where DBN does a better job. Interestingly, alignment seems to drastically improve the unrolling results for our DBN method for TGM going from 24.7% to 78.8%, and mildly helps for the other two datasets. Also, our DBN method seems more stable than the other two, since the much higher average overall bootstrap score indicates that in each bootstrap, the edges learned are consistent with the ones learned in other bootstrap runs. This lower variability across the different random data subsamples used is a clear advantage of our DBN method.

	%UnrolledTGMT	%UnrolledTGT	%UnrolledTMT	overallScore
PyCausal_NoAlignment	0.36751252	0.551476704	0.483516484	0.03430024
PyCausal_Alignment	0.376270805	0.625730994	0.246221323	0.0470768
Tigramite_NoAlignment	0.2	0.721967081	0.944911205	0.01151014
Tigramite_Alignment	0.2	0.666270884	0.918569277	0.01095397
DBN_NoAlignment	0.247210667	0.389020815	0.213623733	0.46395321
DBN_Alignment	0.787923472	0.42515232	0.334329358	0.37356303

Figure 7.3: **Unrolling percentages for all methods averaging over all different parameters.** The heatmap contains information for the percentage of unrolling happening in each of the methods, together with the overall bootstrap score.

7.4 Discussion

The top unrollings and de-confoundings discovered by all the methods were sorted by the combined overall bootstrap score, and other factors like the number of networks they

appear in, or the different network types that supported this particular finding. Below we discuss some particularly interesting results from our analysis above.

7.4.1 Uncovering unrolled biological relationships

The unrolling of the an edge *Eubacterium siraeum* → *Bacteroides thetaiotaomicron* in $G_{\mathbb{T}}$, manifests itself as the unrolled path *Eubacterium siraeum* → uridine kinase → cytidine → *Bacteroides thetaiotaomicron* in $G_{\mathbb{TGM}}$, as shown in Figure 7.4. Interestingly,

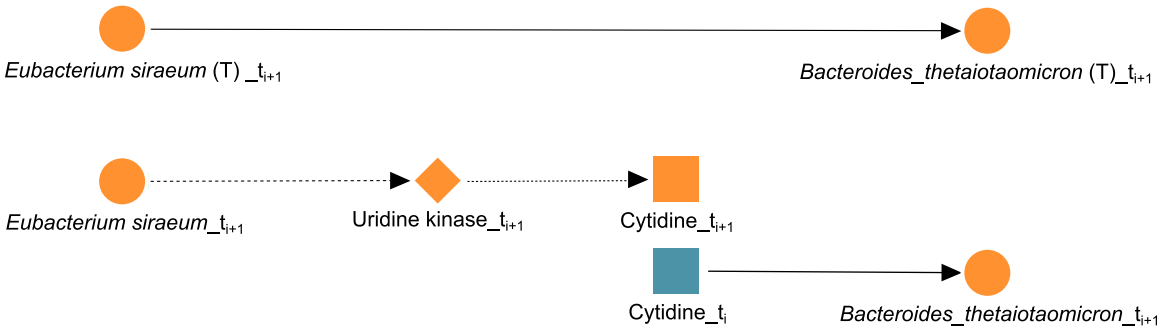

Figure 7.4: **Biologically confirmed unrolling**. The edge *Eubacterium siraeum* → *Bacteroides thetaiotaomicron* learned in $G_{\mathbb{T}}$ (T) is unrolled into *Eubacterium siraeum* → uridine kinase → cytidine → *Bacteroides thetaiotaomicron* in $G_{\mathbb{TGM}}$

each edge in the unrolled path was validated in the literature and the knowledge bases. Both *E. siraeum* and *B. thetaiotaomicron* contain the enzyme uridine kinase [83, 82]. This enzyme can be commonly found in prokaryotes and eukaryotes, and phosphorylates both uridine and cytidine to their mono-phosphate forms, and vice-versa. The specific reactions that this enzyme is capable of performing are the following [186, 123, 166]:

- ATP + Uridine $\Longleftrightarrow$ ADP + UMP
- ATP + Cytidine $\Longleftrightarrow$ ADP + CMP,

where ATP stands for adenosine tri-phosphate, ADP stands for adenosine di-phosphate, UMP stands for uridine mono-phosphate, and CMP stands for cytidine mono-phosphate. Since *B. thetaiotaomicron* also contains uridine kinase, it has the ability to perform the forward reaction and consume it by phosphorylating cytidine to CMP. More importantly,

B. thetaiotaomicron also contains cytidine deaminase, which scavenges exogenous and endogenous cytidine for UMP synthesis [185]. This reaction performed by this enzyme is cytidine + H2O $\Longleftrightarrow$ uridine + Ammonia [189, 168, 192], which validates the third and last edge (cytidine $\rightarrow$ *B. thetaiotaomicron*). In addition, experimental results show that a cytidine-scavenging system confers colonization fitness to *B. thetaiotaomicron*, and therefore positively impact its abundance [59]. Interestingly, uridine may be playing a role in this connection between the two taxa, since both enzymes discussed involve uridine, so both taxa can produce and consume uridine. Reinforcing this argument is the fact that the edge uridine $\rightarrow$ *B. thetaiotaomicron* is also present in the same network $G_{\mathbb{TGM}}$. Moreover, this unrolling can be important for IBD. Treatment for Crohn's disease with live *B. thetaiotaomicron* or its products displays strong efficacy in preclinical models of IBD, with multiple benefits [43]. Similarly, there is precedent to treat gastrointestinal problems with *E. Siraeum* [17], and activation-induced cytidine deaminase seems to prevent colon cancer development despite persistent inflammation in the colon [176].

In summary, our unrolling methods allow us to make biological sense out of a set of related edges in the series of networks generated from the multi-omics data.

As a second example, we can also validate the circular path: *Bacteroides stercoris* $\rightarrow$ uridine kinase $\rightarrow$ cytidine $\rightarrow$ *Bacteroides stercoris*, which can be thought of as an unrolling of the self-loop from *Bacteroides stercoris* to itself in $G_{\mathbb{T}}$ as shown in Figure 7.5. *B. stercoris* contains both uridine kinase [115] and cytidine deaminase [114], so it can both produce and consume cytidine, and since cytidine deaminase can scavenge endogenous cytidine, this lends further support to the self-loop edge from *B. stercoris* to itself; it might be regulating itself through the cytidine or uridine internally. Interestingly, *Bacteroides stercoris* is linked to colorectal cancer [91], and an increased abundance of *B. stercoris* was detected in fecal samples of Crohn's Disease (CD) patients [190]. Also, there is an increased reactivity of Immunoglobulin G from Crohn's Disease patients toward *B. stercoris* and other species of *Bacteroides* in the serum of CD patients [80].

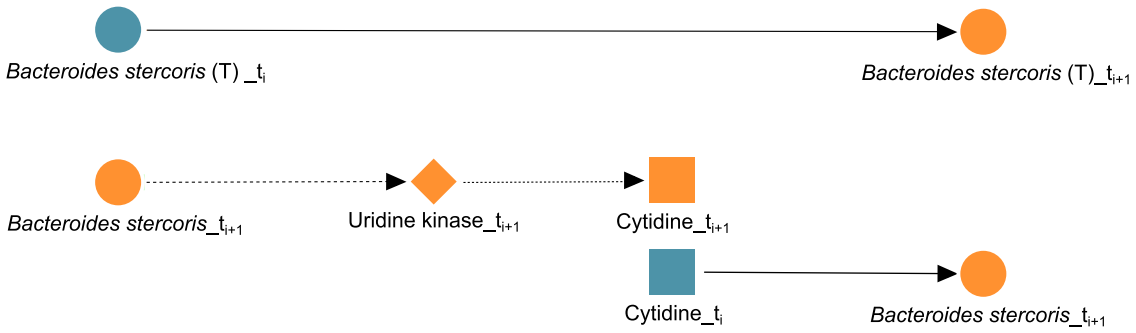

Figure 7.5: **Biologically confirmed unrolling**. The edge *Bacteroides stercoris* → *Bacteroides stercoris* learned in $G_{\mathbb{T}}$ (T) is unrolled into *Bacteroides stercoris* → uridine kinase → cytidine → *Bacteroides stercoris* in $G_{\mathbb{TGM}}$

We provide two examples of partial unrollings from our experiments. The unrolled path *Bacteroides finegoldii* → phosphatidate cytidylyltransferase → Betaine → *Eubacterium ventriosum* was discovered by our search. It first appeared as an edge *Bacteroides finegoldii* → *Eubacterium ventriosum* in $\mathbb{T}$, which then got unrolled in $\mathbb{TG}$, $\mathbb{TM}$, and $\mathbb{TGM}$. *Bacteroides finegoldii* is an anaerobic gram-negative bacteria that has been found to be generally beneficial in the gut [5]. It contains the gene BN532_01044 which expresses the phosphatidate cytidylytransferase protein. This is a membrane-bound enzyme that participates in the glycerophospholipid metabolism and phosphatidylinositol signaling system. Moreover, *Bacteroides finegoldii* is known to produce the metabolite Betaine [37]. Increased levels of betaine have been found to benefit IBD patients, allowing for proper digestion and assimilation of nutrients. Over the last decade, doctors have recommended betaine-rich foods as a way to help IBD patients rapidly absorb and distribute vital vitamins and minerals needed to maintain diversity in the gut [37]. Additionally, recent studies have shown betaine to be correlated to the *Eubacterium* genus and to be of general importance for osmotic adaptation of most species of *Eubacterium* [73]. Even though no specific study was found about the species *Eubacterium ventriosum*, the fact that betaine was found to increase the abundance of the *Eubacterium* genus lends support to the argument that *Eubacterium* members consume betaine through the conversion of Acetate [195], thus validating the unrolling. Moreover, while Acetate was not contemplated in the dataset,

one of its precursors, Choline, was. Many strong unrollings have a link from Choline to a member of the *Eubacterium* genus in the dataset (*E. ventriosum, E. siraeum, E. rectale*), and almost every method learned the edge Betaine → *E. ventriosum* as part of specific unrollings, which could be an indication of a pathway transforming Choline to Acetate to Betaine, which may be facilitated by the taxon, *Eubacterium*.

The path: *Bacteroides ovatus* → DNA helicase → Pyridoxine → *Bacteroides ovatus* in $\mathbb{TGM}$ can be thought of as an unrolling of a self-loop edge in $\mathbb{T}$ from *Bacteroides ovatus* to itself, which got unrolled in $\mathbb{TG}$, $\mathbb{TM}$, and $\mathbb{TGM}$. Moreover, *Bacteroides ovatus* is present in the gut microbiome, and plays a crucial role in the dysbiosis of the gut health. This anerobic bacteria has been found to be significantly elevated in abundance in patients suffering from IBD. Findings suggest that some species of *Bacteroides* injure gut tissue and induce inflammation [149]. This bacteria does contain the gene dnaB which expresses the protein DNA helicase, an enzyme responsible in unpacking genes in an organism and DNA repair. The production of the metabolite pyridoxine has been found in great proportion when there is an abundance of *Bacteroides ovatus* [160]. However, evidence suggesting the consumption of pyridoxine by the taxa could not be found. When pyriodoxine is present in great abundance, it is involved in many biochemical pathways that lead to the synthesis or metabolism of nucleic acids, immune modulatory metabolites and many others [160]. However, when scarce, it leads to inflammation.

7.4.2 Uncovering de-confounded biological relationships

The edge: thymidylate synthase → glutamate dehydrogenase was inferred in the $\mathbb{G}$ network but disappeared in the $\mathbb{TG}$ network, possibly because both genes are present in the taxon *haemophilus parainfluenzae*. This suggests that the relationship between the two genes is spurious and the taxon is the confounder. *Haemophilus parainfluenza* is an opportunistic pathogen that has been found in elevated levels in patients suffering from many diseases including pneumonia and conjunctivitis. Recent studies have shown that high

abundance of this pathogen was found in patients suffering from IBD. Different dynamics have been noted for the abundance of *haemophilus parainfluenza* in the literature. For instance, when IBD patients enter remission, there is a steep decline in this pathogen [154]. Additionally, the two genes that are present in *haemophilus parainfluenzae* were found to produce proteins that help drive diseases including colon cancer.

7.4.3 Limitations and future work

The main limitation of the methods described in this chapter is that they are only applicable to multi-omic datasets, which are relatively uncommon. We, however, expect this to change with the increased effort to understand the underlying mechanisms within biological processes. Secondly, these methods do not provide definitive evidence for the causal chains, but rather lend support to generate hypotheses that would have to be proved with experiments in the laboratory. Also, as larger datasets become available these methods will become increasingly useful. Regarding the future work, the following are research directions that are currently being taken:

- Logically, since the three methods use very different approaches, any edge that is confirmed by more than one method is noteworthy. The validation part could be combined with the other two methods. We are working on combining validation with unrolling and de-confounding, with the intention of testing if the unrolled or de-confounded interactions get validated more frequently. This would be consistent with our hypothesis that interactions inferred by more than one method have a higher chance of being truly causal than others. Also, we can use some of the different network-inferring algorithms to validate the others.
- We are working on combining de-confounding and unrolling in the following synergistic manner: We can start with the $\mathbb{T}$ network and eliminate from it all edges that were de-confounded with $\mathbb{TM}$ and $\mathbb{TG}$. Then, we could perform unrolling with-

out these "spurious" edges in a way that we hope would lead to more meaningful discoveries.

- Y-structures are causally fascinating subgraphs in the sense that one of their edges (the Y-leg) cannot have been confounded. We plan on integrating the *post facto* discovery of these structures in the networks learned by the methods we utilized, and explore the Y-structure unconfoundability claim.
- It would be useful to make the three main methods of this chapter (unrolling, de-confounding, and validating) available to a wider audience. The software PluMA [29] would allow us to develop them in the form of plugins, that could then be integrated into a multitude of bioinformatic pipelines.
- Finally, our largest work-in-progress contribution would be the creation of a holistic novel validation score that could be applied to compare and evaluate biological networks. This score would be a combination of the interaction validation precision, probability, and their average derivatives as a function of a changing bootstrap score threshold. We then would also weight the impact of the amount of unrolling and de-confounding that took place for those networks, and other metrics calculated in this chapter. This will give us a "confidence" measure for the biological validity of the claims made by the network, and allow us to compare different learning algorithms from a biological point of view.

7.5 Conclusion

We have developed three novel biological network analysis algorithms, namely unrolling, de-confounding, and validating. We learned biological networks based on a longitudinal multi-omic IBD dataset, with three state-of-the-art network and causal inference tools. We then applied the three algorithms (unrolling, de-confounding, and validating) to the networks learned by the tools (DBNs, tsGFCI, and Tigramite), and compared their pre-

dictive performance. The top findings by our algorithms were then analyzed, and interesting biological interpretations were found for several of the network-inferred interactions.

CHAPTER 8

CONCLUSIONS

Typical microbiome studies focused primarily on static metagenomics sequence data . More recent data sets are much richer, notably including time series information of host and microbial gene expression, and metabolomics data. The ability to integrate and learn models using these multi-omics, longitudinal data, the inference of direct causal interactions between the omics entities, validating these interactions, and drawing useful biological interpretations remain major challenges for microbiome analysis.

8.1 Longitudinal microbial network inference

In Chapter 5, we developed PRIMAL, a computational pipeline that enables the integration of data across individuals for the reconstruction of dynamic models from time series microbiome data. First, we smoothed each time series using b-splines and the interpolated them, thus addressing irregular sampling rates and missing timepoints. Then our pipeline aligned the data collected for all individuals, to adjust for the different metabolic speeds of each individual and to align the internal biological processes in the different subjects with each other. The aligned profiles were then used to learn a dynamic Bayesian network with the expectation that it represents temporal causal relationships between taxa and clinical variables. We improved the open source CGBayesNets package [106, 108] by adding the capability to learn intra edges (within the same time slice) and implemented better performing network scoring functions to address overfitting, such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). We also implemented a custom network robust to artifacts generated by large differences in the values of the child and parent nodes, and visualization capabilities. Testing our methods on three longitudinal microbiome data sets (infant gut, vagina, and oral cavity) we showed that our pipeline improved upon prior methods developed for this task. We also discussed the biological

insights provided by the models which shed light on several known and novel interactions, in addition to finding interesting interaction relationships.

8.2 Longitudinal multi-omic network inference (PALM)

A key challenge in the analysis of longitudinal microbiome data is the inference of causal interactions between microbial taxa, their genes, the metabolites they consume and produce, and the host genes that are expressed in the environmental niche. To address these challenges, in Chapter 6 we developed a computational pipeline called PALM that first temporally aligned the multi-omics data and then used dynamic Bayesian networks (DBNs) to construct a unified model. Our approach overcame differences in sampling and progression rates and reduced the large number of entities and parameters in the DBNs. It also utilized a biologically-inspired, customizable multi-omic framework, which was provided as an input to the algorithm, and ensured that the inferred interactions represented a flow of information that is consistent with biological realities, prevented spurious connections and greatly reduced computational costs. Moreover, a set of *in silico* validation methodologies were developed to consult with existing databases and help evaluate the biological validity of edges that suggest directed interactions from taxa to genes and taxa to metabolites. Applying PALM to data collected from inflammatory bowel disease (IBD) patients, we showed that it accurately identified known and novel interactions. Targeted experimental validations further supported a number of the predicted novel metabolite-taxa interactions. Moreover, we showed how PALM can be extended to infer not only taxa abundance, but also other types of omics datasets, greatly improving over the state-of-the-art algorithms such as MMvec for metabolite prediction [111], and MTPLasso for taxa prediction [94].

8.3 Longitudinal causal multi-omic network inference

In an effort to improve the state of the art in inferring meaningful multi-omic interactions, in Chapter 7 we addressed some of the most fundamental issues in causal inference.

We developed METALICA, a suite of tools and techniques that inferred strong interactions between microbiome entities. We also developed and applied novel *unrolling* and *de-confounding* techniques to uncovered multi-omic entities that are believed to act as confounders for some of the inferred relationships, thereby lending support for a biological model and process by which two taxa interact with each other. The *unrolling* process helps to find intermediaries to explain interactions, while the *de-confounding* process finds common causes that causes spurious relationships to be inferred. Finally, we showed how to automatically validate such inferences using ground truth databases. We applied our methods to networks learned by causal algorithms such as Tigramite [147] and tsGFCI [101], which we augmented with our restriction framework and alignment techniques, among other improvements. The dataset used was an IBD multi-omic dataset, and the findings were used to compare the inferences of the various methods against the ones of PALM. The top unrollings and de-confoundings were compared against the literature, and partial or total validation for some of them was found.

The problem of holistically analyzing dynamic microbiomes had not been addressed. The work in this dissertation makes a sizable dent in the challenging problem of studying, exploring, and understanding data sets generated by longitudinal microbiome studies. This dissertation also generates a suite of valuable tools for this work, thus addressing the lack of accurate tools for studying dynamic microbiomes and investigating the interactions between their entities.

8.4 Future work

This dissertation addressed several fundamental problems in multi-omics microbiome studies. However, other challenges in this domain still remain unresolved, from both the computational and biological points of view. These challenges span the whole end-to-end pipeline of analysis, from the data collection, study and algorithm design, implementation,

and the validation and usage of the conclusions. Some natural extensions of this dissertation are the following:

8.4.1 Extend the causal tools developed

The application of interventional techniques would help us determine the average causal effect of one entity on any other entity in our dataset. But these effects may not apply to a particular individual. Applying counterfactual theory would allow us to understand how the different entities interact and affect each other at the patient level. There are still many setbacks before this can be applied, but it would open the door to personalized medicine that not only depends on the genome of the patient, but also the person's current microbiome.

8.4.2 Develop a microbiome model to perform causal reasoning

While microbiome analysis can help identify one or more specific microbial taxa within the microbiome as the reason for the condition of a patient, any attempt to address the dysbiosis will also need to take into account the entire web of interactions between the taxa in the microbiome. Otherwise, unexpected side-effects are prone to occur and interventions may not right the dysbiosis. Our vision of a futuristic solution for any dysbiosis is to initiate a sequence of steps that would lead to a new homeostasis in a state that corresponds to a healthy symbiosis. It would allow us to perform personalized reasoning about possible treatments. We believe this to be the future of personalized medicine.

8.4.3 Learning models with even more variables

Ideally, we would like to get a holistic perspective of each subject by integrating as many omics data sets as possible. However, because of the high dimensional nature of these data sets we are facing, several challenges remain unsolved. First, because conditional independence tests available are not reliable with an insufficient sample size, new theoretical frameworks need to be developed. Second, since the structure learning with a

large number of samples and variables is computationally very expensive, the new methods being developed need to take advantage of high performance computing and cloud computing infrastructures.

8.4.4 Address all compositionality problems

One of the main problems with microbiome studies is the inherent compositionality of the data. This could be solved by either developing sequencing techniques that could yield absolute abundance values, or mathematical frameworks and normalizations techniques to address this issue in a fundamental manner. Every step of any microbiome analysis pipeline would benefit from either of these.

BIBLIOGRAPHY

[1] Aguiar-Pulido, V., Huang, W., Suarez-Ulloa, V., Cickovski, T., Mathee, K., and Narasimhan, G. Metagenomics, metatranscriptomics, and metabolomics approaches for microbiome analysis: supplementary issue: bioinformatics methods and applications for big metagenomics data. *Evolutionary Bioinformatics 12* (2016), EBO–S36436.

[2] Akaike, H. A new look at the statistical model identification. *IEEE transactions on automatic control 19*, 6 (1974), 716–723.

[3] Anderson, D. HABs in a changing world: a perspective on harmful algal blooms, their impacts, and research and management in a dynamic era of climactic and environmental change. *Harmful Algae 10* (2012), 3–17.

[4] Aurora, R. Confounding factors in the effect of gut microbiota on bone density. *Rheumatology* (2019).

[5] Bakir, M. A., Kitahara, M., Sakamoto, M., Matsumoto, M., and Benno, Y. Bacteroides finegoldii sp. nov., isolated from human faeces. *International journal of systematic and evolutionary microbiology 56*, 5 (2006), 931–935.

[6] Baksi, K. D., Kuntal, B. K., and Mande, S. S. 'time': a web application for obtaining insights into microbial ecology using longitudinal microbiome data. *Frontiers in microbiology 9* (2018), 36.

[7] Bar-Joseph, Z., Gerber, G. K., Gifford, D. K., Jaakkola, T., and Simon, I. Continuous representations of time-series gene expression data. *J Comput Biol 10*, 3–4 (2003), 341–356.

[8] Bar-Joseph, Z., Gerber, G. K., Gifford, D. K., Jaakkola, T. S., and Simon, I. Continuous representations of time-series gene expression data. *Journal of Computational Biology 10*, 3-4 (2003), 341–356.

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

[10] Bartels, R. H., Beatty, J. C., and Barsky, B. A. *An introduction to splines for use in computer graphics and geometric modeling*. Morgan Kaufmann, 1995.

[11] Bartlett, M., and Cussens, J. Integer linear programming for the bayesian network structure learning problem. *Artificial Intelligence 244* (2017), 258–271.

[12] Bashiardes, S., Zilberman-Schapira, G., and Elinav, E. Use of metatranscriptomics in microbiome research. *Bioinform Biol Insights 10* (2016), BBI–S34610.

[13] Beal, M. J., Falciani, F., Ghahramani, Z., Rangel, C., and Wild, D. L. A bayesian approach to reconstructing genetic regulatory networks with hidden factors. *Bioinformatics 21*, 3 (2005), 349–356.

[14] Beale, D. J., Karpe, A. V., and Ahmed, W. Beyond metabolomics: a review of multi-omics-based approaches. In *Microbial metabolomics*. Springer, Cham, 2016, pp. 289–312.

[15] Berry, A. C. The accuracy of the gaussian approximation to the sum of independent variates. *Transactions of the american mathematical society 49*, 1 (1941), 122–136.

[16] Boekel, J., Chilton, J. M., Cooke, I. R., Horvatovich, P. L., Jagtap, P. D., Käll, L., Lehtiö, J., Lukasse, P., Moerland, P. D., and Griffin, T. J. Multi-omic data analysis using Galaxy. *Nat Biotechnol 33*, 2 (2015), 137–139.

[17] Borody, T. J. Treatment of gastro-intestinal disorders. World Intellectual Property Organization, (1990) WO1990001335A1.

[18] Bouckaert, R. R. Probabilistic network construction using the minimum description length principle. In *European conference on symbolic and quantitative approaches to reasoning and uncertainty* (1993), Springer, pp. 41–48.

[19] Boyer, H. W., and Roulland-Dussoix, D. A complementation analysis of the restriction and modification of dna in Escherichia coli. *J Mol Biol 41*, 3 (1969), 459–472.

[20] Bray, J. R., and Curtis, J. T. An ordination of the upland forest communities of southern wisconsin. *Ecological monographs 27*, 4 (1957), 326–349.

[21] Buntine, W. Theory refinement on bayesian networks. In *Uncertainty Proceedings 1991*. Elsevier, 1991, pp. 52–60.

[22] Canzler, S., Schor, J., Busch, W., Schubert, K., Rolle-Kampczyk, U. E., Seitz, H., Kamp, H., von Bergen, M., Buesen, R., and Hackermüller, J. Prospects and challenges of multi-omics data integration in toxicology. *Arch Toxicol* (2020), 1–18.

[23] Castro-Nallar, E., Shen, Y., Freishtat, R. J., Pérez-Losada, M., Manimaran, S., Liu, G., Johnson, W. E., and Crandall, K. A. Integrating microbial and host transcriptomics to characterize asthma-associated microbial communities. *BMC Med Genomics 8*, 1 (2015), 50.

[24] Chickering, D. M. A transformational characterization of equivalent bayesian network structures. *arXiv preprint arXiv:1302.4938* (2013).

[25] Chirayul. py-causal. `https://github.com/bd2kccd/py-causal`, 2016.

[26] Cho, I., and Blaser, M. J. The human microbiome: at the interface of health and disease. *Nat Rev Genet 13* (2012), 260–270.

[27] Chow, C., and Liu, C. Approximating discrete probability distributions with dependence trees. *IEEE transactions on Information Theory 14*, 3 (1968), 462–467.

[28] Chung, M., Krueger, J., and Pop, M. Identification of microbiota dynamics using robust parameter estimation methods. *Math Biosci 294* (2017), 71–84.

[29] Cickovski, T., and Narasimhan, G. Constructing lightweight and flexible pipelines using plugin-based microbiome analysis (pluma). *Bioinformatics 34*, 17 (2018), 2881–2888.

[30] cmu phil. Tetrad. `https://github.com/cmu-phil/tetrad`, 2015.

[31] Colombo, D., and Maathuis, M. H. A modification of the pc algorithm yielding order-independent skeletons. *arXiv preprint arXiv:1211.3295* (2012).

[32] Colombo, D., Maathuis, M. H., Kalisch, M., and Richardson, T. S. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics* (2012), 294–321.

[33] Cooper, G. F. The computational complexity of probabilistic inference using bayesian belief networks. *Artificial intelligence 42*, 2-3 (1990), 393–405.

[34] Cooper, G. F., and Herskovits, E. A bayesian method for the induction of probabilistic networks from data. *Machine learning 9*, 4 (1992), 309–347.

[35] Cowell, R. G., Dawid, P., Lauritzen, S. L., and Spiegelhalter, D. J. *Probabilistic networks and expert systems: Exact computational methods for Bayesian networks*. Springer Science & Business Media, 2006.

[36] Cowell, R. G., Lauritzen, S. L., and Mortera, J. Maies: A tool for dna mixture analysis. *arXiv preprint arXiv:1206.6816* (2012).

[37] Craig, S. A. Betaine in human nutrition. *The American journal of clinical nutrition 80*, 3 (2004), 539–549.

[38] Dagum, P., Galper, A., and Horvitz, E. Dynamic network models for forecasting. In *Uncertainty in artificial intelligence* (1992), Elsevier, pp. 41–48.

[39] Dagum, P., Galper, A., Horvitz, E., and Seiver, A. Uncertain reasoning and forecasting. *International Journal of Forecasting 11*, 1 (1995), 73–87.

[40] Dagum, P., and Luby, M. Approximating probabilistic inference in bayesian belief networks is np-hard. *Artificial intelligence 60*, 1 (1993), 141–153.

[41] De Campos, L. M., Fernández-Luna, J. M., and Puerta, J. M. Local search methods for learning bayesian networks using a modified neighborhood in the space of dags. In *Ibero-American Conference on Artificial Intelligence* (2002), Springer, pp. 182–192.

[42] de Luis Balaguer, M. A., Fisher, A. P., Clark, N. M., Fernandez-Espinosa, M. G., Möller, B. K., Weijers, D., Lohmann, J. U., Williams, C., Lorenzo, O., and Sozzani, R. Predicting gene regulatory networks by combining spatial and temporal gene expression data in arabidopsis root stem cells. *Proc Natl Acad Sci 114*, 36 (2017), E7632–E7640.

[43] Delday, M., Mulder, I., Logan, E. T., and Grant, G. Bacteroides thetaiotaomicron ameliorates colon inflammation in preclinical models of crohn's disease. *Inflammatory bowel diseases 25*, 1 (2019), 85–96.

[44] Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological) 39*, 1 (1977), 1–22.

[45] DiGiulio, D. B., Callahan, B. J., McMurdie, P. J., Costello, E. K., Lyell, D. J., Robaczewska, A., Sun, C. L., Goltsman, D. S. A., Wong, R. J., Shaw, G., Stevenson, D. K., Holmes, S. P., and Relman, D. A. Temporal and spatial variation of the human microbiota during pregnancy. *Proc Natl Acad Sci 112*, 35 (2015), 11060–11065.

[46] Eberhardt, F. Introduction to the foundations of causal discovery. *International Journal of Data Science and Analytics 3*, 2 (2017), 81–91.

[47] Entner, D., and Hoyer, P. O. On causal discovery from time series data using fci. *Probabilistic graphical models* (2010), 121–128.

[48] Eubank, R. L. *Nonparametric regression and spline smoothing*. CRC press, 1999.

[49] Fabres, P. J., Collins, C., Cavagnaro, T. R., and Rodríguez-López, C. M. A concise review on multi-omics data integration for terroir analysis in vitis vinifera. *Front Plant Sci 8* (2017), 1065.

[50] Fernandez, M., Riveros, J. D., Campos, M., Mathee, K., and Narasimhan, G. Microbial" social networks". *BMC genomics 16*, 11 (2015), S6.

[51] Fiehn, O. Metabolomics—the link between genotypes and phenotypes. In *Functional genomics*. Springer, 2002, pp. 155–171.

[52] Fisher, C. K., and Mehta, P. Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression. *PloS one 9*, 7 (2014), e102451.

[53] Friedman, J., and Alm, E. J. Inferring correlation networks from genomic survey data. *PLoS Comput Biol 8*, 9 (2012), e1002687.

[54] Friedman, N., Nachman, I., and Pe'er, D. Learning bayesian network structure from massive datasets: The" sparse candidate" algorithm. *arXiv preprint arXiv:1301.6696* (2013).

[55] Gajer, P., Brotman, R. M., Bai, G., Sakamoto, J., Schütte, U. M. E., Zhong, X., Koenig, S. S. K., Fu, L., Ma, Z. S., Zhou, X., Abdo, Z., Forney, L. J., and Ravel, J. Temporal dynamics of the human vaginal microbiota. *Sci Transl Med 4*, 132 (2012), 132ra52.

[56] Gao, X., Huynh, B.-T., Guillemot, D., Glaser, P., and Opatowski, L. Inference of significant microbial interactions from longitudinal metagenomics data. *Frontiers in microbiology 9* (2018), 2319.

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

[58] Gibson, T. E., and Gerber, G. K. Robust and scalable models of microbiome dynamics. In *Proc. 35th International Conference on Machine Learning* (2018), PMLR 80, pp. 1763–1772.

[59] Glowacki, R. W., Pudlo, N. A., Tuncil, Y., Luis, A. S., Sajjakulnukit, P., Terekhov, A. I., Lyssiotis, C. A., Hamaker, B. R., and Martens, E. C. A ribose-scavenging system confers colonization fitness on the human gut symbiont bacteroides thetaiotaomicron in a diet-specific manner. *Cell host & microbe 27*, 1 (2020), 79–92.

[60] Gonzales, C., Dubuisson, S., and Manfredotti, C. E. A new algorithm for learning non-stationary dynamic bayesian networks with application to event detection. In *FLAIRS Conference* (2015), pp. 564–569.

[61] Grzegorczyk, M., and Husmeier, D. Non-stationary continuous dynamic bayesian networks. In *Advances in Neural Information Processing Systems* (2009), pp. 682–690.

[62] Halloran, J. T., Bilmes, J. A., and Noble, W. S. Dynamic bayesian network for accurate detection of peptides from tandem mass spectra. *J Proteome Res 15*, 8 (2016), 2749–2759.

[63] Hardy, L., Jespers, V., Abdellati, S., De Baetselier, I., Mwambarangwe, L., Musengamana, V., van de Wijgert, J., Vaneechoutte, M., and Crucitti, T. A fruitful alliance: the synergy between atopobium vaginae and gardnerella vaginalis in bacterial vaginosis-associated biofilm. *Sex Transm Infect 92*, 7 (2016), 487–491.

[64] Heckerman, D. A tutorial on learning with bayesian networks, 2020.

[65] Heckerman, D., and Geiger, D. Learning bayesian networks: a unification for discrete and gaussian domains. *arXiv preprint arXiv:1302.4957* (2013).

[66] Heckerman, D., Geiger, D., and Chickering, D. M. Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning 20*, 3 (1995), 197–243.

[67] Hickey, R. J., Abdo, Z., Zhou, X., Nemeth, K., Hansmann, M., Osborn, T. W., Wang, F., and Forney, L. J. Effects of tampons and menses on the composition and diversity of vaginal microbial communities over time. *BJOG 120*, 6 (2013), 695–706.

[68] Holloway, B. Genetic recombination in Pseudomonas aeruginosa. *Microbiology 13*, 3 (1955), 572–581.

[69] Hughes, D. A., Bacigalupe, R., Wang, J., Rühlemann, M. C., Tito, R. Y., Falony, G., Joossens, M., Vieira-Silva, S., Henckaerts, L., Rymenans, L., et al. Genome-wide associations of human gut microbiome variation and implications for causal inference analyses. *Nature Microbiology 5*, 9 (2020), 1079–1087.

[70] Hymel, G. M. Chapter 2 - research essentials for massage in the healthcare setting. In *Clinical Massage in the Healthcare Setting*, S. Fritz, L. Chaitow, and G. M. Hymel, Eds. Mosby, Saint Louis, 2008, pp. 20 – 53.

[71] iHMP. The integrative human microbiome project. *Nature 569*, 7758 (2019), 641.

[72] (iHMP) Research Network Consortium, T. I. H. The integrative human microbiome project: Dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell Host & Microbe 16*, 3 (2014), 276–289.

[73] Imhoff, J. F., and Rodriguez-Valera, F. Betaine is the main compatible solute of halophilic eubacteria. *Journal of bacteriology 160*, 1 (1984), 478–479.

[74] Imoto, S., Goto, T., and Miyano, S. Estimation of genetic networks and functional structures between genes by using bayesian networks and nonparametric regression. In *Biocomputing 2002*. World Scientific, 2001, pp. 175–186.

[75] James, G. M., and Hastie, T. J. Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63*, 3 (2001), 533–550.

[76] Ji, B. W., Sheth, R. U., Dixit, P. D., Huang, Y., Kaufman, A., Wang, H. H., and Vitkup, D. Quantifying spatiotemporal variability and noise in absolute microbiota abundances using replicate sampling. *Nat Methods 16* (2019), 731–736.

[77] Ji, Z., Xia, Q., and Meng, G. A review of parameter learning methods in bayesian network. In *International Conference on Intelligent Computing* (2015), Springer, pp. 3–12.

[78] Joseph, T. A., Shenhav, L., Xavier, J. B., Halperin, E., and Pe'er, I. Compositional lotka-volterra describes microbial dynamics in the simplex. *PLOS Computational Biology 16*, 5 (2020), e1007917.

[79] Jost, T., Lacroix, C., Braegger, C., and Chassard, C. Assessment of bacterial diversity in breast milk using culture-dependent and culture-independent approaches. *Br J Nutr 110*, 7 (2013), 1253–1262.

[80] Kappler, K., Lasanajak, Y., Smith, D. F., Opitz, L., and Hennet, T. Increased antibody response to fucosylated oligosaccharides and fucose-carrying bacteroides species in crohn's disease. *Frontiers in microbiology 11* (2020), 1553.

[81] Karp, P. D., Riley, M., Saier, M., Paulsen, I. T., Paley, S. M., and Pellegrini-Toole, A. The EcoCyc and MetaCyc databases. *Nucleic Acids Res 28*, 1 (2000), 56–59.

[82] KEGG. Bacteroides thetaiotaomicron 7330: Btheta7330_03179. `https://www.genome.jp/dbget-bin/www_bget?btho:Btheta7330_03179`, Accessed: 2020-10-20.

[83] KEGG. Eubacterium siraeum v10sc8a: Es1_08270. https://www.genome.jp/dbget-bin/www_bget?esr:ES1_08270, Accessed: 2020-10-20.

[84] Kim, S., Imoto, S., and Miyano, S. Dynamic bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data. In *International Conference on Computational Methods in Systems Biology* (2003), Springer, pp. 104–113.

[85] Kolenbrander, P. E., Andersen, R. N., Blehert, D. S., Egland, P. G., Foster, J. S., and Palmer, R. J. Communication among oral bacteria. *Microbiol Mol Biol Rev 66*, 3 (2002), 486–505.

[86] 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 (2014), 12522–12527.

[87] Lähdesmäki, H., Hautaniemi, S., Shmulevich, I., and Yli-Harja, O. Relationships between probabilistic boolean networks and dynamic bayesian networks as models of gene regulatory networks. *Signal processing 86*, 4 (2006), 814–834.

[88] Lauritzen, S. L., and Wermuth, N. Graphical models for associations between variables, some of which are qualitative and some quantitative. *Ann Statist 17*, 1 (1989), 31–57.

[89] Lèbre, S. Inferring dynamic genetic networks with low order independencies. *Statistical applications in genetics and molecular biology 8*, 1 (2009).

[90] Li, S. Z. *Markov random field modeling in image analysis*. Springer Science & Business Media, 2009.

[91] Liu, Z., Cao, A. T., and Cong, Y. Microbiota regulation of inflammatory bowel disease and colorectal cancer. In *Seminars in Cancer Biology* (2013), vol. 23, Elsevier, pp. 543–552.

[92] 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 (2019), 655.

[93] Lo, C., and Marculescu, R. Inferring microbial interactions from metagenomic time-series using prior biological knowledge. In *Proceedings of the 8th ACM Inter-*

national Conference on Bioinformatics, Computational Biology, and Health Informatics (2017), pp. 168–177.

[94] Lo, C., and Marculescu, R. Inferring microbial interactions from metagenomic time-series using prior biological knowledge. In *Proc. 8th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics* (2017), ACM-BCB '17, pp. 168–177.

[95] Love, M., Huber, W., and Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol 15*, 12 (2014), 550.

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

[97] Lynch, K. E., Parke, E. C., and O'Malley, M. A. How causal are microbiomes? a comparison with the helicobacter pylori explanation of ulcers. *Biology & Philosophy 34*, 6 (2019), 62.

[98] Ma, T., and Zhang, A. Integrate multi-omics data with biological interaction networks using multi-view factorization autoencoder (mae). *BMC Genomics 20*, 11 (2019), 1–11.

[99] Madhavan, S., Bender, R. J., and Petricoin, E. F. Integration of multi-omic data into a single scoring model for input into a treatment recommendation ranking, 2019. US Patent App. 16/405,640.

[100] Maier, L., Pruteanu, M., Kuhn, M., Zeller, G., Telzerow, A., Anderson, E. E., Brochado, A. R., Fernandez, K. C., Dose, H., Mori, H., et al. Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature 555*, 7698 (2018), 623–628.

[101] Malinsky, D., and Spirtes, P. Causal structure learning from multivariate time series in settings with unmeasured confounding. In *Proceedings of 2018 ACM SIGKDD Workshop on Causal Discovery* (2018), pp. 23–47.

[102] Mani, S., Spirtes, P. L., and Cooper, G. F. A theoretical study of y structures for causal discovery. *arXiv preprint arXiv:1206.6853* (2012).

[103] Marino, S., Baxter, N. T., Huffnagle, G. B., Petrosino, J. F., and Schloss, P. D. Mathematical modeling of primary succession of murine intestinal microbiota. *Proceedings of the National Academy of Sciences 111*, 1 (2014), 439–444.

[104] Marino, S., Baxter, N. T., Huffnagle, G. B., Petrosino, J. F., and Schloss, P. D. Mathematical modeling of primary succession of murine intestinal microbiota. *Proc Natl Acad Sci 111*, 1 (2014), 439–444.

[105] 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 (2014), e1003676.

[106] 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 (2014), 1–7.

[107] McGeachie, M. J., Sordillo, J. E., Gibson, T., Weinstock, G. M., Liu, Y.-Y., Gold, D. R., Weiss, S. T., and Litonjua, A. Longitudinal prediction of the infant gut microbiome with dynamic bayesian networks. *Scientific reports 6* (2016), 20359.

[108] McGeachie, M. J., Sordillo, J. E., Gibson, T., Weinstock, G. M., Liu, Y.-Y., Gold, D. R., Weiss, S. T., and Litonjua, A. Longitudinal prediction of the infant gut microbiome with dynamic bayesian networks. *Sci Rep* (2016), 20359.

[109] Metropolis, N., and Ulam, S. The monte carlo method. *Journal of the American statistical association 44*, 247 (1949), 335–341.

[110] Moran, M. Metatranscriptomics: eavesdropping on complex microbial communities. microbe mag 4: 329–335, 2009.

[111] Morton, J. T., Aksenov, A. A., Nothias, L. F., Foulds, J. R., Quinn, R. A., Badri, M. H., Swenson, T. L., Van Goethem, M. W., Northen, T. R., Vazquez-Baeza, Y., et al. Learning representations of microbe–metabolite interactions. *Nat Methods 16*, 12 (2019), 1306–1314.

[112] Mounier, J., Monnet, C., Vallaeys, T., Arditi, R., Sarthou, A.-S., Hélias, A., and Irlinger, F. Microbial interactions within a cheese microbial community. *Applied and environmental microbiology 74*, 1 (2008), 172–181.

[113] Navarro Llorens, J. M., Tormo, A., and Martínez-García, E. Stationary phase in gram-negative bacteria. *FEMS Microbiol Rev 34*, 4 (2010), 476–495.

[114] NCBI. Bacste_rs03560 cytidine deaminase [bacteroides stercoris atcc 43183]. `https://www.ncbi.nlm.nih.gov/gene?term=31796333`, Accessed: 2020-10-20.

[115] NCBI. Bacste_rs07450 uridine kinase [bacteroides stercoris atcc 43183]. https://www.ncbi.nlm.nih.gov/gene?term=31797103, Accessed: 2020-10-20.

[116] Nefian, A. V., Liang, L., Pi, X., Liu, X., and Murphy, K. Dynamic bayesian networks for audio-visual speech recognition. *EURASIP J Adv Signal Process*, 11 (2002), 1274–1288.

[117] 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 (2016), e00013–15.

[118] Nugent, R. P., Krohn, M. A., and Hillier, S. L. Reliability of diagnosing bacterial vaginosis is improved by a standardized method of gram stain interpretation. *J Clin Microbiol 29*, 2 (1991), 297–301.

[119] O'Hagan, A., and Forster, J. J. *Kendall's advanced theory of statistics, Vol. 2B: Bayesian inference*, 2nd ed. Edward Arnold Press, London, UK, 2004.

[120] Olivella, S., and Shiraito, Y. Poisbinom: A faster implementation of the poisson-binomial distribution. *R package version 1*, 1 (2017).

[121] Ong, I. M., Glasner, J. D., and Page, D. Modelling regulatory pathways in e. coli from time series expression profiles. *Bioinformatics 18*, suppl_1 (2002), S241–S248.

[122] Opgen-Rhein, R., and Strimmer, K. Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive process. *BMC bioinformatics 8*, 2 (2007), 1–8.

[123] Orengo, A. Regulation of enzymic activity by metabolites i. uridine-cytidine kinase of novikoff ascites rat tumor. *Journal of Biological Chemistry 244*, 8 (1969), 2204–2209.

[124] Palsson, B., and Zengler, K. The challenges of integrating multi-omic data sets. *Nat Chem Biol 6*, 11 (2010), 787–789.

[125] Pearl, J. *Causality*. Cambridge university press, 2009.

[126] Pearl, J. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.

[127] Pearson, K. Determination of the coefficient of correlation. *Science 30*, 757 (1909), 23–25.

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

[129] Perković, E., Kalisch, M., and Maathuis, M. H. Interpreting and using cpdags with background knowledge. *arXiv preprint arXiv:1707.02171* (2017).

[130] Perrin, B.-E., Ralaivola, L., Mazurie, A., Bottani, S., Mallet, J., and d'Alche Buc, F. Gene networks inference using dynamic bayesian networks. *Bioinformatics 19*, suppl_2 (2003), ii138–ii148.

[131] Petersen, M., Schwab, J., Gruber, S., Blaser, N., Schomaker, M., and van der Laan, M. Targeted maximum likelihood estimation for dynamic and static longitudinal marginal structural working models. *Journal of causal inference 2*, 2 (2014), 147–185.

[132] Petrova, M. I., Reid, G., Vaneechoutte, M., and Lebeer, S. Lactobacillus iners: friend or foe? *Trends Microbiol 25*, 3 (2017), 182–191.

[133] Ramoni, M., and Sebastiani, P. Robust learning with missing data. *Machine Learning 45*, 2 (2001), 147–170.

[134] Rangel, C., Angus, J., Ghahramani, Z., Lioumi, M., Sotheran, E., Gaiba, A., Wild, D. L., and Falciani, F. Modeling t-cell activation using gene expression profiling and state-space models. *Bioinformatics 20*, 9 (2004), 1361–1372.

[135] Ranjan, R., Rani, A., Metwally, A., McGee, H. S., and Perkins, D. L. Analysis of the Microbiome: Advantages of Whole Genome Shotgun versus 16S Amplicon Sequencing. *Biochemical and biophysical research communications 469*, 4 (Jan. 2016), 967–977.

[136] Ravel, J., Gajer, P., Abdo, Z., Schneider, G. M., Koenig, S. S. K., McCulle, S. L., Karlebach, S., Gorle, R., Russell, J., Tacket, C. O., et al. Vaginal microbiome of reproductive-age women. *Proc Natl Acad Sci 108*, Suppl 1 (2011), 4680–4687.

[137] Relman, D. A. Thinking about the microbiome as a causal factor in human health and disease: philosophical and experimental considerations. *Current Opinion in Microbiology 54* (2020), 119–126.

[138] Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., and Sabeti, P. C. Detecting novel associations in large data sets. *science 334*, 6062 (2011), 1518–1524.

[139] Riesenfeld, C. S., Schloss, P. D., and Handelsman, J. Metagenomics: genomic analysis of microbial communities. *Annu Rev Genet 38* (2004), 525–552.

[140] Ríos-Covián, D., Ruas-Madiedo, P., Margolles, A., Gueimonde, M., De Los Reyes-gavilán, C. G., and Salazar, N. Intestinal short chain fatty acids and their link with diet and human health. *Frontiers in microbiology 7* (2016), 185.

[141] Robinson, J. W., and Hartemink, A. J. Learning non-stationary dynamic bayesian networks. *J Mach Learn Res 11* (2010), 3647–3680.

[142] Robinson, J. W., Hartemink, A. J., and Ghahramani, Z. Learning non-stationary dynamic bayesian networks. *Journal of Machine Learning Research 11*, 12 (2010).

[143] Rogers, D. F. Mathematical elements for computer graphics. CUMINCAD, 1990.

[144] Ruan, Q., Dutta, D., Schwalbach, M. S., Steele, J. A., Fuhrman, J. A., and Sun, F. Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors. *Bioinformatics 22*, 20 (2006), 2532–2538.

[145] Ruiz-Perez, D., Guan, H., Madhivanan, P., Mathee, K., and Narasimhan, G. So you think you can PLS-DA? *BMC Bioinformatics In Press* (2020).

[146] Runge, J. Causal network reconstruction from time series: From theoretical assumptions to practical estimation. *Chaos: An Interdisciplinary Journal of Nonlinear Science 28*, 7 (2018), 075310.

[147] Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., and Sejdinovic, D. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances 5*, 11 (2019), eaau4996.

[148] Russell, S. J., and Norvig, P. *Artificial Intelligence: A Modern Approach*, 2nd ed. Prentice Hall Press, Upper Saddle River, NJ, USA, 2003.

[149] Saitoh, S., Noda, S., Aiba, Y., Takagi, A., Sakamoto, M., Benno, Y., and Koga, Y. Bacteroides ovatus as the predominant commensal intestinal microbe causing a systemic antibody response in inflammatory bowel disease. *Clinical and diagnostic laboratory immunology 9*, 1 (2002), 54–59.

[150] Sangaralingam, A., Dayem U., A. Z., Marzec, J., Gadaleta, E., Nagano, A., Ross-Adams, H., Wang, J., Lemoine, N. R., and Chelala, C. 'Multi-omic' data analysis using O-miner. *Brief Bioinform 20*, 1 (2017), 130–143.

[151] Sanna, S., van Zuydam, N. R., Mahajan, A., Kurilshikov, A., Vila, A. V., Võsa, U., Mujagic, Z., Masclee, A. A., Jonkers, D. M., Oosting, M., et al. Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. *Nature genetics 51*, 4 (2019), 600–605.

[152] Sazal, M. R., Stebliankin, V., Mathee, K., and Narasimhan, G. Causal inference in microbiomes using intervention calculus. *bioRxiv* (2020).

[153] Scheines, R., Spirtes, P., Glymour, C., Meek, C., and Richardson, T. The tetrad project: Constraint based aids to causal model specification. *Multivariate Behavioral Research 33*, 1 (1998), 65–117.

[154] Schirmer, M., Denson, L., Vlamakis, H., Franzosa, E. A., Thomas, S., Gotman, N. M., Rufo, P., Baker, S. S., Sauer, C., Markowitz, J., et al. Compositional and temporal changes in the gut microbiome of pediatric ulcerative colitis patients are linked to disease course. *Cell host & microbe 24*, 4 (2018), 600–610.

[155] Schwarz, G., et al. Estimating the dimension of a model. *The annals of statistics 6*, 2 (1978), 461–464.

[156] Scutari, M. Bayesian network constraint-based structure learning algorithms: Parallel and optimised implementations in the bnlearn R package. *arXiv preprint arXiv:1406.7648* (2014).

[157] Scutari, M. Dirichlet bayesian network scores and the maximum relative entropy principle. *Behaviormetrika 45*, 2 (2018), 337–362.

[158] Scutari, M., Graafland, C. E., and Gutiérrez, J. M. Who learns better Bayesian network structures: Constraint-based, score-based or hybrid algorithms? In *International Conference on Probabilistic Graphical Models* (2018), pp. 416–427.

[159] Scutari, M., and Lebre, S. Bayesian networks in R: With applications in systems biology, 2013.

[160] Selhub, J., Byun, A., Liu, Z., Mason, J. B., Bronson, R. T., and Crott, J. W. Dietary vitamin b6 intake modulates colonic inflammation in the il10-/- model of inflammatory bowel disease. *The Journal of nutritional biochemistry 24*, 12 (2013), 2138–2143.

[161] Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res 13*, 11 (2003), 2498–2504.

[162] Shaw, G. T.-W., Pao, Y.-Y., and Wang, D. Metamis: a metagenomic microbial interaction simulator based on microbial community profiles. *BMC bioinformatics 17*, 1 (2016), 488.

[163] Silander, T., Kontkanen, P., and Myllymäki, P. On sensitivity of the map bayesian network structure to the equivalent sample size parameter. In *Proc. 23rd Conference on Uncertainty in Artificial Intelligence* (2007), UAI '07, pp. 360–367.

[164] Silverman, J. D., Washburne, A. D., Mukherjee, S., and David, L. A. A phylogenetic transform enhances analysis of compositional microbiota data. *Elife 6* (2017), e21887.

[165] Sims, C. A. Macroeconomics and reality. *Econometrica: journal of the Econometric Society* (1980), 1–48.

[166] Sköld, O. Uridine kinase from ehrlich ascites tumor: purification and properties. *Journal of Biological Chemistry 235*, 11 (1960), 3273–3279.

[167] Smith, A. A., Vollrath, A., Bradfield, C. A., and Craven, M. Clustered alignments of gene-expression time series data. *Bioinformatics 25*, 12 (2009), i119–i127.

[168] Song, B.-H., and Neuhard, J. Chromosomal location, cloning and nucleotide sequence of the bacillus subtilis cdd gene encoding cytidine/deoxycytidine deaminase. *Molecular and General Genetics MGG 216*, 2-3 (1989), 462–468.

[169] Spearman, C. The proof and measurement of association between two things. Appleton-Century-Crofts, 1961.

[170] Spirtes, P., Glymour, C. N., Scheines, R., and Heckerman, D. *Causation, prediction, and search*. MIT press, 2000.

[171] Steck, H. Learning the bayesian network structure: Dirichlet prior vs data. In *Proc. 24th Conference on Uncertainty in Artificial Intelligence* (2008), UAI '08, pp. 511–518.

[172] Stein, R. R., Bucci, V., Toussaint, N. C., Buffie, C. G., Rätsch, G., Pamer, E. G., Sander, C., and Xavier, J. B. Ecological modeling from time-series inference: Insight into dynamics and stability of intestinal microbiota. *PLoS Comput Biol 9*, 12 (2013), 1–11.

[173] Stein, R. R., Bucci, V., Toussaint, N. C., Buffie, C. G., Rätsch, G., Pamer, E. G., Sander, C., and Xavier, J. B. Ecological modeling from time-series inference: in-

sight into dynamics and stability of intestinal microbiota. *PLoS Comput Biol 9*, 12 (2013), e1003388.

[174] Sugimoto, N., and Iba, H. Inference of gene regulatory networks by means of dynamic differential bayesian networks and nonparametric regression. *Genome Informatics 15*, 2 (2004), 121–130.

[175] Sullivan, M., and Verhoosel, J. *Statistics: Informed decisions using data*. Pearson Boston, MA, 2013.

[176] Takai, A., Marusawa, H., Minaki, Y., Watanabe, T., Nakase, H., Kinoshita, K., Tsujimoto, G., and Chiba, T. Targeting activation-induced cytidine deaminase prevents colon cancer development despite persistent colonic inflammation. *Oncogene 31*, 13 (2012), 1733–1742.

[177] Trapnell, C., and Salzberg, S. L. How to Map Billions of Short Reads onto Genomes. *Nature biotechnology 27*, 5 (May 2009), 455–457.

[178] Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. Transcript Assembly and Quantification by RNA-Seq Reveals Unannotated Transcripts and Isoform Switching During Cell Differentiation. *Nature biotechnology 28*, 5 (May 2010), 511–515.

[179] Trosvik, P., Stenseth, N. C., and Rudi, K. Characterizing mixed microbial population dynamics using time-series analysis. *ISME J 2* (2008), 707–715.

[180] Tsamardinos, I., Brown, L. E., and Aliferis, C. F. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning 65*, 1 (2006), 31–78.

[181] Tulchinsky, T. H., and Varavikova, E. A. Chapter 3 - measuring, monitoring, and evaluating the health of a population. In *The New Public Health (Third Edition)*, T. H. Tulchinsky and E. A. Varavikova, Eds., third edition ed. Academic Press, San Diego, 2014, pp. 91 – 147.

[182] Turnbaugh, P. J., and Gordon, J. I. An invitation to the marriage of metagenomics and metabolomics. *Cell 134*, 5 (2008), 708–713.

[183] Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., and Gordon, J. I. The human microbiome project. *Nature 449*, 7164 (2007), 804.

[184] Ulfenborg, B. Vertical and horizontal integration of multi-omics data with miodin. *BMC Bioinform 20*, 1 (2019), 649.

[185] UniProt. Uniprotkb - r9hq62 (r9hq62_bact4). https://www.uniprot.org/uniprot/R9HQ62, Accessed: 2020-10-20.

[186] Valentin-Hansen, P. [39] uridine-cytidine kinase from escherichia coli. In *Methods in enzymology*, vol. 51. Elsevier, 1978, pp. 308–314.

[187] VanderWeele, T. J. Concerning the consistency assumption in causal inference. *Epidemiology 20*, 6 (2009), 880–883.

[188] Verma, T., and Pearl, J. *Equivalence and synthesis of causal models*. UCLA, Computer Science Department, 1991.

[189] Vincenzetti, S., Cambi, A., Neuhard, J., Schnorr, K., Grelloni, M., and Vita, A. Cloning, expression, and purification of cytidine deaminase fromarabidopsis thaliana. *Protein expression and purification 15*, 1 (1999), 8–15.

[190] Walters, S. S., Quiros, A., Rolston, M., Grishina, I., Li, J., Fenton, A., DeSantis, T. Z., Thai, A., Andersen, G. L., Papathakis, P., et al. Analysis of gut microbiome and diet modification in patients with crohn's disease. *SOJ microbiology & infectious diseases 2*, 3 (2014), 1.

[191] Wang, H., Khan, F., Chen, B., and Lu, Z. An approximate modelling method for industrial l-lysine fermentation process. In *Computer Aided Chemical Engineering*, vol. 37. Elsevier, 2015, pp. 461–466.

[192] Wang, T., Sable, H., and Lampen, J. Enzymatic deamination of cytosine nucleosides. *Journal of Biological Chemistry 184*, 1 (1950), 17–28.

[193] Wang, Y., and Blei, D. M. The blessings of multiple causes. *Journal of the American Statistical Association*, just-accepted (2019), 1–71.

[194] Wang, Z., Gerstein, M., and Snyder, M. RNA-Seq: A Revolutionary Tool for Transcriptomics. *Nature reviews Genetics 10*, 1 (Jan. 2009), 57–63.

[195] Watkins, A. J., Roussel, E. G., Parkes, R. J., and Sass, H. Glycine betaine as a direct substrate for methanogens (methanococcoides spp.). *Applied and Environmental Microbiology 80*, 1 (2014), 289–293.

[196] Weber, P., Medina-Oliva, G., Simon, C., and Iung, B. Overview on bayesian networks applications for dependability, risk analysis and maintenance areas. *Eng Appl Artif Intell 25*, 4 (2012), 671–682.

[197] Weiss, S., Van Treuren, W., Lozupone, C., Faust, K., Friedman, J., Deng, Y., Xia, L. C., Xu, Z. Z., Ursell, L., Alm, E. J., et al. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *The ISME journal 10*, 7 (2016), 1669–1681.

[198] Wilczyński, B., and Dojer, N. BNFinder: exact and efficient method for learning bayesian networks. *Bioinformatics 25*, 2 (2009), 286–287.

[199] Wishart, D. S., Jewison, T., Guo, A. C., Wilson, M., Knox, C., Liu, Y., Djoumbou, Y., Mandal, R., Aziat, F., Dong, E., et al. HMDB 3.0—the human metabolome database in 2013. *Nucleic Acids Res 41*, D1 (2012), D801–D807.

[200] Wu, F.-X., Zhang, W.-J., and Kusalik, A. J. Modeling gene expression from microarray expression data with state-space equations. In *Biocomputing 2004*. World Scientific, 2003, pp. 581–592.

[201] Xiao, H. *Network-based approaches for multi-omic data integration*. PhD thesis, University of Cambridge, 2019.

[202] Yugi, K., Kubota, H., Hatano, A., and Kuroda, S. Trans-omics: how to reconstruct biochemical networks across multiple 'omic'layers. *Trends Biotech 34*, 4 (2016), 276–290.

[203] Zhou, W., Sailani, M. R., Contrepois, K., Zhou, Y., Ahadi, S., Leopold, S. R., Zhang, M. J., Rao, V., Avina, M., Mishra, T., et al. Longitudinal multi-omics of host–microbe dynamics in prediabetes. *Nature 569*, 7758 (2019), 663–671.

[204] Zou, M., and Conzen, S. D. A new dynamic bayesian network (dbn) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics 21*, 1 (2005), 71–79.

[205] Zweig, G. *Speech recognition with Dynamic Bayesian Networks*. PhD thesis, University of California, Berkeley, 1998.

VITA

DANIEL RUIZ-PEREZ

2011-2015	B.S., Computer Science University of A Coruña, A Coruña, (Spain)
2014-2015	Undergraduate Research Assistant University of A Coruña, A Coruña, (Spain)
2015-2016	Software Engineer Aldaba IT and services, A Coruña, (Spain)
2015-2020	M.E., Software Engineering University of A Coruña, A Coruña, (Spain)
2016-2019	Graduate Assistant Florida International University, FL (USA)
2016-2019	M.S., Computer Science Florida International University, FL (USA)
2019-2019	Data Analyst Intern Assurant, Miami, FL (USA)
2019-2020	Dissertation Year Fellow Florida International University, FL (USA)
2020-2020	Machine Learning SWE Intern Facebook, Menlo park, CA (USA)

PUBLICATIONS AND PRESENTATIONS

Colbert, B., Coudray, M., Ruiz-Perez, D., Krupp, K., Narasimhan, G., Madhivanan, P., Mathee, K. (2018). To Treat or Not to Treat: Bacterial Vaginosis and its Relationship to Human Papillomavirus. Microbiology Society Annual Conference at Birmingham, UK.

Coudray, M., Ruiz-Perez, D. Colbert, B., Krupp, K., Kumari, H., Narasimhan, G., Mathee, K, Madhivanan P. (2019) Effect of metronidazole on microbiomes associated with asymptomatic bacterial vaginosis (abstract). Access Microbiology, 1(1A).

Coudray, M., Ruiz-Perez, D., Colbert, B., Krupp, K., Kumari, H., Mathee, K., Narasimhan G. (2019). P371 Effect of metronidazole treatment on recurrent and persistent bacterial vaginosis: a pilot study. The BMJ Sexually Transmitted Infections;95:A187

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

Lugo-Martinez, J., Ruiz-Perez, D., Narasimhan, G., Bar-Joseph, Z. (2019). Dynamic interaction network inference from longitudinal microbiome data. Society for Molecular Biology and Evolution (SMBE) at Manchester, UK.

Lugo-Martinez, J., Ruiz-Perez, D., Narasimhan, G., Bar-Joseph, Z. (2019). Dynamic interaction network inference from longitudinal microbiome data. RECOMB at Washington, DC, USA.

Lugo-Martinez, J., Ruiz-Perez, D., Narasimhan, G., Bar-Joseph Z. (2019). Dynamic interaction network inference from longitudinal microbiome data. ACM-BCB at Niagara Falls, USA.

Ruiz-Perez, D., Lugo-Martinez, J., Bourguignon, N., Mathee, K., Lerner, B., Bar-Joseph, Z., Narasimhan, G. (2020). Dynamic Bayesian networks for integrating multi-omics time-series microbiome data. BioRxiv preprint DOI: 10.1101/835124.

Ruiz-Perez, D., Guan, H., Madhivanan, P., Mathee, K., Narasimhan, G. (2019). So you think you can PLS-DA?. In press at BMC Bioinformatics. BioRxiv DOI:10.1101/207225

Ruiz-Perez, D., Lugo-Martinez, J., Bar-Joseph, Z., Narasimhan, G. (2020). Application of Bayesian Techniques to Multi-omic Longitudinal Data. International Conference on Machine Learning (ICML) (LXAI) at Vienna, Austria.

Ruiz-Perez, D., Lugo-Martinez, J., Bourguignon, N., Lerner, B., Mathee, K., Bar-Joseph, Z., Narasimhan, G. (2019). Temporal interactions of genes, taxa, and metabolites of the microbiota in patients with inflammatory bowel disease. Asian Conference on Transcription at Dunedin, Otago, New Zealand.

Ruiz-Perez, D., Guan, H., Madhivanan, P., Mathee, K., Narasimhan, G. (2018). So you think you can PLS-DA?. International Conference on Computational Advances in Bio and medical Sciences (ICCABS) at Las Vegas, NV, USA.

Ruiz-Perez, D., Colbert, B., Coudray, M., Mathee, K., Madhivanan, P., Narasimhan, G. (2018). Vaginal microbial profile of women with asymptomatic bacterial vaginosis in US. Microbiology Society Annual Conference at Birmingham, UK.

Ruiz-Perez, D., Sazal, M., Park, Cickovski, T., Lee, H., Cho, H., Hwang, D., Narasimhan, G. (2019). Role of gut microbiota and their temporal interactions in kidney transplant recipients. LXAI at NeurIPS, Vancouver, Canada.

Sazal, M., Ruiz-Perez, D., Cickovski, T., Narasimhan, G. (2019). Inferring Relationships in Microbiomes from Signed Bayesian Networks. In press at BMC Bioinformatics. BioRxiv preprint DOI: 10.1101/2020.02.18.955344.

Sazal, M., Ruiz-Perez, D., Valdes, C., Cickovski, T., Stebliankin, V., Mehta, A., Mathee, K., Narasimhan, G. (2019). Signed Causal Bayesian Networks for Microbiomes. LXAI at NeurIPS, Vancouver, Canada.

Sazal, M., Ruiz-Perez, D., Cickovski, T., Narasimhan, G. (2018). Inferring Relationships in Microbiomes from Signed Bayesian Networks. ICCABS, Las Vegas, NV, USA.

Suarez-Ulloa, V., Aguiar-Pulido, V., Ruiz-Perez, D., Narasimhan, G., Eirin-Lopez, JM. (2016). Network-based analysis of chromatin-associated gene expression dynamics in response to environmental stress. ISMB at Orlando, USA.