Pattern Recognition for Complex Heterogeneous Time-Series Data: An Analysis of Microbial Community Dynamics

Rajith Vidanaarachchi

A thesis submitted for the degree of
Doctor of Philosophy at
The Australian National University

September 2021

Except where otherwise indicated, this thesis is my own original work.

Rajith Vidanaarachchi
14 September 2021

To my parents

Acknowledgments

"It takes a village to raise a child"

—African Proverb

First and foremost, I acknowledge and pay respects to the traditional owners of the land on which I conducted my research. My acknowledgement extends to the Ngunnawal and Ngambri people, on whose land the Australian National University is built upon, and the Wurundjeri people of Kulin nation, on whose land the University of Melbourne is situated. As Aboriginal people continue their struggle for equity and respect in the a that was taken from them, I recognise my role as an academic—and a human—to support their interests, as well as the interests of the others whose voices are suppressed.

I would sincerely like to thank Prof Saman Halgamuge, my primary supervisor, for his guidance and for being there for me. I would also like to thank my other supervisors, Dr Marnie Shaw, for her guidance and kindness, and Prof Sen-Lin Tang for his insights in the field of microbiology. My research was conducted initially in Canberra at the Australian National University and then at the University of Melbourne. I would like to thank the academic and professional staff at both universities who assisted me in many ways. A special shout-out is due to the ANU Academic Skills team for helping me develop my writing skills—arguably the most challenging part of the PhD.

I would like to thank my colleagues and friends, Dr Damayanthi Herath, Dr Damith Senanayake, and Dr Chathurika Mediwaththe. They guided me when I felt stuck in my research journey. I would also like to thank the Pattern Recognition and Optimisation group members, Nanduni Nimalsiri, Maneesha Perera, Tamasha Malepathirana, Rashindrie Perera, and Dr Richard Wang, for their friendship (and proofreading). I would also like to thank all my lecturers, teachers and mentors I had in my life, especially Dr Chandana Gamage from the University of Moratuwa, Sri Lanka, for encouraging me to do a PhD.

Just as it takes a village to raise a child, it takes a village to produce a PhD—My village, listed here, helped keep my sanity during the past four years. Hence, my gratitude goes to family and friends for supporting me during my candidature. First, to my parents for their love and sacrifice, to my brother for being there, for being reliable. To Ama for being my friend and the best sister-in-law, one could have. To Byron for his love and companionship, for going on adventures with me and introducing me to art, which helped me keep sanity during the stressful periods of my candidature and the Covid-19 lock-downs. To Patrick, for being a good friend,

a good flatmate and a fellow interdisciplinary academic who never shied away from listening to my PhD problems. To Thanuja and other Sri Lankan friends in Canberra. Firstly, for their love, friendship and familiarity. Secondly, for making sure that I was well-fed during the summer in which I fractured my knee. To Andy, Cedric, and Judit for their love, for being my friends and keeping me company in my Canberra days. To the Canberra Underdogs—the coolest Underwater Rugby team ever, for making me feel at home in my early days in Australia and for the camaraderie. To Sanka, Sachini, Sachith and Vipula, for being my peer support network, for the years-long friendship, and calling me on every birthday. To Sabra, Kulakshi, Ridwan, Danula and other friends from the University of Moratuwa's Computer Science and Engineering, who answered my requests to proofread my thesis chapters at the last minute. To Leo, Bruno and other canine good boys who gave me licks of support. Cheers to all good human (and other) beings who brought me joy and comfort in this journey!

Last but not least, I would like to pay my gratitude to the people of Sri Lanka who funded my education from kindergarten until university and to the people of Australia who funded my PhD. Without Sri Lanka's free education system and Australia's scholarship schemes, my education would have only been a dream.

Thank you!

Love,
Rajith

Abbreviations

AR	Auto-Regressive
ARIMA	Auto-Regressive Integrated Moving Average
ARMA	Auto-Regressive Moving Average
AUPR	Area Under Precision-Recall curves
BCD	Bray–Curtis Dissimilarity
CSL	Competitive Synergistic Links
CST	Community State Type
CoPR	Collective Pattern Recognition
DBSCAN	Density Based Spatial Clustering of Applications with Noise
DNA	Deoxyribonucleic Acid
DTW	Dynamic Time Warp [Distance]
DWT	Discrete Wavelet Transform
EKF	Extended Kalman Filter
EM	Expectation Maximisation
ESABO	Entropy Shifts on Abundance vectors under Boolean Operators
GA	Genetic Algorithm
GMM	Gaussian Mixture Models
gLV	Generalised Lotka–Volterra
HMM	Hidden Markoc Model
ICU	Intensive Care Unit
IMPARO	Inferring Microbial Interactions through Parameter Optimisation
KDE	Kernel Density Estimation
KS	Kolmogorov–Smirnov [Test]
LIMITS	Learning Interactions from Microbial Time Series
LOESS	Locally Estimated Scatterplot Smoothing
LPWC	Lag Penalised Weighted Correlation
LSTM	Long–Short Term Memory [Networks]
LV	Lotka–Volterra
MA	Moving Average
MIN	Microbial Interaction Network
MSE	Mean Squared Error
NGS	Next Generation Sequencing
NICU	Neonatal Intensive Care Unit
NRO	Non-Linear Regulatory OTU-triplet
OPTICS	Ordering Points To Identify Cluster Structure
OTU	Operational Taxonomic Unit

PCR	Partial Component Regression
PLS	Partial Least Squares
PLSR	Parial Least Square Regression
RAPA	Reconstructed Abundance Profile Accuracy
RMN	Rule-based Microbial Network
RMSE	Root Mean Square Error
RNA	Ribonucleic Acid
rRNA	Ribosomal Ribonucleic Acid
SCOW	Shorting Correlation Optimal Warping
SONG	Self-Organising Nebulous Growths
SPIEC-EASI	Sparse Inverse Covariance Estimation for Ecological Association Inference
STARS	Stability Approach to Regularisation Selection
SgLV	Stochastic Generalised Lotka–Volterra
SparCC	Sparse Correlations for Compositional Data
t-SNE	t-distributed Stochastic Neighbour Embedding
TV-DBM	Time-Varying Dynamic Bayesian Networks
TVAP	Temporal Variation of Abundance Profile
UMAP	Uniform Manifold Approximation and Projection

Abstract

Microbial life is the most widespread and the most abundant life form on earth. Microbes exist in complex and diverse communities in environments from the deep ocean trenches to Himalayan snowfields. Microbial life is essential for other forms of life as well. Scientific studies of microbial activity include diverse communities such as plant root microbiome, insect gut microbiome, and human skin microbiome. In the human body alone, the number of microbial life forms surpasses human body cells. Microbial communities are known to affect and shape the host ecosystem with their influence. Hence, it is essential to understand microbial community dynamics. With the advent of 16S rRNA sequencing, we have access to a plethora of data on the microbiome, warrantying a shift from *in vitro* analysis to *in silico* analysis. This thesis focuses on challenges in analysing microbial community dynamics through complex, heterogeneous, and temporal data.

Firstly, we look at the mathematical modelling of microbial community dynamics and the problem of inferring microbial interaction networks by analysing longitudinal sequencing data. We look at this problem to minimise the assumptions involved and improve the accuracy of the inferred interaction networks. Secondly, we explore the temporally dynamic nature of microbial interaction networks. We look at the fallacies of static microbial interaction networks and approaches suitable for modelling temporally dynamic microbial interaction networks. Thirdly, we study multiple temporal microbial datasets from similar environments to understand macro and micro patterns apparent in these communities. We explore the individuality and conformity of microbial communities through visualisation techniques. Finally, we explore the possibility and identify challenges in representing heterogeneous microbial temporal activity in unique signatures.

In summary, we have explored various aspects of complex, heterogeneous, and time-series data through microbial temporal abundance datasets and have enhanced the knowledge about these complex and diverse communities through a pattern recognition approach.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

"I'm not talking about people [...] I'm talking about bacteria, the real rulers of the Earth."

—Sean Williams,
The New Venusians

In this Chapter, I introduce the motivation behind my research in using pattern recognition techniques for analysing complex microbial data and provide an outline of the thesis. An introduction to the field is given in Section 1.1. The manuscripts which were prepared during my PhD candidature are listed in Section 1.3 and an outline of the remaining chapters are given in Section 1.2. Lastly, Section 1.4 summarises the contributions of the thesis.

1.1 Background

Given the interdisciplinary nature of this research, I provide a background to the microbiology aspect and data analysis aspect separately. In Section 1.1.1, we will discover three questions we would like to ask about the microbiome, and how have they been addressed so far, and the benefits of these questions being addressed. Next, in Section 1.1.2, we will explore the nature of microbial data available to us and the main challenges we face in analysing it.

1.1.1 Understanding the Microbiome

Microbial life exists all around us. They live in water, air and earth. They live in harsh environments, just as they do in lush environments. They live in vast deserts, and they live in the guts of tiny insects. In fact, in the human body, the number of microbial organisms exceeds the number of individual cells of the human. Not only do they live in the human body, but they are known to affect the hosts in many ways [DeSalle and Perkins, 2015]. Because of these far-reaching effects of microbial communities, it is essential that we study their behaviour. Then, the first question to arise is, what "should" we study about microbial life. Boon et al. [2013] provide an

exciting direction for this. They separate this quest for knowledge into three different questions.

- Who is there?
- What are they doing?
- How will they respond [to external stimuli]?

They further suggest shifting the focus of exploration from the first two questions into the third. However, gaps still exist in our understanding of the second question, "What are they doing?". In this thesis, I focus on exploring these gaps and attempt to enhance the scientific understanding in answering the question of "what they [microbial organisms] are doing [in their communities]".

1.1.1.1 Who is there?

The first question proposed by Boon et al. [2013] is arguably the most important one. This straightforward question inquires about the composition of a microbial community. The seemingly simple task of visually observing the microbial community and identifying its constituents is not practical, as it is impossible to discern this visually. Due to the inability to visually observe, we rely on data-driven methods to answer the question of "who is there [in those communities]". Fortunately, DNA sequencing methods have been greatly improved in the recent past [Reis-Filho, 2009]. 16S rRNA sequencing is commonly used for this task. However, with microbial life, we face an additional challenge—microbial life evolves faster than macro life [Pepper, 2014]. Hence, taxonomic differentiation is unclear. To address this challenge, the commonly accepted method is to consider Operational Taxonomic Units (OTUs).

Simply put, OTUs are clusters of similar sequence variants identified through processing 16S rRNA data. Usually, the sequences are clustered at a 97% identity threshold. The pipeline for obtaining OTU data from 16S rRNA data is as follows. Firstly, sequencing data–DNA contig data–are preprocessed to ensure quality. This step usually involves combining forward and reverse reads into contigs and cleaning data to remove ambiguous reads and reads that deviate significantly from the expected read length (250 base pairs). Secondly, unique reads are identified, aligned, and re-cleaned. Thirdly, chimaeras (any combined sequences or sequences with anomalies) are removed. Lastly, these reads are clustered and matched with known reads to the satisfaction of 97% identity [Hiltemann et al., 2019]. 16S rRNA sequencing methods, also referred to as Next Generation Sequencing (NGS), are preferred over whole-genome sequencing as they are faster and less costly [Reis-Filho, 2009].

1.1.1.2 What are they doing?

We are simply not satisfied with only knowing the composition of a microbial community. We are also interested in the activity within these communities. To illustrate the point, let us take the example of a microbial community in a human gut.

Just knowing the composition of the gut microbial community is illuminating, but it is only a start. In a clinical context, the activity of these microbial communities would also be of great interest. We noted earlier that microbial community activity is not directly observable. The barriers to direct observation raise the complexity of the second problem—"What are they doing"? Simply expressed—it is practically impossible to discern the nature of microbial activity precisely. However, we can approximate, infer, and make educated guesses about what is happening within these communities.

Microbial 'activity' is considered to be a temporal concept (Quevauviller [2004]). Exploiting our knowledge from the answer to the first question, we can consider microbial activity to be reflected in the temporal change in the composition of microbial communities. To approximate microbial community activity, microbial communities are sampled at regular intervals. The tabulation of microbial community composition against time is what I refer to as an abundance profile of a community. To illustrate with the previous example, to see what 'they [microbes] are doing' in a human gut microbial environment, we can sample faeces from a host body at regular intervals and sequence it to get the microbial abundance profile. This microbial composition, which varies with time, provides us with an indication of microbial activity. The abundance profile itself is useful to see the compositional change in the microbiome. However, we are interested in exploring further. Some questions we are interested in asking are—How did these temporal changes come to be? What made one OTU's abundance increase while the abundance of another decreased? Are OTU behaviours inter-connected to each other? Many similar questions have been raised in the literature. The entirety of this thesis is on exploring this second question of 'What are they doing?' through microbial abundance profile data.

There are multiple approaches to process the data from abundance profiles to infer microbial activity. One popular strategy is to infer quantified microbial interactions by studying the abundance profiles Shaw et al. [2016]. A summary introduction to this process is to consider the abundance profiles to reflect the actual interactions. Any change in an OTU abundance is attributed to an interaction with another OTU or the environment. Observing the parallels between macro and microbiological communities, we apply ecological models in inferring these interactions. Analysing the inferred interactions, we delve further into quantified interactions by exploring the temporal changes in the interactions themselves. Having performed a primarily quantitative analysis to answer the question of 'What are they doing?', we move on to qualitative analysis of microbial community activity beyond numerical evaluations by visualising the data obtained through microbial abundance profiles. In this analysis of microbial activity, we consider the heterogeneity of microbial activity and consider concepts of individuality and conformity in microbial behaviour. Finally, we explore visual representations of microbial activity, visualising the answer to 'What they are doing?'.

1.1.1.3 How will they respond [to external stimuli]?

Now that we know how they [microbes] interact and what they are doing in their communities, we want to explore further about them. Answering the first two questions would, of course, make us understand better what is happening in the microbial communities; our curiosity exceeds that. We can gain an even deeper understanding if we figure out how the microbial life would respond to external stimuli — allowing us to engage in measured interactions with the microbiome around us. To illustrate the excitement of the ability to interact with the microbiome, we only need to look at how the discovery of penicillin changed contemporary medicine enormously (Ligon [2004]). We were able to change the composition of our body's resident bacteria. We now have in our arsenals not just antibiotics but also probiotics. What if we can eliminate harmful bacteria and support useful bacteria? What if we can come up with individually tuned concoctions of pro/anti-biotics so that we can achieve the perfect human gut microbiome? These are the motivations for us to explore a microbial community's response to external stimuli. How to do that is unfortunately out of the scope of this thesis. However, it would be exciting follow-up work. Hence we will briefly revisit this for a discussion at the end of the thesis.

1.1.2 Challenges Present in Microbial Data

In Sub Section 1.1.1 we talked about understanding the microbiome through data. Specifically, we talked about the deluge of data related to the microbiome and how we use data analysis methods to further our knowledge about microbial community activity. This Sub Section will discuss what makes this analysis challenging due to various aspects of the available data.

1.1.2.1 Data Complexity

We consider the data we receive in the format of microbial abundance profiles to be complex data. 'Complex' is an umbrella term used to describe data that has several – usually unfavourable – characteristics. The time-varying nature and heterogeneity, which are discussed next, are also characteristics of complex data. Apart from this, the large size of data, unavailability of specific data points (missing data), connectivity within data points indicating an underlying structure, inclusion of significant noise, irregular sampling when considering time-series data, high dimensionality, and continuously growing nature are some other characteristics of complex data. The above-mentioned characteristics can be observed in microbial abundance data sets, necessitating the consideration of such complexity in the process of data analysis.

1.1.2.2 Data Heterogeneity

Heterogeneity is the diverse nature of various aspects of data. In microbial data, we observed heterogeneity in multiple aspects (González-Cabaleiro et al. [2017]). Firstly,

microbial communities are composed of different microbial taxa, which creates a degree of heterogeneity in their activities. Within these taxa, we categorise OTUs according to their abundance. OTUs with high abundance (>1%), OTUs with low abundance (<1% but >0.1%), rare OTUs (<0.1%) are usually considered separately (Shaw et al. [2018]). Even within the same taxon, cellular heterogeneity may exist, which will translate into behavioural heterogeneity. Since we are considering temporal data, we are also open to observing behavioural changes with time. Temporal heterogeneity may arise due to changing environmental factors or the fast evolution of microbial communities, depending on the timelines we are considering. Chapters 3, 4, 5, and 6 all consider heterogeneity of the data in the analysis and discussions.

1.1.2.3 Analysing Time-Series Data

The temporal nature of the data brings its inherent challenges in addition to already existing challenges of complexity and temporal heterogeneity. Biological data is usually prone to temporal noise (Tsimring [2014]). Causes for these can range from irregular data collection, changing environmental conditions affecting populations and equipment, as well as inherent uncertainties. Apart from this, modelling temporal data with ecological models – which I do in the Chapters 3 and 4 – is prone to accumulate errors with time. In my work, I attempt to alleviate some of these adverse effects by utilising techniques suitable for time-series data.

1.2 Thesis Outline

An overview of the thesis outline is illustrated in Fig. 1.1. The following Sub Sections detail each Chapter.

1.2.1 Chapter 1

This chapter provides a broad introduction to the overall thesis, including the motivations, challenges and contributions of the thesis.

1.2.2 Chapter 2

Chapter 2 provides a detailed background to the problem of analysing microbial community dynamics and contains a compilation of reviewed literature for each of the research questions.

1.2.3 Chapter 3

In Chapter 3, I present IMPARO: Inferring Microbial interaction networks through PARameter Optimisation, a novel method for inferring microbial interactions. This Chapter also discusses the traditional assumptions involved with inferring microbial interactions and how we can loosen the assumptions to improve the quality of the

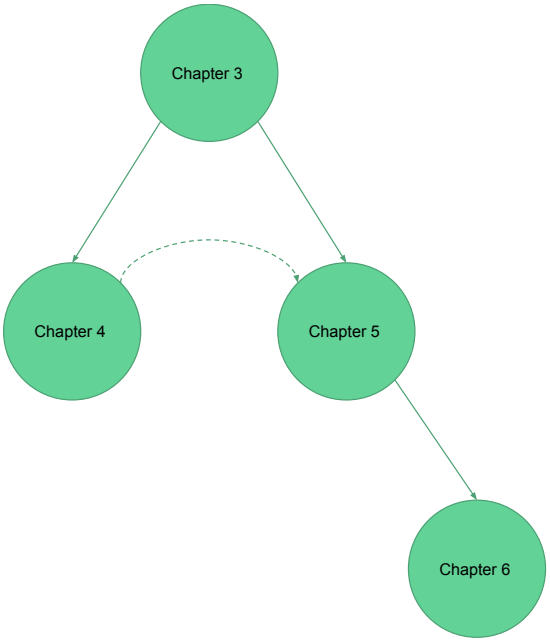

Figure 1.1: Illustrated thesis outline. Chapters pertaining to research questions are represented by circles, while logical flow of ideas is represented with solid arrows. Dashed arrows indicate lateral influence.

inferred Microbial Interaction Networks. We further discuss the ambiguity in Microbial Interaction Networks inferred through model-based and statistical methods and the possibility of multiple solutions. The findings presented in this Chapter resulted in a journal publication (Vidanaarachchi et al. [2020]).

1.2.4 Chapter 4

In Chapter 4, I present an exploration of the temporal dynamics of microbial interaction networks, arguing that microbial interactions themselves are dynamic. We discuss the importance of considering not only the composition but also the underlying network of a microbial community to be temporally dynamic for an accurate representation of the community dynamics. We discuss alternate modelling approaches suitable for temporally dynamic microbial communities and show probable evidence for heterogeneity in the said dynamics. The findings presented in this Chapter resulted in a publication in IEEExplore indexed conference proceedings (Vidanaarachchi et al. [2019]).

1.2.5 Chapter 5

In Chapter 5, we take a step back and look at the nature of the data available and how to develop robust methods which will be capable of achieving qualitatively and quantitatively superior results by collectively processing multiple datasets. This

Chapter also introduces our novel visualisation pipeline, which provides insights into microbial organisms' individualist and conformist nature. I present results that illustrate how some OTUs tend to be limited to a small number of Temporal Variation of Abundance Profile) TVAP patterns amongst other particular trends. I also raise the idea that sometimes the TVAP patterns may be connected to clinical factors. The findings of this Chapter has been compiled into a manuscript aiming for a journal publication.

1.2.6 Chapter 6

In Chapter 6, we explore the heterogeneous nature of microbial dynamics and visual representations of these high-dimensional heterogeneous time-series data. I raise exciting questions regarding unique representations of OTU dynamics and propose interpretations to dimensionally reduced incremental OTU signatures. Initial findings of this exploration has been compiled into a journal manuscript. However, the visualisation techniques presented here warrant further interpretation and are ideally tested with a varied collection of microbial datasets.

1.2.7 Chapter 7

Finally, Chapter 7 offers a summary of the thesis, my conclusions and identified future work in this field of study.

1.2.8 Chapter Outlines

In Chapters 3, 4, 5 and 6, I adhere to a uniform structure as follows:

Background	These sections contextualise the research question, present the previous work done pertaining to the question, and explain my motivations and contributions.
Results	These sections provide the results that I have obtained in each study.
Discussion	These sections provide an in-depth discussion of each study, including identified future work.
Conclusion	These sections briefly summarise each study and present their conclusions
Methods	Please note that following the conventions of bioinformatics reporting, we provide the methods sections at the end of each chapter. These sections contain a thorough explanation of the computational methods used in each study and the datasets and terminology where required.

1.3 Manuscripts

Some material in this thesis proposal is contained in the following manuscripts.

- Vidanaarachchi, R., Shaw, M., Tang, S.L., and Halgamuge, S.K., 2020. IMPARO: Inferring Microbial Interactions through Parameter Optimisation. BMC Molecular and Cell Biology. (Vidanaarachchi et al. [2020])
- Vidanaarachchi, R., Shaw, M., Halgamuge, S.K., 2019. Computational Inference of Microbial Interactions and their Dynamics. 14th IEEE International Conference on Industrial and Information Systems (ICIIS). (Vidanaarachchi et al. [2019])
- Vidanaarachchi et al. 2021. CoPR: Collective Pattern Recognition—a Framework for Microbial Community Activity Analysis. (Manuscript being finalised for submission to a journal)
- Vidanaarachchi et al. 2021. Incremental Microbial OTU Signatures. (Manuscript being finalised for submission to a journal)

My contribution is between 90%–95% for the four manuscripts listed above. I conceptualised the research problems, implemented and tested codes, carried out the experiments and wrote the initial manuscript. My PhD supervisors, who are the co-authors of the manuscripts, advised me to develop and formalise the ideas, frame the concept, and improve the presentation of the reports.

The following manuscript was prepared in collaboration during my PhD candidature but did not form a part of this thesis.

- Faleel, A., Vidanaarachchi, R., Shaw, M., Halgamuge, S.K., 2021. From The International Space Station To Tropical Rainforests And Polar Ice Caps: Microbial Communities Foretell The Effects Of Climate Change. 10^{th} IEEE International Conference on Information and Automation for Sustainability (ICIAfS) *(Pending Publication)*

1.4 Contributions

My main contributions in this thesis are two-fold. Firstly, I apply data analysis methods and build pipelines for the analysis of longitudinal microbial abundance data. Secondly, I present novel ideas to push the boundaries of modelling dynamic microbial communities.

Pertaining to data analysis, firstly, I use a genetic algorithm (GA) to optimise the answer to a question for which the ground truth is not known. I highlight the importance of not discarding the non-optimal answers in the GA and show how they can be potentially valid answers within a reasonable error margin given the noise-addled nature of our data. Secondly, I critically evaluate the shortcomings of the

Lotka–Volterra model for temporally dynamic biological processes and discuss possible alternatives. Thirdly, I introduce a pattern recognition pipeline for parallel sets of heterogeneous data and engage unsupervised learning methods for information extraction. Lastly, I explore methods of representing heterogeneous, high dimensional time-series data in unique signatures.

My contributions pertaining to systems biology are as follows. Firstly, I describe the current assumptions involved in inferring microbial interaction networks (MINs). Then I explore the possible scenarios for relaxing the assumptions involved and whether the relaxation of said assumptions can indeed improve the results quantitatively and qualitatively. I also reiterate the fundamental issue of assuming a unique, verifiable solution to the problem of inferring microbial interactions. Secondly, I question the use of first order differential equations to model microbial communities and suggest that they may be better modelled through second order differential equations. Although Lotka-Volterra equations have been used for more than a century to model ecological community interactions, in the scope of microbial communities, assuming the interaction parameters themselves stay static can result in loss of essential information. Thirdly, through the results of the visualisation, I question the strategy behind striving to find a unique pattern in OTU temporal variation patterns. I contrast the unique pattern approach, where I obtain multiple signals in an unsupervised approach. I also discuss the concepts of individuality and conformity in the context of microbial community dynamics. Lastly, I explore the options for obtaining a unique signal for OTU community dynamics.

Chapter 2

Related Work

In this chapter, I present the existing body of literature under four main themes. Throughout this thesis, I have drawn inspiration not only from the literature of microbiology and bioinformatics but also from the literature pertaining to other complex systems. In Section 2.1, I introduce the literature pertaining to the inference of static microbial interaction networks in a single environment. In Section 2.2, I explore the studies on time-varying systems. Then, in Section 2.3 I discuss the literature and ideas related to collective pattern recognition. Lastly, in Section 2.4, I present existing literature on characterising temporal processes in general and the dynamic characterisations of the microbiome.

2.1 Microbial Interaction Networks and their Inference

In this section of our literature survey, we broadly cover the topics related to Chapter 3. First, we will look at the microbial interaction network inference as a whole in Section 2.1.1. Then we will look at methods that we can use in inferring microbial networks through parametric models in Section 2.1.2. We think this approach is the most suitable as most microbial interaction inference methods use some form of the Lotka-Volterra (LV) equations to model the interaction networks.

2.1.1 Inference Approaches

Table 2.1 summarises previous work of ten microbial interaction inference algorithms. More extensive details about each of these methods are presented in Section 3.1.1. Most methods utilise a model-based approach rather than relying on correlation based analysis. LIMITS [Fisher and Mehta, 2014] present proof for correlation not being equal to the interaction between species. A majority of methods test on both real-life and simulated data. As mentioned above, most methods also use a variant of the LV equations to model the interaction network. Durán et al. [2021] shine a new light on the problems discussed in this thesis in their network analysis pipeline using machine learning technologies. They successfully challenge linear techniques for multivariate analysis and reveal bacterial network reorganisations due to external factors.

Table 2.1: Review of existing algorithms for inferring microbial interactions

Algorithm	Publication	Biological Model	Statistical / Optimisation Approach	Verification	Real Life Data	Simulated Data	Nature of Output
Unnamed	Barberán et al. [2012]	*Not Applicable—Statistical Method*	Checker-board Score	Non-Randomness	Soil microbiome	*Not Available*	A Co-occurrence Network
SparCC	Friedman and Alm [2012]	*Not Applicable—Statistical Method*	Approximating Correlations	Root Mean Squared Error	Human Microbiome Project	By sampling joint abundances of real life data	A Correlation Network
LIMITS	Fisher and Mehta [2014]	Discrete-time Lotka Volterra (dLV)	Linear Regression	R^2	Moving Pictures of the Human Microbiome	dLV simulation	Interaction matrix
RMN	Tsai et al. [2015]	OTU Triplets	Tanh Functions	Accuracy (Binary Classification)	An infant gut data set	Tanh simulation	A Microbial Regulatory Network
SPIEC-EASI	Kurtz et al. [2015]	*Not Applicable—Statistical Method*	Stability approach to regularisation selection	Precision Recall Curves	American Gut Project	Based on American Gut Project Data	Microbial Associations
Boolean Dynamics Model	Steinway et al. [2015]	Subsystem Enrichment Model	Perturbation Analysis	*In-vitro* Validation	A data set of *Clostridium difficile* infection in the gut	*Not Available*	Infers specific interactions

continues on next page

Table 2.1: Review of existing algorithms for inferring microbial interactions

Algorithm	**Publication**	**Biological Model**	**Statistical / Optimisation Approach**	**Verification**	**Real Life Data**	**Simulated Data**	**Nature of Output**
MetaMIS	Shaw et al. [2016]	Lotka Volterra (LV)	Partial Least Square Regression	Recreation of the abundance profile and Bray-Curtis dissimilarity	Moving Pictures of the Human Microbiome	*Not Available*	Consensus Network
Boolean Abundance Analysis	Claussen et al. [2017]	Competitive Synergistic Links (CSL)	Entropy Shifts on Abundance vectors under Boolean Operators (ESABO)	Jaccard Index	A human gut data set	CSL simulation	Species Interaction Network
SgLV-EKF	Alshawaqfeh et al. [2017]	Stochastic Generalised LV (SgLV)	Extended Kalman Filter	Mean Squared Error	A data set of *Clostridium difficile* infected murine gut	SgLV simulation	Microbial Interaction Network
Unnamed	Gao et al. [2018]	Generalised LV	Forward Stepwise Regression	Previously known interactions	A cheese microbial community	*Not Available*	Combined Interaction Network
IMPARO	Vidanaarachchi et al. [2020]	Generalised LV and Community Dynamics Model	Genetic Algorithm	Recreation of abundance profile and Bray Curtis dissimilarity	Moving Pictures of the Human Microbiome	GLV and Community Dynamics Model simulation	Microbial Interaction Network

2.1.2 Parameter Optimisation

When modelling microbial interaction networks with Lotka–Volterra equations (The LV equations are detailed in Chapter 4), we are faced with a large parameter optimisation problem, as the number of parameters in an LV model is quadratic to the number of taxa represented in the model. Some methods use regression to estimate the parameters, while others choose evolutionary algorithms. In this section, we will summarise some of the available options.

One of the most straight forward approaches to optimise parameters is the use of regression analysis. Linear Regression, Partial Component Regression (PCR), and Partial Least Squares (PLS) are some methods that can be used. Apart from these, Bayesian networks and various evolutionary algorithms have been used. As these are well-established methods, we will not go into great detail about them in our literature survey.

2.2 Behavioural Dynamics of Microbial Interaction Networks

In this section, which broadly covers the topics related to Chapter 4, we discuss the literature related to behavioural dynamics of microbial activity. Firstly, we will discuss the literature on time-varying systems in Section 2.2.1. Secondly, we will summarise literature pertaining to the dynamics of microbial communities in Section 2.2.2.

2.2.1 Time-Varying Systems

As there is a limited amount of studies exploring the microbial interactions' temporal dynamics, we will first look at time-varying interactions and their inference as recorded in other areas. Primarily we will consider literature from the area of gene expression analysis, as there are similarities observable in gene expression analysis and microbial interaction analysis.

Song et al. [2009b] present one of the early works in modelling time-varying processes. They propose time-varying dynamic Bayesian networks (TV-DBN) to model the structurally-varying directed dependency structures. Applying their method to yeast cell cycle gene expression datasets, they also mention the suitability of their method for data with sample scarcity. Data scarcity, again, is a common challenge gene expression data analysis shares with longitudinal microbial data analysis.

Song et al. [2009a], the same authors from the above study present another method—a kernel re-weighted logistic regression method—to reverse engineer dynamic interactions between genes based on their longitudinal expression values. Again, this is a very similar problem to the analysis of dynamic microbial interactions.

Kolar et al. [2012] present a very interesting study in estimating time-varying networks. They introduce two methods based on temporally smoothed l_1-regularised logistic regression approaches. They study gene expression data from *Drosophila*

melanogaster and temporally re-wiring political networks by examining voting patterns in the US Senate.

The three methods discussed above are all applied in gene expression network analysis. The methods vary from dynamic Bayesian networks to logistical regression. We note that the results are analysed in the context of biology to determine whether the inferred gene expression networks make sense—this validation issue is also present in the OTU interaction inference problem. However, translation of this into the area of microbial interaction networks has been lacking.

2.2.2 Temporal Dynamics of Microbial Interactions

There is limited work done on the temporal dynamics of microbial interactions. Most work, as we discussed in Section 2.1 focus on the temporal dynamics of the microbial composition. We will highlight a few key papers which we consider to be important.

In their review of microbial interactions Layeghifard et al. [2017] considers time-varying microbial interactions as well. In their conclusion, they propound that the use of network biology is still in its infancy. They propose that the complexity of conceptualisation as well as implementing suitable models contribute to this. They suggest reexamining the interaction inference processes—including the data collection—to suit novel network modelling practices.

Faust et al. [2015] highlight the importance of implementing time-varying networks constructs to infer microbial interactions. They also suggest building static networks, which infer the interactions in overlapping time-segments as a starting point. They also point to results from an analysis of the human gut microbiome, which shows that taxon associations, including their stability and strength, vary over time.

Hosoda et al. [2021] introduce unsupervised learning-based microbial interaction inference method using Bayesian estimation (Umibato) to improve the estimation of time-varying microbial interactions. They use Gaussian process regression and a continuous-time regression hidden Markov model. This method, so far, is the best work available in the field of dynamic microbial interaction inference.

2.3 Concepts of Collective Pattern Recognition

In this section, which broadly covers the topics related to Chapter 5, we discuss the literature related to collective pattern recognition. Firstly, we summarise the literature pertaining to the availability of datasets in Section 2.3.1. Secondly, we discuss the literature about the concept of collective pattern recognition and computational algorithms used in various approaches in Section 2.3.2. Thirdly, in Section 2.3.3, we explore the literature on the concepts of individuality and conformity, as this encompasses the central philosophy behind Chapter 5. Lastly, we look at the literature to discover the potential applications of our research in the field of precision medicine in Section 2.3.3.2.

2.3.1 Availability of Datasets

In any data analysis problem, it is essential that we explore the availability of data and the nature of available data prior to the analysis. In solving the problems of understanding microbial communities' temporal behaviours, the analysis heavily depends on the available data. Even with the advances in technology, some data collection approaches are too costly or too time-consuming to be practical. Primarily as we rely on time-series data, the quality of available data could vary.

2.3.1.1 Favourable Qualities of Time-Series Data

Time-series data collection, in other words, record keeping, has been done for millennia. Especially with the advent of naval trade, regular numeric data collection has been a common practice [Wilkinson et al., 2011]. However, not all datasets are collected equally. We will look at some aspects of time-series data that affect their quality.

Numerousness For time series data analysis, it is of utmost importance that we have a significant number of data points be available for the analysis. In fact, Hanke and Wichern [2013] suggests at least 50 data points available for any time-series data analysis task. This number may vary depending on the nature of the research problem at hand, but the quantity of data is essentially important.

Sampling Frequency The next important quality of a good dataset is that the sampling has been performed at a satisfactory period. It is hard to pinpoint an exact number, as this varies vastly with the nature of the problem [Nason et al., 2017]. In the context of microbial abundance data, the nature of the dynamics we capture depends on the sampling frequency. For example, if we only sample a gut microbiome daily, we will not be able to identify the diurnal effects. However, higher sampling frequencies are practically impossible sometimes, and we have to rely on interpolation methods [Castiglioni et al., 2003].

Consistency of Sampling Even when the data is numerous, a relative consistency of sampling intervals is helpful. Unfortunately, with biological data achieving this consistency is more complicated than with data from the fields such as physics or economics. Studies such as Nason et al. [2017] suggest that inconsistent sampling is more of a problem pertaining to classical time-series analysis and that a regular sampling frequency is not essentially required for Bayesian and other model-based methods. Nevertheless, we find that microbial abundance data sampling rates vary heavily. We will illustrate some examples in the following sections.

2.3.1.2 Longitudinal Microbial Datasets

In this section, we explore some of the temporal microbial abundance datasets available and discuss the suitability according to the criteria we described above.

Moving Pictures of the Human Microbiome Published in Caporaso et al. [2011], Moving Pictures of the Human Microbiome is one of the largest human microbial time-series datasets available. Their data comes from two individuals—a male and a female. The microbial samples are collected from four body sites—faecal matter, mouth, left and right palms. The data from the male subject were collected for 15 months, with an almost daily sampling rate. The data from the female were collected for six months at the same rate. Considering this information, we can conclude that this dataset is numerous and has a relatively consistent and high sampling frequency making it an ideal dataset for analysis.

American Gut Project The American Project [McDonald et al., 2018] is an open platform for citizen science, where individuals had sampled themselves and sent in their microbial samples for analysis to the authors. The sample collection and dispatch were performed at regular temperatures. Their impressive collection has microbial sequencing data from over 15000 samples from more than 11000 participants from all over the world. Their collection includes longitudinal data from individuals who have contributed multiple samples over different time intervals. Again this is a fascinating dataset purely due to the vast amounts of data collected. As for time-series studies, the sampling was not highly frequent not consistent.

Murine Gut Microbial Samples Marino et al. [2014] collected gut microbial samples from five germ-free adult female mice. At the beginning of the study, caecal contents from an adult mouse were homogenised at inoculated into the female mice. Then the mice's faecal matter were obtained daily over 21 days. The collected samples were immediately frozen as well. This study is relatively short. However, the data collection process minimised external influences and was fairly consistent.

An Infant's Gut Microbiome In their study, Koenig et al. [2011] collected more than 60 faecal samples from an infant—from birth up until the age of 2.5 years. Furthermore, this infant's dietary, social and clinical information was recorded to study the impact of life events on the microbiome. The male infant, delivered via vaginal delivery, was immunised and healthy other than an ear infection. The report recorded that he was given antibiotics in that incident. Also of interest is the change in diet from breast milk, formula milk, solid food, etc. This study is interesting as it collects background information on potential events which could affect the microbiome. This study also illustrates the difficulty of adequately quantifying 'pure' microbial interactions from experiments conducted *in situ*.

Smear Cheese Microbiome In this interesting experiment by Mounier et al. [2008], they analysed the microbial composition of the smear cheese ripening process. Inoculating pasteurised milk with a frozen starter culture which consisted of nine known microorganisms isolated from various cheese batches, this experiment also

shows the advantages of a controlled *in vitro* environment in analysing the microbiome. The cheese was sampled daily for 21 days in duplicate. In a later stage of the experiment—after five months, they again sample the cheese on four different days. This dataset is an example of the inconsistent sampling approaches present in microbiology, which introduces difficulty in modelling the microbial dynamics.

Oropharyngeal Microbiome In this study on the human oropharynx, Bach et al. [2021] sampled the composition of the oropharyngeal microbiome of 18 adults at a weekly interval for 40 weeks. The otherwise healthy adults self-reported disease during the time frame of the experiment, which would provide interesting insights regarding external/environmental factors. This research generated OTU tables with 97% similarity. The researchers provide a Phylum-level analysis of the data as well.

Bodily Microbiome of Horses O'Shaughnessy-Hunter et al. [2021] present a study of the bodily microbiome of horses, collected from four body sites — dorsum, ventral abdomen, pastern and groin. The study has included 12 healthy horses from the same farm. This study is interesting because the four samples were collected in four seasons over a year, reflecting the weather's effect on the microbiome. This study also uses 16S rRNA data and identifies OTUs based on a 97% similarity. Unfortunately, the sampling frequency and the number of samples are lacking for an adequate longitudinal analysis.

2.3.1.3 Parallel Collections of Longitudinal Datasets

In the previous section, we explored some datasets and discovered the difficulties in collecting the perfect dataset for microbial behavioural analysis. This difficulty is why we are interested in studies where data is collected for a shorter duration but in several similar host environments. In this section, we will explore several dataset collections. In these collections, if we were to select a particular dataset representing one microbial environment, it would be insufficient for the needs of our analysis. However, as the collection contains parallel data from a significant number of parallel sources, we can augment the lack of individual data. Availability of datasets of this nature was crucial for our work in Chapter 5.

Premature Infants' Gut Microbiome La Rosa et al. [2014] compiles a dataset from 58 premature infants' gut microbiome during a stay at a neonatal intensive care unit (NICU). The NICU is a controlled environment where exposure to external microbiota is minimal. They have collected 922 samples averaging just over 15 data points per infant. However, the data sampling depended on the stool passing of the infants, rendering it non-consistent. The study also reported other clinically relevant information as well. This collection of datasets is a perfect example of data augmentation with parallel datasets. Usually, the low number of samples and the inconsistent rate of sampling would be a deterrent. However, with a high number of parallel datasets (58), we can look into the prospect of recognising patterns collectively.

Vaginal Microbiome of Reproductive-Age Women Gajer et al. [2012] presents a collection of datasets from 32 reproductive-age women's vaginal microbiome. The samples were collected twice weekly for 16 weeks, with 937 samples averaging 29 per woman. This data is also accompanied by longitudinal background data on vaginal activity in terms of health and behaviour. Again, this is an excellent dataset where we can use collective pattern recognition approaches. The longitudinal background data provides essential information in aligning the datasets according to key events—such as menstrual cycles (in the context of this collection).

Human Microbiome Related to Pregnancy DiGiulio et al. [2015] present a comprehensive study of the human microbiome related to pregnancy. They collected a total of 3767 samples from 49 women during and after pregnancy. The collection sites included the vagina, distal gut, saliva and tooth/gum—the study averages 20 samples per site per woman. Comprehensive longitudinal studies such as this are essential to recognise the connections between the behavioural dynamics of multiple body sites' microbial communities. It is also an interesting dataset as the data has been collected around an important life event—pregnancy. In their analysis, the authors identify specific microbial trends associated with pre-term birth.

Neonatal Gut and Respiratory Microbiome In another comprehensive study, Grier et al. [2018] presents data from 82 infants, both pre-term and full-term, collected from the gut, nasal cavity and throat. The data were collected weekly during their hospital stay and monthly after discharge. This study, which we use for our analysis in Chapters 5 and 6 is an interesting dataset due to the high number of similar subjects and the distinct nature of the body-sites. While the gut microbiome is primarily anaerobic, the respiratory (nasal and throat) microbiome is aerobic. However, the average samples per body site are limited, with 13 samples on average for the gut, 12 and 6 for the nasal and throat environments.

Goat Kids' Gut Microbiome In this study by Zhuang et al. [2020], the researchers collected eight samples of the gut microbiome of 48 goat kids. The samples were collected almost weekly, from birth until post-weaning. The selected time frame allowed the researchers to identify gut microbial compositions during the colostrum phase, breast milk phase, combined breast milk and starter feed phase, and the starter feed only phase. OTUs were calculated at the 97% similarity rate by analysing 16S rRNA sequencing. This study consisted of 48 healthy goats. Again, this dataset is exemplary to study with collective pattern recognition techniques, as the eight sampling points would otherwise limit identifying temporal dynamics.

Supragingival Plaque Microbiome Study of Twins In this experiment by Freire et al. [2020], supragingival plaque samples were collected from 70 sets of twins and one set of triplets (totalling 143 participants). There were 62 monozygotic, 36 opposite-sex dizygotic, and 45 same-sex dizygotic participants. They ranged from

5.5–12 years in age with a median age of 9. The sampling frequency is very low at six-monthly intervals, and only three sampling points are available. 16S rRNA sequencing was used, and OTUs were generated directly from raw reads. Although this is a fascinating study with many participants, it is difficult to carry a longitudinal microbial analysis due to the lack of sampling points.

Faecal Microbiome of Humans with Inflammatory Bowel Disease Clooney et al. [2021] presents a study on the faecal microbiome of humans with Inflammatory Bowel Disease. This study boasts a reasonably large sample size of 692 individuals (303 with Crohn's Disease, 228 with ulcerative colitis, and 161 controls). Unfortunately, the sampling rate is at 16-week intervals, with a total of three sampling events per individual. They also collect interesting background data including, geographic locations, surgical histories, alcohol consumption, medication, and diets. However, they also report that the compositional variance is unexplained with regard to the background and clinical data. Microbial data were collected with 16S rRNA sequencing.

2.3.2 Collective Pattern Recognition Approaches

In this section, we will discuss, compare, and contrast the existing approaches where we identify the notion of collective pattern recognition. We will also explore common aspects of these approaches and the algorithms that have been utilised.

2.3.2.1 Existing Methods

There is existing literature that exploits the availability of multiple datasets to improve the inferring processes. However, most such methods seem to be used in analysing gene expression datasets. As we note in several points of this thesis, methods used for gene expression data and methods used for microbial abundance data share some commonalities. We have provided a summary of these methods in Table 2.2.

2.3.2.2 Curve Fitting

Curve fitting approaches are a standard part of many collective pattern recognition approaches. Biological datasets are usually sparsely and inconsistently sampled. In a collection of datasets, sampling points rarely overlap each other. Hence, a curve fitting approach is taken in most methods we reviewed, with the notable exception of Chandereng and Gitter [2020]. In this section, we will discuss some popular approaches to curve fitting.

Interpolation Interpolation estimates the points in-between the sampled points by approximating a polynomial. For microbial data, this would usually be a higher degree polynomial, which allows capturing temporal trends.

Table 2.2: An overview of a selection of collective pattern recognition approaches

Algorithm	**Publication**	**Datasets Used**	**Curve Fitting**	**Alignment Methods**	**Other**
Unnamed	Lugo-Martinez et al. [2019]	Longitudinal microbial abundance data	B-Spline	Using a first degree polynomial function	Combined with a Dynamic Bayesian Network
genewarp	Aach and Church [2001]	RNA and protein expression data	Can be use with or without interpolation	Inflated Alignment Score	Presented as a package of executables
TimeFit	Bar-Joseph et al. [2003]	Time-series gene expression data	B-Spline	Using a first degree polynomial function	Clustered using k-nearest neighbours
Unnamed	Smith et al. [2009]	″	*Multi-segment method*	Shorting Correlation Optimal Warping (SCOW)	Clustered using k-means
GenT$_X$*Warper*	Criel and Tsiporkova [2006]	″	*None*	Dynamic Time Warping	*Not Applicable*
Unnamed	Dong et al. [2020]	″	*Not Applicable*	Bayesian Multiple Kernel Learning	Introduced as a statistical framework for predicting viral exposure
Unnamed	Somani et al. [2020]	″	Gaussian proccess interpolation	Central Composite Design	Identifies disease-relevant pathways
Lag Penalised Weighted Correlation (LPWC)	Chandereng and Gitter [2020]	Gene expression and Protein phosphorylation data	No interpolation as a design decision to conserve originality of data	LPWC	Adjusted Rand Index (ARI) clustering
TimeMeter	Jiang et al. [2020]	Time-series transcriptomic data	*Unspecified*	Dynamic Time Warping	Uses four alignment quality metrics

Regression Regression differs from interpolation by finding the best line that fits the available data points rather than approximating the values in-between the data points. Regression can be used with higher degree polynomials or non-linear functions.

B-spline B-spline, or basis spline a type of interpolation where the polynomial is piece-wise. The spline usually contains low-degree polynomials which fit smoothly together [De Boor, 2001].

LOESS Locally Estimated Scatterplot Smoothing (LOESS) is a local regression method built on classical regression methods, with local weighting, based on the concepts of the Savitzky-Golay filter [Savitzky and Golay, 1964]. It obtains a smooth plot from a set of scattered data points while respecting local trends.

2.3.2.3 Alignment Approaches

In many collective pattern recognition approaches, an alignment of the different samples followed the curve fitting. Various temporal alignment approaches could be used for this step.

Linear Transformation A linear transformation is one of the most common approaches to aligning temporal datasets. They usually involve shifting and scaling on one or both axes. For the case of two datasets, this can be intuitively understood as applying a linear transformation for one dataset to minimise the difference between the two datasets. When considered for multiple datasets, the most common dataset is selected, and the rest are transformed to match the pattern of the selected one.

Dynamic Time Warp distance Dynamic Time Warp (DTW) distance calculates an optimal match between two different temporal datasets. This distance metric is helpful as it can identify matches between sequences that are temporally "warped" in a non-linear manner. Although it does not provide a direct alignment *per se*, this measure is a valuable tool to substitute alignment where a metric of similarity at the best alignment is required.

2.3.2.4 Clustering Approaches

Some collective pattern recognition approaches choose to cluster the data after alignment, depending on the study's aims. In our work of Chapter 5, we use clustering; hence, we will mention some clustering mechanisms.

k-means clustering A straightforward clustering algorithm, k-means, identifies k number of centroids in the data and allocates every other point to the closest centroid. This method is usually inadequate for most complex applications [Teknomo, 2006].

Gaussian Mixture Model Clustering A problem with k-means clustering is that it does not account for the variance of a dataset. In a two dimensional visualisation, k-means clustering places clusters in perfect circles. Gaussian mixture model (GMM) clustering is capable of identifying non-circular clusters as well. It employs an Expectation Maximisation (EM) algorithm in determining the cluster membership of data points. GMM clustering also has the advantage of providing the likelihood of cluster membership; hence it can also be called a soft classification method [Bilmes, 1998].

Mean Shift Clustering Mean shift clustering is a parallelisable clustering approach, with the added advantage of not needing a cluster number in advance. It uses a Kernel Density Estimation (KDE) mechanism to identify the densest areas of the data as peaks. Then, it lets the other datapoints assign themselves to one of the peaks depending on a metric of distance and the density of the peak [Comaniciu and Meer, 2002].

DBSCAN Density Based Spatial Clustering of Applications with Noise (DBSCAN) is a clustering approach proposed by Ester et al. [1996]. DBSCAN algorithm can be summarised in three steps. Firstly, it identifies data points within a certain radius to each point and considers them to be core points if they have more than a certain number of neighbours. Secondly, it finds connected components made up of core points. Thirdly, it associates every non-core points to the identified connected components (clusters) and disregards the rest as noise [Schubert et al., 2017]. This algorithm is excellent in recognising non-linear clusters.

OPTICS Ordering Points To Identify Cluster Structure (OPTICS) is an extension of DBSCAN which introduces two metrics for determining the cluster membership—core distance and reachability distance Ankerst et al. [1999].

2.3.3 Individuality and Conformity

Individuality is a quality, trait or behaviour that sets some entity apart from other similar entities. In contrast, conformity is the opposite, where some entity's qualities, traits and behaviours can be expected to be predicted according to set standards. The constant contrast between individuality and conformity has been discussed as a topic of philosophy for millennia. We will briefly explore literature pertaining to the concepts of individuality and conformity in this section.

2.3.3.1 General Notions

This concept has been applied in the field of social science to describe human populations [Mughal, 2014; Wilson, 2009]. Kingsbury [1997] presents the balance between individuality and conformity as an academic discussion in the area of law and compliance. There are many discourses of this pertaining to fashion, lifestyle choices etc.

However, these concepts have been noted in nature as well. The concept of individuality and conformity has been discussed in the literature of ecology [Hull, 1980] and made a comeback only recently in the field of microbiology [Montassier et al., 2018].

2.3.3.2 Notions in Microbiology

As mentioned earlier, a discussion on individuality and conformity in the microbiome is presented in the work of Montassier et al. [2018]. The individuality of microbial communities has been identified in the literature [Martins and Locke, 2015]. Especially in research on the gut microbiome, precision medicine has been proposed and successfully used in several studies De Filippis et al. [2018]; Cammarota et al. [2020]. Conforming behaviour has also been reported in the literature Gong et al. [2016]. Based on the ideas summarised above, we define individuality and conformity as a fuzzy concept in the context of microbial behavioural patterns.

Community State Types In the case of OTU communities, the concept of Community State Types Ravel et al. [2011] is an existing approach to explaining the balance of individuality and conformity in microbial communities. This concept has been used in many publications to date. Grier et al. [2018]; DiGiulio et al. [2015] and other studies report community state types (CSTs) in various microbial communities. The idea of CST is based on the composition of the constituent OTUs and is defined for a snapshot in time.

2.4 Characterisation of Temporal Dynamics of the Microbiome

In this section, which broadly covers the topics related to Chapter 6, we discuss the literature related to the characterisation of the microbiome (Section 2.4.1) and the characterisation of temporal dynamics in other systems (Section 2.4.2). Apart from this, we discuss the literature pertaining to the study of the characterisation of temporal dynamics (Section 2.4.3).

2.4.1 Characterisation of the Microbiome

The literature pertaining to the characterisation of the microbiome is immense. Hence we describe a selection of approaches that have been used to characterise the microbiome. Both static and dynamic characterisations are described here. We also discuss Community State Types (CSTs), which have been used successfully to characterise the dynamics of the vaginal microbiome.

2.4.1.1 Static Microbial Signatures

First, we should clarify that the microbial signatures here refer to the characterisation of the microbiome through compositional profiling, which is distinct to the meaning

we use in Chapter 6 where the signature refers to the low dimensional mapping. However, both are characterisations of microbial communities in their own rights.

In Desikan [2017] the author reports the use of the microbial composition of an environment as a uniquely identifiable signature for that environment. Various bodily microbial community signatures have been proposed to have to potential to identify humans individually [Tridico et al., 2014]. Banerjee et al. [2018] suggests the use of microbial signatures to identify the state of the host environment. In their work, they show that microbial signatures have specific associations with different types of breast cancer. Romero et al. [2014] show that the vaginal microbiota of pregnant women is different to that of a non-pregnant woman. These example studies show that even the microbiota's static composition is an indicative signature or a characterisation of the host environment.

2.4.1.2 Dynamic Microbial Signatures

Literature also has interesting findings pertaining to the dynamic microbial signatures (or characterisations). In their study Gerber et al. [2012] show that by analysing temporal microbial abundance data through a computational framework that uses continuous-time dynamic models and Bayesian dimensionality adaption methods, they are able to successfully characterise the microbial community's reaction to the use of antibiotics. In Yang et al. [2019] they show that dynamic signatures of the infant gut microbiome are able to capture information about the delivery and feeding modes of the infants. They use feature-based characterisations, where the features included microbial composition as well as bacterial richness, bacterial diversity etc.

Knights et al. [2011] suggests using machine learning methods to harness the value of microbial signatures for various clinical prediction tasks, including personalised medicine, treatment prognosis, forensic identification etc. They suggest the use of supervised learning for feature selection and signature discovery. Sanguinetti et al. [2019] and Zheng et al. [2020] are two interesting studies that explore dynamic microbial signatures in the identification and/or classification of illnesses. They respectively studied the relationship of microbial signatures to reduced memory and cognitive functions and the classification of unipolar versus bipolar depression using microbial signatures.

2.4.1.3 Community State Types and Representation of Heterogeneity

Ravel et al. [2011] presents a landmark paper in the characterisation of microbial communities, where they classify the vaginal microbiota into Community State Types (CSTs). They identify five different CSTs the vaginal microbiome exists in. They also identify that these CSTs change over time. This is important because this characterisation explicitly identifies the heterogeneous and time-variant nature of the microbial communities. Since then, these CSTs have been used in multiple studies pertaining to the human vaginal microbiome. [DiGiulio et al., 2015; Stewart et al., 2018b; Romero et al., 2014; De Seta et al., 2019; Ma and Li, 2017; Brooks et al., 2017; Mitra et al., 2015]

2.4.2 Characterisation of Temporal Dynamics in Other Systems

Not just microbial systems, but many other systems share the problem of characterisation of the temporal dynamics. While the characterisation of the microbiome has its unique challenges due to the nature of microbiological data, many studies have been done in the fields of physics and economics. Especially, the studies from physics relate more to biological studies as they are interested in multivariate characterisations and dealing with imperfect data (such as data with noise, measurement errors etc. [Bradley and Kantz, 2015; Zou et al., 2019]).

Teodorescu [2012] in their review show that characterisation of time-series data pertaining to nonlinear dynamic systems has been used in multiple fields of engineering. In the field of hydrology, Toth [2013] successfully uses the characterisation of time-series data coming from stream flow and precipitation to classify catchment sites. In the field of finances, Pecar [2004] uses time-series characterisation in a feature-based approach (using white noise and Wiener process plots) to successfully characterise eight real-life data sets from the New York stock exchange.

2.4.3 Utilities for Characterising Temporal Dynamics

In this section, we will explore the literature about non-linear time-series analysis and dimension reduction approaches suitable for characterising multivariate time-series in the low dimension.

2.4.3.1 Non-Linear Time-Series Analysis

Time-series analysis is inherently complex due to the natural temporal order. This distinctly places time-series analysis as a separate category within data analysis, as the directionality and order of data are to be preserved in the analysis process. Characterising time-series data hence has this additional challenge [Zou et al., 2019]. Early approaches in this characterisation involved linear stochastic models such as autoregressive (AR) and moving average (MA) and combinations such as autoregressive moving average (ARMA), autoregressive integrated moving average (ARIMA)models. However, for most natural processes, the set of linear stochastic models prove to be inadequate [Bradley and Kantz, 2015].

[Fulcher, 2017] summarises the features utilised in time-series characterisation argues that the feature selection is ultimately more useful than the categorisation of a time-series to be linear or non-linear. Some popular characterisations are in terms of features are value distribution, entropy, correlation properties, stationarity etc. which honour the directional nature of a time series. We can also find more nuanced features when we specifically consider non-linear models.

Time-series analysis starts with a hypothesis on the appropriate model for the data. As such, modelling natural processes involves many assumptions in the hypothesis on the appropriate model. Packard et al. [1980], and Takens [1981] in their landmark publications introduces the non-linear time-series analysis in the field of

fluid analysis. Non-linear time-series analysis improves our hypothesis by broadening our models and through the ability to extract non-linear features of naturally occurring time-series data. [Bradley and Kantz, 2015]

Another separation for time-series analysis is the univariate or multivariate nature of the data [Fulcher, 2017]. In this work, we are solely interested in multivariate data as microbial behavioural dynamics, and associated abundance profile data are always multivariate.

2.4.3.2 Dimension Reduction Approaches

As mentioned earlier, most time-series data related to natural processes are multivariate, as is the case with microbial abundance data. Hence, to visualise these in a low dimensional space, dimension reduction approaches are often taken. Often, temporal data are collected once and then visualised, but rarely there are instances where continuous visualisation is important as well. This is especially true with fast advancing technology, such as next-generation sequencing (NGS), resulting in a plethora of data. The approaches mentioned below are well suited for obtaining a visual characterisation of high-dimensional, complex, time-series data. They are trained on a neural network or related models that can capture complex non-linear features.

Parametric t-SNE Van Der Maaten [2009] introduces parametric t-SNE, which learns a parametric mapping between the high-dimensional space and the low-dimensional space while preserving the local structure of the data is preserved in the latent space. With a parametric mapping, more data points could be added continuously, allowing the visualisation of time-series multivariate data.

Parametric UMAP Sainburg et al. [2020] introduces a parametric version of the Uniform Manifold Approximation and Projection (UMAP) algorithm. The original UMAP algorithm is non-parametric and graph-based. Summarily it operates in two steps. First, it computes a graphical representation of the dataset. Then it learns a set of neural network weights that preserves the structure of the original graph. The neural network weights are now the parameters of the trained parametric-UMAP model. Original UMAP algorithm [McInnes et al., 2018] used to learn an embedding which preserved the structure of the graph via optimisation.

SONG Senanayake et al. [2019] introduces Self Growing Nebulous Growths(SONG), which is again a parametric, non-linear dimension reduction technique that presents a topology-preserving lower-dimensional visualisation. It is capable of handling both homogeneous and heterogeneous data increments into its mapping, making it more suitable for complex data such as microbial abundance data.

Chapter 3

IMPARO: Inferring Microbial Interactions through Parameter Optimisation

This chapter is partially composed of material which were published in a manuscript titled "IMPARO: Inferring Microbial Interactions through Parameter Optimisation" by Vidanaarachchi R., Shaw, M., Tang, S.L., Halgamuge, S.K. in BMC Molecular and Cell Biology [Vidanaarachchi et al., 2020]

Summary

Background Microbial Interaction Networks (MINs) provide important information for understanding bacterial communities. MINs can be inferred by examining microbial abundance profiles. Abundance profiles are often interpreted with the Lotka Volterra model in research. However existing research fails to consider a biologically meaningful underlying mathematical model for MINs or to address the possibility of multiple solutions.

Results In this chapter we present IMPARO, a method for inferring microbial interactions through parameter optimisation. We use biologically meaningful models for both the abundance profile, as well as the MIN. We show how multiple MINs could be inferred with similar reconstructed abundance profile accuracy, and argue that a unique solution is not always satisfactory. Using our method, we successfully inferred clear interactions in the gut microbiome which have been previously observed in *in vitro* experiments.

Conclusions IMPARO was used to successfully infer microbial interactions in human microbiome samples as well as in a varied set of simulated data. The work also highlights the importance of considering multiple solutions for MINs.

3.1 Background

Microbes are the most abundant, widespread organisms on Earth. They can be found in the biosphere, including all animals and plants, and most habitats in the oceans [Blaser et al., 2016; Apprill, 2017], on land, or in air. Many studies show that microbes play a important role in the health and well-being of the hosts they are associated with. For example, in the human body, imbalances or changes in microbial communities correlates to various illnesses and other complications [Clemente et al., 2012; Cho and Blaser, 2012; Khanna and Tosh, 2014; Singh et al., 2017; Hibberd, 2013; Funchain and Eng, 2013; Kumar and Chordia, 2017]. In plants, microbes provide essential nutrients, including all economic crops [Fitzpatrick et al., 2018; Finkel et al., 2017; Mueller and Sachs, 2015].

In the past, studying microbial communities through cultivation in laboratories was challenging [Hiergeist et al., 2015]. Also, as over 99% [Amann and Rosselló-Móra, 2016; Locey and Lennon, 2016] of microbial species on earth are yet to be identified, the inability to cultivate and separate some microbial species in a laboratory environment has hindered progress on the study of microbiota.

Due to recent advances in 16S rRNA sequencing and high throughput sequencing, scientists can now explore the nature of real-world microbial samples and recognise individual species in these samples. 16S ribosomal RNA has been used by many scientists in order to identify, categorise and classify microbes.

Microbial networks are inherently complex in nature. With longitudinal studies, for example, it has become clear that the composition of microbial communities is constantly changing. Now, in order to properly understand these communities, it is important to study how they are changing, why they are changing, and how they interact with each other. To do so, it is important to acknowledge the following dynamics which play a part in the microbial composition changes. There could be temporal changes that are caused by external factors such as temperature variations [Minich et al., 2018], diurnal cycles [Thaiss et al., 2014] or seasonal variations [Smits et al., 2017]. In addition to these, other non-random co-occurrence patterns have been observed. Like in any other community, organisms in microbial communities interact in various ways with each other. Some of these interactions could be categorised under mutualism, competition, parasitism, predation, commensalism and amensalism. [Faust and Raes, 2012]

Some important questions to ask about any biological community include, 'Who is there?', 'What are they doing?', and 'How will they respond?' [Boon et al., 2013]. While 16S ribosomal RNA sequencing can answer the first question, the latter two questions require an understanding of the interactions between different bacteria, hence the importance of inferring microbial interactions. These answers will improve our understanding of the human gut, the world's oceans, plant root systems, lakes, etc.

3.1.1 Related Work

With the advance of high throughput sequencing, high throughput inferring approaches have also been recently proposed. These are shown to be more successful than *in vitro* analysis of interaction patterns [Yokobayashi, 2019]. Some of these approaches are Metagenomic Microbial Interaction Simulator (MetaMIS) [Shaw et al., 2016], Rule-based Microbial Network (RMN) algorithm [Tsai et al., 2015], Sparse Inverse Covariance Estimation for Ecological Association Inference (SPIEC-EASI) [Kurtz et al., 2015], Learning Interactions from Microbial Time Series (LIMITS) [Fisher and Mehta, 2014], Boolean Abundance Analysis [Claussen et al., 2017], Boolean Dynamic Model [Steinway et al., 2015], Stochastic Generalised Lotka-Volterra and Extended Kalman Filter (SgLV-EKF) [Alshawaqfeh et al., 2017] and Sparse Correlations for Compositional Data (SparCC) [Friedman and Alm, 2012]. These algorithms mainly take two approaches [Shaw et al., 2016], correlation-based analysis and model centred analysis. Often algorithms combine the two approaches to come up with a more robust method of inferring microbial interactions.

MetaMIS [Shaw et al., 2016] uses a model-based approach where microbial interactions are assumed to abide by the biologically-inspired Lotka Volterra Model. The parameters of the Lotka Volterra model, which elucidate the interaction coefficients, are then approximated through a Partial Least Square Regression (PLSR). With these coefficients in place, the initial population is repopulated to recreate the community abundance profile. The accuracy metric is the Bray–Curtis Dissimilarity between the original and recreated abundance profiles. The authors do not use any simulated data in their results but report inferences from male and female gut microbial communities. Their reported accuracy is 78% to 82%.

RMN [Tsai et al., 2015] introduces its own model of Non-linear Regulatory OTU-triplet (NRO) model. This is a model for three OTUs which supposedly interact with each other. This assumption of interaction is then tested on the temporal abundance profile by a hyperbolic tangent based lack-of-fit function which they have introduced. The accuracy of the model is calculated based on correct inferences and correct non-inferences as a fraction of all inferences and non-inferences. Their reported accuracy is approximately 75% on simulated data. The authors use their method on infant gut data and infer previously known interactions.

SPIEC-EASI [Kurtz et al., 2015] is a correlation-based statistical method, which uses a Stability Approach to Regularisation Selection (STARS) to recreate the interaction correlations in form of a weighted undirected graph. Although this method does not indicate the nature of the interaction between two OTUs, it does give an idea of how close the OTUs are. The verification has been done through simulated data, and accuracy is measured with the Precision-Recall (P-R) curves and Area Under P-R Curves (AUPR). The authors have also presented the results from applying their method to the American Gut Project [McDonald et al., 2018] data.

LIMITS [Fisher and Mehta, 2014], yet another model-based algorithm, uses the discrete-time Lotka Volterra equations as the central microbial interaction model in its approach. The parameters of the Lotka-Volterra model are approximated through

linear regression with an iterative bootstrapping approach. The verification is done through simulated data where the authors report a specificity of 60%–80% and a sensitivity of 70%–80%. They also analyse two individuals' gut samples with the LIMITS algorithm. The major use of the LIMITS algorithm is to deduce keystone species.

Gao et al. [2018], in their work, use a model-based approach. They use a Lotka-Volterra model, fitted with abundance data using the non-linear least squares minimisation technique. Then they use a forward step-wise regression method with bootstrap aggregation to select candidate models. These models are then filtered through a Bayesian information criterion which results in multiple models being selected. They aggregate the models into a single network as the output. The algorithm is tested on a cheese microbial community. The authors also apply the method on the gut microbiome of children with Type 1 diabetes. They do not present accuracy numerically, but confirm that their method was successful in inferring experimentally confirmed microbial interactions.

Boolean Analysis [Claussen et al., 2017] uses an interesting model-based approach. The underlying biology is assumed to be forming either competitive links or synergistic links. Pairs of abundance vectors are analysed with the ESABO (Entropy Shifts on abundance vectors under Boolean operators) to confirm either a competitive or a synergistic link. Using a Jaccard index of the difference between the normalised number of correctly and incorrectly classified links, with their simulated data, they have achieved indexes ranging from 0.1–0.6 on competitive links and 0.1–0.9 on synergistic links. Their approach is also applied to a Human gut data-set.

Boolean Dynamic Model [Steinway et al., 2015] does not contain an embedded biological model but assumes a binary relationship among OTUs. First, this method binarises the abundance data with a k-means binarisation, which allows binarisation with a threshold value, but with a stochastic element. Then it uses a recapitulating approach of updating and maintaining binary rules. The last part is a perturbation analysis, where it analyses the effects of removal (knock-out) or addition (forced over-abundance) on the created model. This method is effective for the work's purpose of analysing *Clostridium difficile* infection in the gut. The finding is that *Barnesiella intestinihominis* hinders the growth of *Clostridium difficile*. This has been confirmed in *in vitro* experiments.

SgLV-EKF [Alshawaqfeh et al., 2017] model is a straightforward approach of using the Lotka Volterra equations as the underlying biological model. But it improves the generalised Lotka Volterra (LV) system by introducing a Gaussian noise term, making it stochastic. Then the LV parameters are estimated using an Extended Kalman Filter (EKF), giving it the name SgLV-EKF. This algorithm is tested on Monte-Carlo simulated data, and shows an accuracy of 75%, with Mean Square Error (MSE) being the indicator of accuracy. The authors also apply the method on two mouse gut systems infected by *Clostridium difficile*, one being treated with clindamycin.

SparCC [Friedman and Alm, 2012] is a co-occurrence based method which iteratively finds non-random co-occurrence patterns in microbial data. One of the first methods proposed in inferring microbial interactions, SparCC has shown a consid-

erable improvement from the Pearson Correlation method. On simulated data, it has shown to achieve root mean squared errors (RMSE) as low as 0.02. The authors also apply the method on Human Microbiome Project data to show its usability on real life data.

Barberán et al. [2012] presented an early study in which they used a Checkerboard-score along with Spearman's correlation coefficient to uncover non-random co-occurrence patterns in soil microbiome data. They stop short on identifying patterns of community co-existence but raise the need for more focused experiments to study environmental and community shifts over time.

Considering the literature, there seems to be a shift towards using model-based systems, with the support of statistical methods, rather than depending purely on statistical methods. An explanation of this is that, due to the complex nature of the microbial communities, purely mathematical methods, which ignore the underlying biology, would be prone to overlooking important biological constraints. Microbial communities have biologically specific behavioural dynamics, which cause non-independence between adjacent time-steps. Hence models which take into account these behavioural dynamics are useful in inferring the interactions.

On examining existing model-based work, it is notable that Lotka Volterra Equations or one of its adaptations has been used in many approaches as the underlying biological model. The major reason for this use is that it has been shown that Lotka-Volterra Model can successfully simulate a microbial community when applied to different scenarios such as Lake Ecosystems [Dam et al., 2016], Human and murine intestinal microbial systems [Stein et al., 2013; Marino et al., 2014] or the microbial ecosystem which occurs in the process of ripening of smear cheese [Mounier et al., 2008]. The generalised Lotka Volterra equations have the capacity to capture the growth rates and the pairwise interactions of the OTUs, which are the important coefficients estimated in the process of inferring Microbial Interaction Networks (MINs).

Many of these studies have applied a new methodology to simulated data as well as real-life data. This is important because data simulations always assume a known biological model, and the inherent noise in a biological system is not always present in artificially simulated data. Our work and the majority of other works are also guilty of using the same biological model in the inference algorithms, as well as in the data simulations. Hence some sort of verification with real-life data is obviously important. The problem with using real-life data for verification is that sans *in vitro* studies, it is difficult to discern whether the inferred interactions are in fact *bona fide* interactions found in that microbial system. One potentially useful verification strategy is to highlight the overlap between identified interactions and interactions that were previously known. MetaMIS [Shaw et al., 2016] uses an abundance profile reconstruction strategy to confirm their results. This system of verification influenced our method.

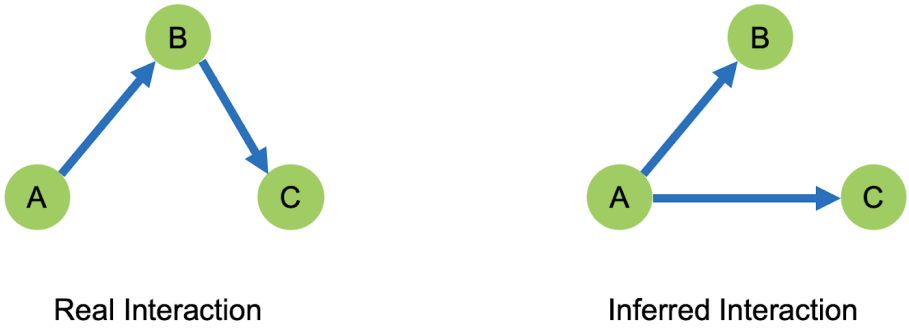

Figure 3.1: Although the real interactions are $A \rightarrow B$ and $B \rightarrow C$, through A's influence on B, A has an indirect influence on C. When these interactions are inferred through an abundance profile, the indirect interaction $A \rightarrow C$ may be inferred instead.

3.1.2 Motivation & Contributions

It was interesting to note that the above mentioned methods imply a unique solution to the problem of inferring a microbial interaction network, given a particular abundance profile. In their work addressing pitfalls in inferring microbial dynamics, however, [Cao et al., 2017] demonstrate that multiple interaction networks can lead to the same abundance profile. This is supported by the simple scenario of three OTUs with indirect interactions, as shown in Figure 3.1.

In this chapter, we present IMPARO (Inferring Microbial interactions through PARameter Optimisation), an algorithm for microbial interaction inference which incorporates biologically meaningful models for the interaction network as well as the abundance profile.

IMPARO is the first inference method to not make the assumption of a unique inferred solution, and to explore multiple solutions with similar accuracy levels. Because of the inherent noise in microbial abundance data, it is reasonably assumed that small changes in accuracy do not necessarily mean superior MINs.

It is also the first to assume an underlying biological model for a microbial interaction network (MIN), by using the microbial community dynamics model introduced in [Gibson et al., 2016]. The shift from statistical methods to model-based methods was inspired by using an underlying biological model for the Abundance Profile, and models such as gLV, SgLV, NRO and entropy shift of competitive synergistic links were used. Our work goes a step further in introducing an underlying biological model for the MIN, which reduces the optimiser search space by pruning solutions which are less feasible biologically.

It also contains a Monte Carlo approach [Metropolis and Ulam, 1949] for the purpose of encompassing the effect of rarer OTUs into the inferred MIN. Most statistical

methods fail to do justice to the effects of rarer OTUs simply because their presence is overwhelmingly shadowed by the other OTUs. And most model-based solutions use filtering processes which favour higher-ranked (in terms of abundance) OTUs before the inference process. But in fact, the majority of OTUs in a community are rarer OTUs. [Shaw et al., 2016; Caporaso et al., 2011]

Our results are verified through both simulated and real-life data. Our simulations take into account the diversity of microbial communities. Community dynamics models are used to ensure different types of communities are included in our testing. We compare the results from IMPARO with results reported in the literature.

Key Contributions Summarised:

- Inference of interactions without the assumption of a unique solution.
- Consideration of an underlying biological model for the MIN.
- Using a Monte Carlo approach to ensure a better representation of rarer OTUs.
- Verification of the algorithm on real life and simulated data.
- Comparison of results with that of existing methods.

3.2 Results

IMPARO was used to infer interaction parameters in both simulated and real life data. We present the overall results in this section. [1]

3.2.1 Simulated Data

Data simulation was performed using the microbial community dynamics model described above, and focuses on heterogeneity and sparsity variation. Nominal component $\mathbf{N}$ is sampled from a normal distribution $\mathcal{N}(0,1)$. Initial abundance values were sampled randomly from a uniform distribution $\mathcal{U}(0,1)$, as suggested in Gibson et al. [2016]. In this study, we are interested in examining how IMPARO handles data-sets with varying heterogeneity and sparsity. For the purpose of the simulated study, we used ten species.

For the heterogeneity study, we use $P(\alpha)$ s.t. $\alpha \in [0.2, 0.4, 0.6, 0.8, 1.0]$, so that communities with a heterogeneity favouring a minority of highly influential OTUs are considered.

For the sparsity study we use $G(n,p)$ s.t. $p \in [0.2, 0.4, 0.6, 0.8, 1.0]$. This would include communities which are very sparse ($p = 0.2$) to fully connected ($p = 1.0$).

The Mean Squared Error (MSE) between the ground truth and the inferred parameters in each case as described above are shown in Table 3.1. We observe that lower p values and higher α values—highly sparse and highly heterogeneous instances—result in lower errors.

[1] Additional results and snapshots of simulated data are available in Additional file 3 of Vidanaarachchi et al. [2020].

Table 3.1: MSE values from the heterogeneity and sparsity study. Heterogeneity and sparsity were varied—through varying α and p respectively—to investigate how IMPARO responded to microbial samples of varying nature. Mean Squared Error (MSE) indicates how far the inference is from the ground truth.

$\sigma = 1$	**P**				
	$p = 0.2$	$p = 0.4$	$p = 0.6$	$p = 0.8$	$p = 1.0$
$\alpha = 0.2$	0.05	1.32	1.36	2.55	1.99
$\alpha = 0.4$	0.61	0.63	1.36	0.66	1.02
$\alpha = 0.6$	0.42	0.57	1.54	1.98	1.81
$\alpha = 0.8$	0.09	0.57	1.14	0.79	1.51
$\alpha = 1.0$	0.34	0.28	0.71	0.73	1.28

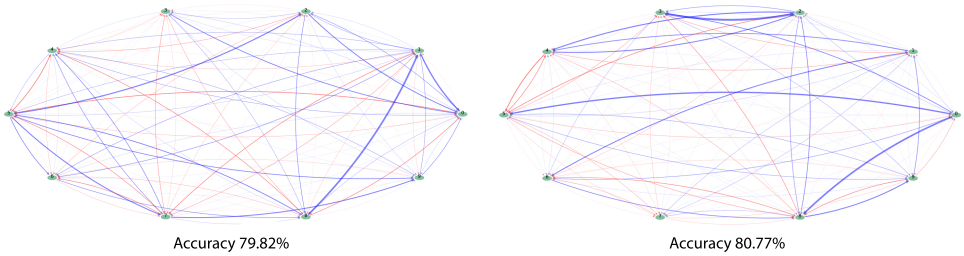

Figure 3.2: An example of two distinct solutions for the same simulated data-set. The MINs corresponding to each solution, when evaluated with reconstructed abundance profile accuracy were within 1% of each other.

Tested for robustness with Gaussian noise ($\mu = 0.0$, $\sigma = 0.01$), IMPARO returns solution clusters which are within mean squared errors of 0.4 - 0.5 of each other, suggesting the solutions are robust.

3.2.2 Existence of Multiple Solutions

As we have mentioned in the literature review, it is possible to find multiple solutions for the problem of inferring microbial interactions when the accuracy is measured through reconstructed abundance profiles.

In Figure 3.2 we present two MINs inferred from the same abundance profile, which—after recreating the abundance profile and measuring for accuracy using Bray–Curtis metric—returns accuracies within 0.1% (79.82% and 80.77% respectively). Compared to the true values used in simulating the data, they indicate mean squared errors of 0.59 and 0.58 respectively.

Table 3.2: The results for the female faecal microbiome sample showing reconstructed abundance profile accuracy values for varying numbers of highest-ranking OTUs.

No of Highest-Ranking OTUs	Reconstructed Abundance Profile Accuracy
5	85.42%
10	84.22%
20	82.77%
30	79.93%
40	81.86%
50	82.08%
60	74.83%
69	80.11%

3.2.3 Tests on Real Life Data

For this study we use the data from human faecal microbiome samples collected from a healthy male and a female for time spans of 15 months and 6 months respectively [Caporaso et al., 2011]. This data is publicly available at MG-RAST:4457768.3-4459735.3.

On female faecal microbiome analysing the 10 highest-ranking OTUs, our method achieves a 84.22% reconstructed abundance profile accuracy. On the male faecal microbiome OTU rankings, our method achieves a 81.60% accuracy. It should be noted that in the female sample, 185 time points were taken into account. In the male sample 442 time points were considered. In both instances, the sparsity of the connections were assumed to be 50% for the inference process.

The results for the female faecal microbiome sample showing reconstructed abundance profile accuracy values for varying numbers of highest-ranking OTUs are tabulated in Table 3.2.

As a further analysis, we inferred MINs at different taxonomic resolution levels—from Phylum to Genus. The reconstructed abundance profile values of this study performed on the female faecal microbiome are tabulated in Table 3.3. The ten highest-ranking OTUs were considered in this study.

3.2.4 Inference of Rarer OTU Interactions

In order to understand how our method works for rarer OTUs, we processed randomly selected samples from the female faecal microbiome with at least 50% of the considered OTUs from the rare range (average abundance lower than 0.1%). In some studies, [Tsai et al., 2015; Shaw et al., 2016] these rare OTUs are discarded while favouring the most abundant OTUs. But we show that rarer OTUs can indeed be considered in the inference process, and give satisfactory results. Our samples provided an average accuracy (reconstructed abundance profile accuracy) in the order

Table 3.3: Inspecting the reconstructed abundance profile accuracy with varying taxonomic resolution levels in the female faecal microbiome.

Taxonomic Resolution Level	Reconstructed Abundance Profile Accuracy
Genus	76.30%
Family	84.22%
Order	87.22%
Class	87.54%
Phylum	87.63%

of 60%.

3.3 Discussion

In this section, we analyse the results obtained by IMPARO.

3.3.1 Simulated Data

The simulated study indicates that IMPARO works better with data samples with low heterogeneity and high sparsity (low p-value). When considering highly heterogeneous samples, we attribute the larger errors to the difficulty in inferring near-zero values. For less sparse data-sets this can be attributed to the difficulty in inferring a fully connected MIN. The best case as seen in Table 3.1 being the most heterogeneous and sparsest instance can be attributed to it being close to the trivial case of all zeros. It is indeed expected to have better results in the more sparse samples, as Genetic Algorithms (GAs) tend to converge faster when the dimensions of the parameter space are lower. Achieving better results on low heterogeneous and moderately sparse samples in the simulated data explains the better results obtained in real-life samples with the higher ranking OTUs, which are more homogeneous and are assumed to be moderately connected.

3.3.2 Existence of Multiple Solutions

Although the reconstructed abundance profile accuracy is indicative of the prediction accuracy of the interaction parameters, there seem to be multiple distinct solutions for interaction matrices resulting in similar abundance profile accuracies. Also to be noted is that these distinct solutions are within 1–2% of reconstructed abundance profile accuracy. Because of the high noise in microbial data-sets, a solution which is only 1–2% better in recreated abundance profile accuracy cannot be considered to be a superior solution. A possible cause for multiple solutions could be the optimiser being stuck at local optima. However as the parameter space has too many dimensions to permit visualisation, the methods need to rely on results obtained

from multiple initialisations. While recognising GA is particularly challenged with overcoming local optima, it is worth looking into other explanations possible. One cause for multiple distinct solutions is the possibility that indirect interactions are being inferred incorrectly through these methods.

We may conclude that good reconstructed abundance profile accuracy is a necessary condition for a precise prediction although it is not a sufficient condition by itself. Hence we highlight the need to widen the search for all such instances where the reconstructed abundance profile accuracy is higher than a threshold value. An optimisation approach which provides multiple answers is, therefore, important.

3.3.3 Tests on Real Life Data

First, we note that the inference of the male faecal microbiome resulted in a lower accuracy compared to the female faecal microbiome. This might be due to the fact that the male sample covers a greater time period than the female sample. (442 time points over 15 months in comparison to 185 time points over 6 months).

Apart from the increased difficulty in predicting a longer time series, it can also be hypothesised that the inherent changes in the microbiome itself over a longer period of time could be a reason for the reduced predictive accuracy. Microbes, as any other community of living organisms, change over time, which includes changes in the nature of their interactions.

In Table 3.2 we observe a trend towards the accuracy decreasing as the number of OTUs included is increased. The reasons for this could be two-fold. Firstly, as the number of OTUs increases, the number of parameters to be estimated grows quadratically. Secondly, as more lower-ranked—and rarer—OTUs are considered, the difficulty level of inference increases.

We observe that higher accuracy levels correspond to higher taxonomic ranks in Table 3.3. Considering that the number of OTUs remained constant in this study, we conjecture that as abundances get more numerous for each OTU with each higher taxonomy level, abundance profiles become less disorderly. This could have resulted in better reconstructed abundance profile accuracies for higher taxonomic resolution levels.

Of mutualism interactions inferred by our algorithm, some have been shown to exist in previous studies as shown in Figure 3.3. The population of bacterial families of *Prevotellaceae* and *Rikenellaceae* has shown to increase simultaneously in immune impaired Nod2(-/-) mice faecal microbiome [Hasegawa and Inohara, 2014]. The populations of *Rikenellaceae* and *Verrucomicrobiaceae* have been shown to simultaneously increase in another study of mice faecal bacteria studying diet induced obesity [Clarke et al., 2013]. Both these results were inferred from the female faecal microbiome sample.

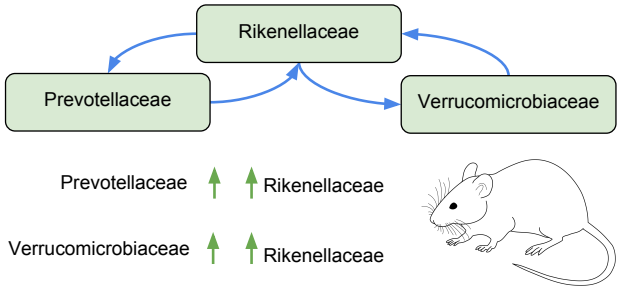

Figure 3.3: Strong microbial interactions inferred from the female faecal microbiome have been previously observed in *in vitro* studies of murine microbiome.

3.3.4 Consideration of Rarer OTUs

From the results, it could be seen that when the rarer OTUs are taken into account, the predictive power is significantly less. Even though the predictive power is less, the approximately 60% reconstructed abundance profile accuracy suggests promise in exploring the question of inferring interactions for rarer OTUs further. Also, when combined with higher ranking OTUs, rarer OTUs do not significantly reduce the accuracy of the whole sample, as indicative from the results in Table 3.2.

Analysis of Errors

We consider the reasons for the 20% error margin of IMPARO to be threefold. Firstly, microbial interactions are prone to change over time. When interactions are inferred over multiple points covering a large time interval, this could add a significant error. Secondly, the high dimensionality of the search space increases the chance of local optima, thus resulting in higher errors. Thirdly, as the input data is acquired through experimental means, we expect the errors from the experimental procedures and data collection to have contributed to the overall error.

To understand the implications of the errors, we have to look at the error calculation. We depend on the derived error metric RAPA as it is impossible to know the ground truth for MIN parameters (Section 3.5.4). As the abundance profile is reconstructed recursively with the inferred interactions, the error present at the RAPA level reflects a propagated error from the errors at the MIN level. Hence, the error margin of 20% is acceptable and expected. Due to the nature of the calculation, the error is representative of all the MIN parameters proportionately.

Future Work

There are several possible ways of extending IMPARO, to alleviate some of its weaknesses. IMPARO attempts to infer a single interaction parameter for each OTU couple for the entire time-line. We note that, as microbial interactions are prone to change over time, it can be beneficial to infer interactions over separate time intervals, which

could allow better abundance profile recreation and exploration of interaction parameter dynamics over time. Also, IMPARO currently lags at inferring rarer OTUs, as compared to higher ranking OTUs. Supplementing genomic data with transcriptomic data in the inference process can potentially increase the prediction quality. It is also worth exploring how IMPARO can be improved to deter the disruption of the community dynamics model by zero and non-zero values.

3.4 Conclusions

Inferring microbial interactions will advance our understanding of microbial communities. We have presented IMPARO, a microbial interaction inference algorithm based on parameter optimisation. We have conducted studies on simulated microbial communities and on real-life data. IMPARO has shown to successfully infer interaction parameters corresponding to microbial systems in the human body. We also emphasise the importance of considering multiple solutions for the MINs.

3.5 Methods

In this section, we present the methods used in IMPARO.

3.5.1 Generalised Lotka Volterra Model

The Generalised Lotka-Volterra Model (GLV) is a system of Ordinary Differential Equations. In inferring interactions the GLV is used in its discrete form, where each time point represents a sample in the temporal abundance profile. The differential equations describe the difference of a single OTU's abundance levels in two adjacent time points, and how it is dependant on the growth rate and its interaction coefficients with the other OTUs.

$$\frac{d}{dt}x_i(t_k) = r_i x_i(t_k) + x_i(t_k)\sum_{j=1}^{L} A_{ij}x_j(t_k) \tag{3.1}$$

In Equation 3.1 $x_i(t_k)$ describes the relative abundance of the i^{th} OTU at time t_k. The growth rate of the i^{th} OTU is described by r_i. $\mathbf{A}$ is the overall interspecific interaction matrix, where $\mathbf{A}_{ij}$ describes the effect on the j^{th} OTU by the i^{th} OTU. ($\mathbf{A}_{ij} < 0$ represents a negative effect on the j^{th} OTU by the i^{th} OTU). The saturation terms have not been included as we do not consider communities to have known carrying capacities (different types of LV equations are outlined in Chapter 4).

We use the above framework as it is in our implementation and add a noise term afterwards to compensate for inherent and experimental noise in microbial data. All the abundance values are normalised for each time point.

3.5.2 Community Dynamics Model

Introduced by Gibson et al. [2016], the community dynamics model is best described as a Mathematical Model consisting of a set of Matrices which represent different qualities in microbial interactions.

$$A = NH \circ Gs \tag{3.2}$$

In Equation 3.2 **A** is the microbial interaction matrix, **N** is the nominal interspecific interaction matrix, **H** is the heterogeneity matrix and **G** is the adjacency matrix of the underlying ecological network. s is a scaling coefficient. The operator $\circ$ represents the Hadamard product (element-wise multiplication of matrices).

$\mathbf{N} \in \mathbb{R}^{n \times n}$, the nominal interspecific interaction matrix has a normal distribution with a mean of 0, and a standard deviation of σ^2, i.e. $\mathbf{N}_{ij} \sim \mathcal{N}(0, \sigma^2)$. This matrix warrants that the interactions are fair in the absence of an influencing factor, which is introduced in the next component. $\mathbf{H} \in \mathbb{R}^{n \times n}$, the heterogeneity matrix is a diagonal matrix with a power-law distribution, with an exponent of α, i.e. $\mathbf{H}_{ii} \sim \mathcal{P}(\alpha)$. This matrix simulates the difference in the interspecific influence levels. It is believed that in a typical community there are a small number of highly influential species [Dawson et al., 2017]. Together with the interspecific interaction matrix, the heterogeneity matrix assures a balanced community dynamics model. The next step is defining the connectedness, as MINs are generally not fully connected but sparse. $\mathbf{G} \in \mathbb{R}^{n \times n}$ is a binary matrix where $\mathbf{G}_{ij} = 1$ represents that the OTU i is affected by OTU j and $\mathbf{G}_{ij} = 0$ represents otherwise. This matrix follows an Erdős–Rényi model with $G(n, p)$ where n is the number of OTUs and p is the probability of an edge which also represents the sparsity of **G**. [2]

3.5.3 Bray–Curtis Dissimilarity

Bray–Curtis dissimilarity [Bray and Curtis, 1957] is used in our work to determine the dissimilarity between two samples, specifically, the dissimilarity between corresponding time-points in original and recreated abundance profiles. However, a limitation of using the Bray–Curtis Dissimilarity is that the dissimilarity metric is biased towards more abundant species.

$$BCD(\mathbf{x}_{(t_k)}, \mathbf{x}^*_{(t_k)}) = \frac{\sum_{i=1}^{L} \left| x_{i_{(t_k)}} - x^*_{i_{(t_k)}} \right|}{\sum_{i=1}^{L} \left(x_{i_{(t_k)}} + x^*_{i_{(t_k)}} \right)} \tag{3.3}$$

$$BCD_{overall} = \frac{\sum_{k=0}^{T} BCD(\mathbf{x}_{(t_k)}, \mathbf{x}^*_{(t_k)})}{T} \tag{3.4}$$

where $\mathbf{x}_{(t_k)}$ and $\mathbf{x}^*_{(t_k)}$ represent relative abundances of the original and recreated abundance profile, at time k. $x_{i_{(t_k)}}$ represents the relative abundance of the i^{th} OTU

[2] An illustrated numerical example is given in the Additional file 1 of Vidanaarachchi et al. [2020].

of the original abundance profile at time point k and $x^*_{i_{(t_k)}}$ represents the same in the recreated abundance profile. L is the number of OTUs in the sample, while T is the total number of time-points in the abundance profile.

3.5.4 Reconstructed Abundance Profile Accuracy

The reconstructed abundance profile accuracy is a metric of how accurately the original abundance profile can be reconstructed with the inferred MIN. Using the original initial conditions, $\mathbf{x}_{(t_0)}$, the subsequent microbial community compositions are calculated using the generalised Lotka-Volterra model. This reconstructed microbial community abundance profile is then compared to the original abundance profile using the Bray–Curtis Dissimilarity. This metric reflects the quality of the inferred MIN.

3.5.5 Kolmogorov–Smirnov Test

We use the Kolmogorov–Smirnov Test as a goodness-of-fit test to compare the empirical distribution of the inferred MIN to a model empirical distribution which follows the Community Dynamics Model.

$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)| \tag{3.5}$$

where $F_{1,n}(x)$ and $F_{2,m}(x)$ are the empirical distribution functions for the parameters of the microbial interaction networks. Here parameters of the interaction networks are considered as one-dimensional probability distributions. (i.e. each interaction is considered to be independent). sup is the supremum function [Dodge, 2008].

3.5.6 Inferring MINs from Abundance Profile

We are viewing the inference of MINs as an optimisation problem. As our aim is to estimate the elements of the matrix $\mathbf{A}$, the overall interspecific interaction matrix, this can specifically be described as a large parameter optimisation problem, because the parameters we are estimating is in the order of N^2, where N is the number of OTUs taken into consideration. The interaction coefficients of the bacteria community are considered to be the parameters. In the simplest case, the value we are optimising is the averaged Bray–Curtis Dissimilarity over the time axis, for the original abundance profile and the recreated abundance profile from generated with the parameters. We later take the statistical similarity of the parameter set (interaction coefficients) to the theoretical distribution of interaction coefficients according to the microbial community model.

MINs are estimated to be sparse in nature [Chen et al., 2017]. This information can be used to our advantage in optimising the parameters because the adjacency matrix of a sparse MIN contains many zero values. But what we do not know is

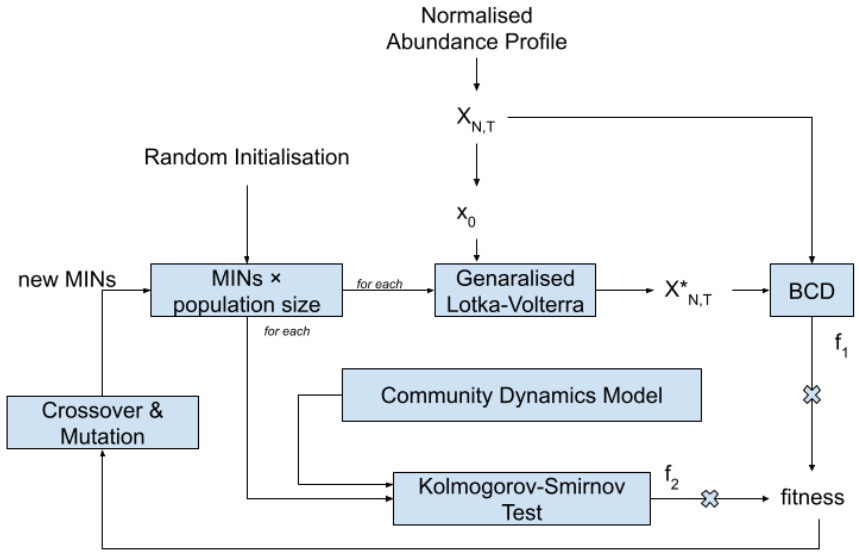

Figure 3.4: The process of IMPARO includes a Genetic Algorithm, which takes into account the Bray–Curtis Dissimilarity (BCD) and the Kolmogorov–Smirnov Test to calculate the fitness of a solution. The combined score ensures that the Microbial Interaction Networks (MINs) provided by the algorithm are feasible solutions. $X_{N,T}$ is the microbial abundance profile, with N OTUs, and T time points. X_0 is the microbial abundances at the initial time point. $X^*_{N,T}$, is the recreated abundance profile. f_1 and f_2 respectively are the factors BCD and KS Test scores counting towards the overall score.

which parameters should be set to zero, and which parameters should be set a non-zero value. Here we use a GA [Sastry et al., 2005; Holland, 1992] based approach whose Monte-Carlo simulation of Adjacency Matrices for MINs allow an estimated percentage of values to be set to zero, and to reevaluate that based on the BCD, which we are trying to minimise.

For the purpose of the GA, we consider each element in the matrix $\mathbf{A}$ to be a gene, and a collection of elements to be a chromosome. Because we are expecting sparse MINs, the chromosomes do not contain N^2 number of genes. This reduces the computational complexity. The algorithm makes mutations to the genes, which affect both row (i), column (j), and numeric effect ($\mathbf{A}_{ij}$). The crossover operation is a single-point crossover, where a randomly selected part of a single chromosome is replaced by the corresponding part of another chromosome.

The algorithm uses a two-fold fitness function where a score is assigned to each chromosome based on the BCD and a penalty is assigned based on the likelihood of being compatible with the community dynamics model. Thus, our algorithm considers underlying biological compatibility for both the abundance profile - in terms of OTU propagation through the generalised Lotka Volterra Equations, and the Adjacency Matrix for MIN with the community dynamics model.

The first part of the score is straightforward, with the BCD. For the penalisation step, it is important to explore the probability distributions of the community dynamics model. The matrix $\mathbf{A}$'s near zero values are identified and zeroed at first, to satisfy sparseness. The generated matrix is checked for compliance with expected statistical properties using the Kolmogorov–Smirnov (KS) test, and penalties are applied according to the KS statistic [Dodge, 2008]. Thus a combination score makes sure that future generations of solutions are compatible with the underlying biological models in terms of MIN and abundance profile. This process is illustrated in Figure 3.4. [3]

The GA approach in IMPARO which uses Monte Carlo methods for gene introduction allows rarer OTUs a better representation in the solution.

[3] Important code segments are provided in Additional file 2 of Vidanaarachchi et al. [2020].

Chapter 4

Exploratory Study of Temporally Dynamic Microbial Interaction Networks

"I am not now
That which I have been."

—Lord Byron,
Childe Harold's Pilgrimage

This chapter is partially composed of material which were published in a manuscript titled "Exploring Computational Inference of Microbial Interactions and their Dynamics" by Vidanaarachchi R., Shaw, M., Halgamuge, S.K. in proceedings of the 14th IEEE International Conference on Industrial and Information Systems 2019. [Vidanaarachchi et al., 2019]

Summary

Background Next Generation Sequencing has increased detailed studies of microbial communities. Understanding microbial interactions is crucial for understanding these communities. Microbial Interaction Networks (MINs) have been inferred using microbial abundance profiles, using multiple methods and frameworks.

Results In this chapter we explore current approaches in inferring microbial interactions and demonstrate that they are dynamic in nature, and propose a visualisation approach for observing dynamic nature of MINs. Our work highlights that even very rare microbiota have bursts of high abundance at some time points. Further, we suggest possible improvements in the field of computational inference of MINs.

Conclusion In conclusion, this work highlights the importance of considering microbial interactions to be dynamic. It also emphasises the need for furthering the study of microbial communities past current assumptions and limitations.

4.1 Background

As we discussed in Chapter 3, microbial communities are abundant in nature. They are the most numerous of living organisms. As such, they interact and affect many other systems [Apprill, 2017]. They can be found in water [Blaser et al., 2016], land [Fenchel et al.] and air [Bryan et al., 2019], and in humans, animals and plants. The presence or the absence of microbial organisms affect the health of humans and animals [Clemente et al., 2012; Cho and Blaser, 2012; Khanna and Tosh, 2014; Singh et al., 2017; Hibberd, 2013; Funchain and Eng, 2013; Kumar and Chordia, 2017]. In plants, the microbial communities provide essential nutrients and directly affect the yield [Fitzpatrick et al., 2018]. In aerosphere, microbes are responsible for airborne diseases [Bryan et al., 2019; Finkel et al., 2017; Mueller and Sachs, 2015]. In oceans, microbial disturbances have been linked to extreme weather and climate change [Cavicchioli et al., 2019] and also play an indispensable role in protecting coral reefs [Yang et al., 2017]. Related to climate change, it has been shown how thermal stress on corals are reflected on the coral microbiome [Shiu et al., 2017; Lee et al., 2015]. Indeed, solutions for multiple biological problems will have a connection to understanding the microbiome of ecosystems.

In Chapter 1 we discussed the three questions proposed by Boon et al. [2013], to further our understanding of the effect of microbiota in the ecosystems mentioned above. The research on microbial organisms traditionally relied on *in vitro* studies [Hiergeist et al., 2015] but had the limitation of the inability to cultivate most microbial species in laboratory conditions [Amann and Rosselló-Móra, 2016; Locey and Lennon, 2016]. However, with the advent of Next Generation Sequencing (NGS), the first question was answered by analysing the 16S rRNA to identify the different species in a microbial sampling [Janda and Abbott, 2007]. With high throughput 16S rRNA sequencing approaches that allowed the faster processing of time-series microbial samples, more studies can be done to understand the dynamic aspects of microbial communities. This allows us to answer the second question by observing dynamic compositional changes in the microbiome.

External and internal influences affect the composition of microbial communities. External influences can include variations in the temperature [Minich et al., 2018], diurnal cycles [Thaiss et al., 2014], and seasonal changes [Smits et al., 2017]. Despite these external factors, non-random co-occurrence patterns have been identified in microbial samples, which are attributed to internal influences. In addition, both trophic and non-trophic relationships have been observed in microbial communities [Wang et al., 2019; Tiede et al., 2016]. Faust et al. [2015] have categorised microbial interactions under mutualism, competition, parasitism, predation, commensalism and amensalism. Our interest lies in understanding these interactions of microbial communities.

It has been recorded that the external factors mentioned above affect the nature of trophic and non-trophic activities, in both macro- and microbial life. [Kim and Or, 2017; Lovett et al., 2009; Ciechanowski et al., 2007]. However, in the study of these microbial dynamics, especially where the Microbial Interaction Networks (MINs) are

quantified, the interactions themselves are considered to be static.

In this exploratory study, we provide an overview of existing literature and provide a clear explanation on different Lotka–Volterra models. Then we present evidence for the dynamic nature of microbial interactions. We also propose a novel way of visualising dynamic MINs. Then we provide an extensive discussion on future research directions with dynamic interactions. Lastly, we explain the methods utilised in this chapter. Apart from the exploration, this chapter has tutorial value in studying the processes of microbial interaction inference.

4.1.1 Related Work

In the following sub-sections we discuss Lotka–Volterra [Wangersky, 1978] equations, and their derivations in detail. We hope to clear some confusions regarding the various uses of the Lotka–Volterra equations present in the literature.

4.1.1.1 Lotka–Volterra Equations

On examining existing model-based work, it is notable that Lotka–Volterra Equations or one of its adaptations has been used in many approaches as the underlying biological model. The major reason for the use of the Lotka–Volterra Model is that it can successfully simulate a microbial community when applied to different scenarios such as Lake Ecosystems [Dam et al., 2016], human and murine intestinal microbial systems [Stein et al., 2013; Marino et al., 2014] or the microbial ecosystem which occurs in the smear cheese ripening process [Mounier et al., 2008].

Predator–Prey Equations The predator–prey equations are two first-order nonlinear differential equations which describe the dynamics of a system with two interacting species [Wangersky, 1978]. They form the basis for Lotka–Volterra equations.

$$\frac{dx}{dt} = \alpha x - \beta xy \tag{4.1}$$

$$\frac{dy}{dt} = \delta xy - \gamma y \tag{4.2}$$

In Equations 4.1 and 4.2, x is the number of prey, y is the number of predators. $\frac{dx}{dt}$ and $\frac{dy}{dt}$ represent the instantaneous growth rates of the two populations. α, β, δ, γ are positive real parameters describing the interaction between the two species.

Generalised Lotka–Volterra Equations The generalised Lotka–Volterra equations have the capacity to capture the growth rates and the pairwise interactions of multiple Operational Taxonomic Units (OTUs), which are the important coefficients estimated in the process of inferring MINs.

$$\frac{d}{dt}x_i(t_k) = r_i x_i(t_k) + x_i(t_k) \sum_{j=1}^{L} \mathbf{A}_{ij} x_j(t_k) \tag{4.3}$$

In Equation 4.3 $x_i(t_k)$ describes the relative abundance of the i^{th} OTU at time t_k. The growth rate of the i^{th} OTU is described by r_i, which is the generalisation of the parameters α and γ in Equations 4.1 and 4.2. Generalisation of β and δ is $\mathbf{A}$, which is called the overall interspecific interaction matrix, where $\mathbf{A}_{ij}$ describes the effect on the j^{th} OTU by the i^{th} OTU. ($\mathbf{A}_{ij} < 0$ represents a adverse effect on the j^{th} OTU by the i^{th} OTU).

Competitive Lotka–Volterra Equations and the Saturation Term Some methods use a version of the Lotka–Volterra equation with a saturation term [Gao et al., 2018]. This version is sometimes called the competitive form of the Lotka–Volterra equations for species competing for a common resource and with known carrying capacities [Wangersky, 1978]. The saturation term acts as a damping coefficient.

$$\frac{d}{dt}x_i(t_k) = r_i x_i(t_k)\left(1 - \frac{\sum_{j=1}^{L}\mathbf{A}_{ij}x_j(t_k)}{K_i}\right) \tag{4.4}$$

However, we also identify certain issues with the use of Lotka–Volterra equations in modelling microbial interactions, some of which are already discussed in literature [Momeni et al., 2017; Gilpin and Ayala, 1973]. Firstly, with static interactions, pairwise modelling is not sufficient to capture the intricacies of microbial interactions. Secondly, even if we are to consider pairwise abundance variations, as the Lotka–Volterra equations are a set of first order linear differential equations, any dynamic interactions are not captured from these equations.

4.1.1.2 Temporally Dynamic Interactions in Other Systems

Microbial ecological models can be compared and contrasted with many other natural and man-made systems. Literature pertaining to other systems have indications of time-varying interactions. Song et al. [2009a] discuss time-varying interactions between genes in gene regulatory networks. In their approach these dynamic interactions were inferred through a kernel-reweighted logistic regression approach based on time-series gene expression values. Tan et al. [2016] discuss dynamic interactions in modelling stock market interactions, in terms of a specific stock's influence on others.

4.1.2 Motivation and Contributions

From literature we noted that the most frequent modelling approach for microbial interactions are not suited for non-linear modelling or for representing temporally varying interaction parameters. Furthermore we noted temporally varying parameters were successfully used in other similar systems. With this in mind, we engaged in an exploratory study to investigate the temporally varying nature of microbial interactions.

Our contributions from this study lie in the area of exploring temporal dynamics of microbial communities. Key contributions summarised:

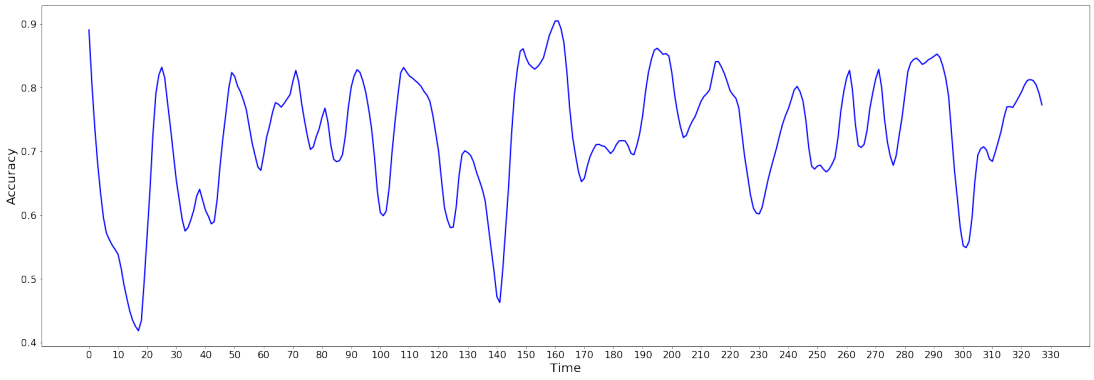

Figure 4.1: MIN inferred from the first ten data-points is verified for accuracy with the subsequent time-intervals of ten data-points. LOESS trend line is shown.

- Exploration into the time-variant nature of microbial interactions
- Visualisation approach for dynamic Microbial Interaction Networks (MINs)
- Investigation of the nature of the temporal behaviour of the individual interactions, and the relationship between the temporal variations of different pairwise interactions.

4.2 Results

We investigated the dynamic nature of MINs and propose a new visualisation approach for dynamic MINs.

4.2.1 Dynamic Nature of Microbial Interactions

Using our extension on IMPARO on the data from Caporaso et al. [2011] at the family taxonomic level, we inferred a MIN for the time interval comprising of the first ten data points. This resulted in an Reconstructed Abundance Profile Accuracy [Vidanaarachchi et al., 2020] of nearly 90%. However using the same MIN for different sliding window intervals of ten data points resulted in the results shown in Figure 4.1. Average accuracy for the entire data set was 72.42%.

4.2.2 Visualisation of Dynamic MINs

Traditional visualisation techniques for microbial interaction networks, such as heat maps and directed graphs, are not well equipped to convey temporal dynamics of MINs. Hence we propose the use of circular flow diagrams [Abel and Sander, 2014] to represent MINs. Snapshots of a dynamic MIN inferred on the female faecal microbiome of Caporaso et al. [2011] is shown in Figure 4.2. Each chord represents a single

interaction, whose magnitude is represented through the width of the chord. The direction is represented by the arrow, while a red border around the chord indicates negative values. Multiple such diagrams are dynamically connected to visualise the trends of change in MINs.

4.2.3 Further Insights into Interaction Dynamics

We further investigated the nature of the temporal variance of the microbial interactions. In Figure 4.3 we have indicated how a single-pairwise interaction undergoes a significant change with time. The indicated values were obtained by using IMPARO [Vidanaarachchi et al., 2020] in a sliding window approach as described in Section 4.5.

4.2.4 Categorisation of Temporal Behaviour of Microbial Interactions

After obtaining temporal variation of pairwise microbial interactions, we investigated the relationships among the temporal variation patterns for different interactions. To visualise these relationships we used UMAP [McInnes et al., 2018] on the temporal interaction strength variation curves to obtain points in a reduced dimension. We can observe clear clusters as presented in Figure 4.4.

4.3 Discussion

In this section we discuss the results that we obtained regarding dynamic microbial interaction networks and future research directions it leads to.

4.3.1 Dynamic Nature of Microbial Interactions

Many algorithms consider MINs to be static. However, in reality this is not the case. Microbial communities change over time, and the nature of their interactions may also change over time. As we showed in Figure 4.1, the inferred interactions on one time interval do not fit similarly to all time intervals. In further investigations, we observed that single pair-wise interactions change over time, and that temporal variation could potentially be classified into two main categories.

Dynamic interactions are found in other biological systems as well. In [Song et al., 2009a] the authors show that interactions in gene expression during yeast cell cycles and EEG data during motor imagination tasks can be successfully expressed using Time Varying Dynamic Bayesian Networks.

4.3.2 Parallels with Stock Market Systems

There are non-biological systems which have temporally dynamic interactions. [Tan et al., 2016] presents the time varying nature of stock interactions in the Shanghai stock market and discuss techniques associated with the inference of the interactions.

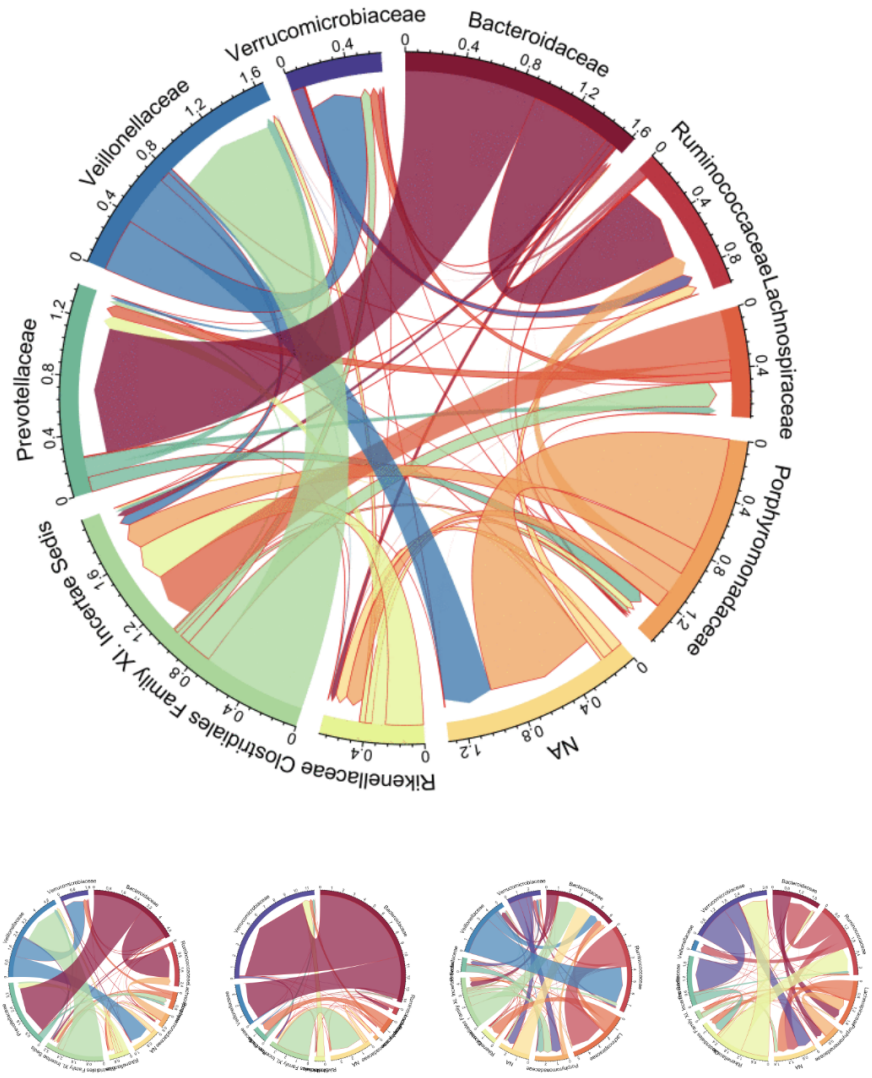

Figure 4.2: A dynamic Microbial Interaction Network inferred on the female faecal microbial sample of Caporaso et al. [2011] at family taxonomic level using IMPARO [Vidanaarachchi et al., 2020]. Interactions with red outlines are negative interactions. All interactions are directed in the arrow direction. Width of the chord represents the magnitude of the interaction, while a red border around the chord indicates negative values. Sampled snapshots are shown. (This dynamic MIN visualisation is originally presented in the format of an animated GIF.)

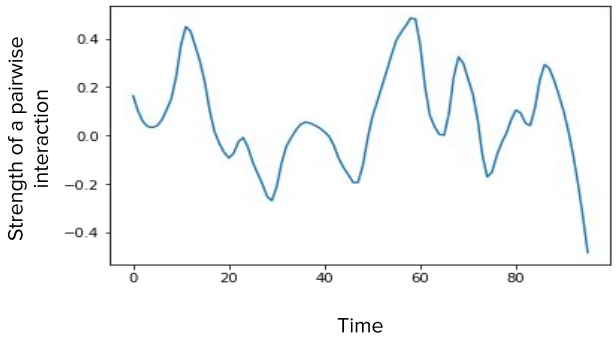

Figure 4.3: A single pairwise interaction's temporal change. The LOESS trendline is shown here against sampling time-points. We note that this interaction oscillates with time, with changes from positive to negative and *vice versa*. The peaks of the interaction strength show a similar notion of periodicity as seen in the Figure 4.1.

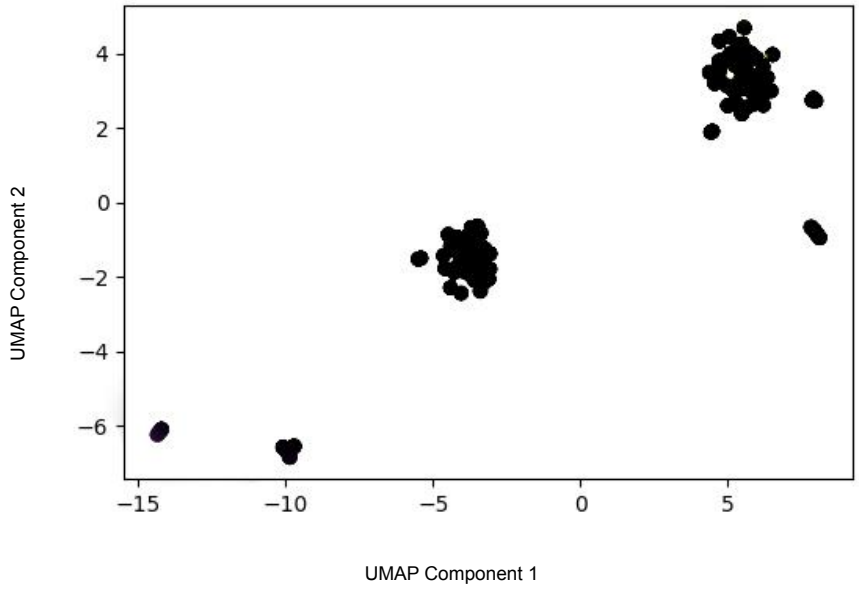

Figure 4.4: UMAP visualisation of microbial interaction dynamics. This figure shows the temporal variation patterns of 400 pairwise interactions (between 20 OTUs) are plotted using the Uniform Manifold Approximation and Projection (UMAP) dimension reduction technique. Interaction strengths at each time point were considered to be the dimensions in the higher-dimensional space. UMAP was used to reduce this into the lower-dimensional space. As UMAP is a neighbourhood preserving dimensional reduction technique, the neighbourhoods in the 2D plot are indicative of the neighbourhoods in the higher dimension. This result indicates that the temporal behaviour of microbial interactions can potentially be classified into a limited set of categories.

This work is an ideal example to observe time varying interactions in a complex system with multiple interacting units. Stock market systems contain data in a scale not available in microbial systems; hence observations from this parallel system are interesting. Despite the fundamental differences in a stock market system and a microbial community, at an abstract level these two systems have a multitude of parallels. When we consider stock price and microbial abundance to be parallel quantifiers in each system, we observe other parallels in the way that these two dynamic systems operate. Hence, exploring other similar systems would be of interest.

4.3.3 Future Work

In this subsection, we discuss some of the future research directions that can follow the ideas investigated in this exploratory study.

4.3.3.1 Testing on Real Life Data

When considering real life data, the actual microbial interactions are not well known. Hence verification of inference methods is not straight forward. Different methods use various solutions to overcome this issue. Some use the inferred interactions fed into a mathematical model for microbial community propagation to recreate the longitudinal microbial abundance profiles. Others prefer *in-vitro* methods, where they cultivate selected species under controlled environments to observe whether the change in one species is proportional to the change in another. Some studies simply consider the inference of interactions previously reported in literature as sufficient verification. However, considering the temporal dynamics of the interactions will further complicate this verification process.

Some publicly available real data sets, which are suitable for this kind of a study due to their high number of data points are listed below:

- Moving Pictures of the Human Microbiome [Caporaso et al., 2011]
- American Gut Project Data [McDonald et al., 2018]
- Murine Gut Data [Marino et al., 2014]
- Infant Gut Data [Koenig et al., 2011]
- Microbial Community on Cheese[Mounier et al., 2008]

4.3.3.2 Testing on Simulated Data

Many methods are verified against simulated data, as well as real life data. Simulated data have two advantages. Firstly, the actual interactions are known for verification purposes. Secondly, data could be synthesised for a variety of scenarios in high volume. The disadvantage is that the data is generated with certain assumptions in place and using mathematical models which are typically the same as the models used in the inference process. To circumvent issues with circularity, some methods

sample real life data sets to create varied simulated data which do not conform to a mathematical model. However this approach has the disadvantage of not knowing the underlying interactions.

4.3.3.3 Quantifying the effect of external factors

As mentioned in Section 4.1, changes in the microbial abundances are not exclusively due to internal factors. Yet in many inference algorithms, it is assumed that there are no external influences on the abundance of OTUs. With Hidden Markov Models (HMM), it would be possible to quantify various external factors as a hidden state.

4.3.3.4 Existence of Multiple Solutions within a Dynamic System

The assumption of the existence of a unique solution does not always hold. [Cao et al., 2017; Vidanaarachchi et al., 2020]. IMPARO seeks to loosen this assumption and to provide multiple solutions. Gao et al. [2018] combine multiple intermediate solutions into a unique MIN. Further research is possible in examining the possibility of multiple solutions and their interpretations, noting that the existence of multiple solutions is a separate question from that of dynamic solutions to the MINs. Specifically, investigating how the probability of different solutions changes over time would be interesting.

4.3.3.5 Rarer OTUs and their Effect

Most microbial interaction inference methods do not treat rare OTUs equally. However, as microbial communities are heterogeneously influential, the low abundance of a certain OTU is not a guarantee that the species is not influential. Likewise, not all abundant OTUs will have a great influence on the community. It would be of interest to consider techniques which allow the exploration of the influence of rarer OTUs. In some microbial profiles, rare OTUs seem to have abundance bursts as shown in Figure 4.5. As a result of considering microbial interactions to be dynamic, we can now investigate whether the rarer OTUs have a higher influence at peaks of abundance.

4.3.3.6 Collective Pattern Recognition

There is a possibility of improving the quality of microbial inferences by supplementing a singular data set with other parallel data sets. Lugo-Martinez et al. [2019] suggest promising results in using parallel data sets (for example, data sets from multiple newborn infants, aligned using time of birth as reference). Computational challenges associated with this approach include aligning the timescale and recognising patterns which are common across the set of samples. We take on this challenge in the following chapter (Chapter 5).

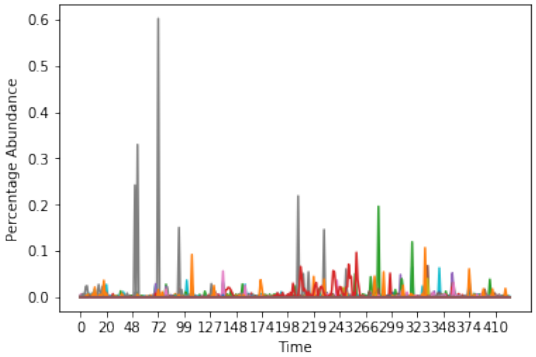

Figure 4.5: Abundance profiles of very rare OTUs (average abundance $\leq 0.01\%$) in a male faecal data set at family taxonomic level from [Caporaso et al., 2011]. Even the very rare OTUs have bursts of high abundance at some timepoints.

4.4 Conclusion

We compared and contrasted existing work and methodologies used in inferring microbial interactions in detail. We proposed that Microbial Interaction Networks (MINs) should be considered as dynamic systems, and presented a new visualisation. Then we explored the nature of the temporal variance of the interactions, and suggested that it shows signs of categorisation. Furthermore, we discussed future research possibilities in computational inference of microbial interactions.

4.5 Methods

We extended IMPARO [Vidanaarachchi et al., 2020] to investigate the dynamic nature of MINs. IMPARO has reported its accuracy in the order of 80% for real-life data sets. The authors explain that the high error margin is partly due to their assumption of microbial interactions as static. Our method of investigating the dynamic nature of microbial interactions is described below.

- Let X be a microbial abundance profile of time duration T for N OTUs, where relative abundance of each species at time points $(0, 1, 2...T)$ are given.
- Let $X[t_1, t_2)$ be the time interval of X, such that, for any given time point, $t_1 \leq t < t_2$.
- Let $A = f_{IMPARO}(X[t_1, t_2))$ be the MIN inferred through IMPARO by examining the time interval $X[t_1, t_2)$.
- Let $acc = f_{RAPA}(A, X[t_1, t_2))$ be the Reconstructed Abundance Profile Accuracy (Sub Section 4.5.1), when A is applied to the time interval $X[t_1, t_2)$.
- Let Δt be a small time difference.

First we consider $A_1 = f_{IMPARO}(X[t_1, t_1 + \Delta t_a))$. Then we consider the indexed set Θ, consisting of sliding window time intervals of size Δt_b, such that, $\Theta = \{X[0, 0 + \Delta t_b), X[1, 1 + \Delta t_b), ..., X[T - \Delta t_b, T)\}$. Lastly we plot the series of accuracy metrics, where, $acc_i = f_{RAPA}(A_1, \Theta_i)$, for $0 \leq i \leq T - \Delta t_b$.

4.5.1 Reconstructed Abundance Profile Accuracy (RAPA)

The reconstructed abundance profile accuracy [Shaw et al., 2016; Vidanaarachchi et al., 2020] is a metric of how accurately the original abundance profile can be reconstructed with the inferred MIN. This is useful with real-life data, because the ground truth is not known in such cases. MINs which are capable of recreating the original abundance profile closely are considered to be be good approximations. Using the original initial conditions, $X[0]$, the subsequent microbial community compositions are calculated using the generalised Lotka–Volterra model. This reconstructed microbial abundance profile is then compared to the original using the Bray-Curtis Dissimilarity [Bray and Curtis, 1957].

4.5.2 Locally Estimated Scatterplot Smoothing (LOESS)

Locally Estimated Scatterplot Smoothing (LOESS) is a local regression method, built on classical regression methods, with local weighting, based on the concepts of the Zavitzky-Golay filter [Savitzky and Golay, 1964]. It obtains a smooth plot from a set of scattered data-points while respecting local trends.

4.5.3 Uniform Manifold Approximation and Projection (UMAP)

Uniform Manifold Approximation and Projection (UMAP) [McInnes et al., 2018] is a manifold learning technique for dimension reduction. It is considered to have high visualisation quality and preserve more global structure than other dimension reduction techniques such as t-SNE [Van Der Maaten and Hinton, 2008]. For our application of comparing microbial temporal dynamics, UMAP is especially suitable because of its global structure-preserving nature.

Chapter 5

CoPR: Collective Pattern Recognition—a Framework for Microbial Community Activity Analysis

This chapter is partially composed of material that appears in a manuscript titled "CoPR: Collective Pattern Recognition—a Framework for Microbial Community Activity Analysis" by Vidanaarachchi R., et al. that is being finalised for submission to a journal.

Summary

Background Microbial community activities provide essential information on understanding bacterial communities. Unfortunately, they are generally not directly observable. We rely on longitudinal abundance profiles to get insight into microbial community activities. Often datasets do not have sufficient longitudinal sampling points to successfully apply our algorithms. Hence, in this chapter, we are interested in analysing multiple datasets from similar environments to alleviate the aforementioned problem. Furthermore, we wish to see whether collective pattern recognition would enhance our understanding of microbial community activities.

Results In this chapter, we present CoPR, a framework for collective microbial longitudinal abundance data. Our visualisation shows that a single pattern for temporal abundance variation does not exist. However, it also indicates that even complete individuality does not exist. Consequently, our visualisation highlights the individuality and conformity in the temporal variation of abundance profiles of similar host environments. We also identify different characteristics in the TVAP (Temporal Variation of Abundance Profile) patterns with regards to cohesion and separation.

Conclusions CoPR helps gain essential insights into the microbial communities and their heterogeneity. This chapter also highlights the choice between individuality and conformity in microbial community data analysis.

5.1 Background

In the previous chapters, we discussed that we cannot generally observe microbial activity in the host environment and that we rely on longitudinal observations of microbial abundance profile data to infer their activity. With IMPARO [Vidanaarachchi et al., 2020], we explored techniques for using mathematical models and optimisation to interpret the temporal variation of the abundance profile and infer interactions in the microbial communal activity.

In this chapter, we continue our quest for answering the question of "What are they doing [in the microbial communities]?" [Boon et al., 2013]. In doing so, rather than quantifying the microbial behaviour as in the previous chapters, we look at the patterns in the temporal variation of the abundance profile (TVAP patterns). We can define a TVAP pattern as a particular pattern observable in the graph of abundance against time. It can be unique to a certain OTU or a specific host environment. We believe that comparing and contrasting TVAP patterns can infer insights into microbial community activity.

The inference of microbial activity through interaction inference algorithms is heavily reliant on the quality of available datasets. A useful dataset's favourable qualities are high sampling frequency, consistent sampling frequency, and numerous sampled time points. Unfortunately, most of the available datasets do not feature these qualities.

For example, let us examine the Moving Pictures of the Human Microbiome study [Caporaso et al., 2011], which was analysed in Chapter 3. This study was conducted over up to 15 months over 396 time points and provided time-series microbial abundance data of two individuals at four body sites [Caporaso et al., 2011]. The sampling frequency of this study is daily. We achieved high accuracy in inferring microbial activity in this dataset [Vidanaarachchi et al., 2020].

However, such datasets are not very commonplace. To illustrate the difficulty in collecting such a dataset, consider the scenario of a longitudinal study of the gut microbiome of a healthy individual. An individual available for daily stool analysis for six months would be the primary requirement of such a study. Furthermore, there will be barriers in terms of cost in studies that require specialist sample collection, such as coral microbiome collection requiring expert divers. On the other hand, some studies naturally have a shorter duration of interest, as in the menstrual vaginal microbiome—where the sampling needs to happen during the menstruation period [Song et al., 2020].

Literature shows that many studies collected data parallelly from similar microbial communities–we consider microbial communities inhabiting the same type of host environments to be similar microbial communities. When a single dataset is not

sufficient for inferring microbial communal activity, it has been shown that multiple datasets from similar communities could supplement the lack of data and provide more accurate inferences [Lugo-Martinez et al., 2019]. In this chapter, we explore the idea of collective pattern recognition for enhancing our understanding of microbial community dynamics.

For example, one such dataset we explore in this chapter comes from a study by La Rosa et al. [2014]. They have sequenced the microbial communities in the guts of multiple premature infants in a neonatal Intensive Care Unit. We consider the premature infant gut as the type of host environment. Hence we use collective pattern recognition on microbial abundance profile datasets from all the infants in the study. With an average of 15 time-points per dataset, each dataset lacks sufficient data to successfully infer the microbial community dynamics. However, given that there are data from 58 infants, we can compare and contrast the different infants' TVAP patterns and figure out common patterns of microbial behaviour—thus, this chapter's motivation is to combine datasets to recognise patterns collectively.

A second issue prevalent in analysing microbial community activities is the lack of independence from clinical or environmental factors. While the longitudinal abundance variation patterns reflect microbial activity (interactions), it is hard to discern the difference between the change in abundance due to internal—trophic or non-trophic—microbial activity and the change due to clinical factors and external influence. Even if the clinical data and a subset of environmental details are available, it is difficult to eliminate their factor into the abundance variation completely.

Using collective pattern recognition also allows discarding external factors up to an extent. Unless similar external factors affect all the host environments, collectively looking at the TVAP would identify patterns common to the host environment itself. If there are limited ways external factors affect the microbiome, collective pattern recognition allows identifying which host environments have been affected by the external factors.

Given the above reasons, we believe that collective pattern recognition will improve our understanding of microbial community activity.

5.1.1 Related Work

We explore related work under four main themes. Firstly, we look into the nature of the datasets available to gain insights into the requirements of working with similar data. Secondly, we look into microbial activity inference methods, as our end goal is to be facilitating the activity inference processes. Thirdly, we look into collective pattern recognition and clustering approaches, which plays a crucial role in our research. Fourthly, we look into the existing literature on individuality and conformity in microbial communities, under which we explore the ideas of microbial signatures, community state types, and precision medicine.

5.1.1.1 Microbial Abundance Datasets

In this section, we will summarise some datasets where 16S rRNA sequencing has been used to collect time-series data from multiple similar host environments, which we already mentioned in Chapter 2.

Premature Infants' Gut Microbiome La Rosa et al. [2014] presents data from 58 neonatal infant gut microbial communities. This dataset was of interest as all the samples were collected while the infants were undergoing care at the neonatal intensive care units, which limited the gut microbiome's interaction with the outside world, thereby limiting the external factors into the dynamics of the microbial community. They have collected 922 samples with an average of just over 15 per infant, up until 36 weeks of post-conception age, for all stool passings for each infant.

Vaginal Microbiome of Reproductive-Age Women Gajer et al. [2012] presents a dataset from 32 reproductive women's vaginal microbial communities. They collected 937 samples over 16 weeks, with twice a week sampling frequency, averaging just over 29 samples per woman.

Human Microbiome Related to Pregnancy DiGiulio et al. [2015] have collected over 2500 samples from 49 pregnant women, pre– and post–delivery. They collected microbial community samples from the vagina, distal gut, saliva and tooth/gum just under 20 samples per site per woman. The collection frequency was weekly during gestation and monthly after the delivery.

Neonatal Gut and Respiratory Microbiome Grier et al. [2018] have compiled another infant dataset from 82 infants. The data was collected over up to a year after birth with a weekly sampling frequency at the hospital and monthly thereafter. They have collected data from the gut (average of 13 per person) and respiratory tracts (nasal - an average of 12 and throat - an average of 6).

Availability of data of this nature makes our framework necessary and feasible.

5.1.1.2 Microbial Community Activity Inference

Out of the many microbial community activity inference approaches, most depend on high-frequency datasets with a higher number of data points. IMPARO [Vidanaarachchi et al., 2020] uses an evolutionary algorithm with Lotka-Volterra equations to approximate microbial interaction parameters. MetaMIS [Shaw et al., 2016] uses partial least square regression to estimate interaction parameters. Other methods include SparCC [Friedman and Alm, 2012], which uses statistical methods, and the method by Lugo-Martinez et al. [2019], which uses integrated data from multiple subjects in a Dynamic Bayesian network-based model. More about microbial community activity inference were discussed in Chapters 2, 3 and 4.

The majority of the above methods use time-series data from a single host environment, and the results are significantly impacted by the availability of a large number of frequently sampled data points. Also, they do not utilise the availability of datasets with time-series samples of multiple similar host environments.

5.1.1.3 Collective Pattern Recognition, Clustering, and Temporal Aligning Approaches

The idea of collective pattern recognition has been previously discussed in multiple works. Less so in the field of microbial interaction inference or related to microbial abundance data, but mainly in the area of gene expression analysis. As we can draw parallels between many biological data types, we will be looking at collective pattern recognition (including clustering and temporal aligning focused) work covering various applications.

Lugo-Martinez et al. [2019] has explored collective pattern recognition for inferring microbial abundance patterns successfully. They align the temporal variation patterns coming from multiple host environments and define a typical pattern. They use this common pattern together with a dynamic Bayesian network to successfully predict microbial composition.

Bar-Joseph et al. [2012] reviewed clustering mechanisms to explore the response to external signals in time-series gene expression data. They also reviewed combining time-series data with other dynamic and static genomics data to better model gene expression patterns.

Bar-Joseph et al. [2003] used dataset alignment techniques and clustering to estimate unobserved data points in gene expression data. With datasets aligned by modelling them as piecewise polynomials, they had been able to achieve biologically meaningful results.

Smith et al. [2009] also used time-series aligning techniques for temporal gene expression data. What makes their work interesting is that they present clustered data alignment, removing the assumption that all genes share the same alignment. Theirs is the first work to not treat gene expression data as homogeneous. Their inter-cluster independence in aligning temporal data provides more accurate alignments than earlier methods.

Aach and Church [2001] used time warping algorithms when working with RNA and protein expression datasets. They show that time-warping clustering is superior to standard clustering using both interpolative and simple time-warping techniques. Their work is interesting in not assuming time-series biological data to be homogeneous in their temporal variation.

Criel and Tsiporkova [2006] also used time-warping techniques for alignment and template matching of time-series gene expression data. In their work, they have adapted dynamic time warping techniques from speech recognition research.

Dong et al. [2020] introduce a statistical framework for co-expression networks. They use a kernel function to measure the similarity between subjects, which we identify as a collective pattern recognition technique. Their method, applied to time-series gene expression profiles of a group of subjects with respiratory virus exposure, produced early and accurate results.

Somani et al. [2020] also analyses gene-expression data in their recent work. While using data from multiple subjects to model disease-relevant pathways, they also allow personalisation through a Gaussian process to identify differentially expressed genes. Their work claims to be more robust for identifying disease-relevant pathways in heterogeneous diseases.

Chandereng and Gitter [2020] recently used time-series clustering techniques with lag penalisation for gene expression and protein phosphorylation datasets. They successfully identify clusters with distinct temporal patterns in both yeast osmotic stress response and axolotl limb regeneration studies. This study exemplifies that heterogeneous temporal variation behaviour is observed across various biological processes.

Jiang et al. [2020] used dynamic time warping techniques for comparative time-series transcriptome analysis in their recent work, TimeMeter. They were successful in characterising complicated temporal gene expression associations. They uncover exciting patterns in mouse digit restoration and axolotl blastema differentiation datasets.

5.1.1.4 Individuality and Conformity

Individuality is identified as a microbial community behaviour, which is not uniform across communities of similar nature. Conformity is the opposite when similar OTU communities behave in a set predictable pattern. The individuality of microbial communities has been identified in the literature [Martins and Locke, 2015]. Especially in research on the gut microbiome, precision medicine has been proposed and successfully used in several studies [De Filippis et al., 2018; Cammarota et al., 2020]. Conforming behaviour has also been reported in the literature [Gong et al., 2016]. In the case of OTU communities, the concept of Community State Types [Ravel et al., 2011] is an existing approach of explaining the balance of individuality and conformity in microbial communities.

Introduced by Ravel et al. [2011], Grier et al. [2018], DiGiulio et al. [2015] and other studies report community state types (CSTs) in various microbial communities. The

idea of CST is based on the composition of the constituent OTUs; as such, it is defined for a snapshot in time. DiGiulio et al. [2015] further notes that some communities show different state types, while some may show the same CST throughout the entire sampling period.

However, most microbial activity inference methods consider a single microbial community in their inference process, thus taking a highly individualistic approach. Some reasons for this could be the complexity of the external factors affecting the microbial community dynamics [Vidanaarachchi et al., 2020].

Lugo-Martinez et al. [2019], however, uses a unified model obtained by aligning different microbial community TVAPs and assumes that a general microbial community of that particular type will take a particular pattern. DiGiulio et al. [2015] considers a vaginal community signature in their analysis of the vaginal microbiome.

Looking at the literature cited above, we can acknowledge that both approaches in considering individuality and conformity in the microbial community analysis process have their own merits.

5.1.2 Motivation and Contributions

It was interesting to note that collective pattern recognition goes hand-in-hand with individuality and conformity of microbial community dynamics. Considering the availability of studies where multiple temporal datasets of similar environments are available and the successful prior use of collective pattern recognition for biological data analysis, we further explored collective pattern recognition for microbial community dynamics analysis. We were motivated to use the collective pattern recognition techniques to shed light on individuality and conformity in microbial community dynamics and the heterogeneous nature of microbial abundance datasets.

We use unsupervised learning and visualisation techniques to analyse microbial abundance datasets and examine the TVAP patterns. We also talk about individuality and conformity, and heterogeneity of data in relation to microbial abundance datasets in our work.

In this chapter, we present CoPR (Collective Pattern Recognition), a framework for analysing microbial community activities, which aims to address problems in lack of temporal abundance data and effects of external factors. CoPR primarily clusters OTU communities based on their TVAP patterns.

Ours is the first work to analyse the balance between individuality and conformity in microbial community activity patterns. Many existing works attempt to isolate a single pattern for temporal activity in microbial communities when working with multiple datasets. However, we consider microbial communities to be individualistic to a certain degree and look at multiple temporal activity patterns microbial communities can follow. Thus we believe our framework allows a more accurate analysis of microbial community activity.

CoPR also considers the heterogeneity of microbial datasets. We talk about microbial heterogeneity on multiple aspects and show that microbial abundance data, when treated as non-homogeneous, can uncover important details about the tempo-

ral community activity.

We present the analysis of multiple real-life datasets and simulated data. Our analysis identifies that the qualities of individuality and conformity in microbial communities are present across varying taxonomic resolutions, abundance levels and host environment types.

Key contributions summarised:

- Identifying correlations between different OTU TVAP pattern clusters.
- Framing the discourse around the balance between individuality and conformity, which we believe is essential to understanding microbial community activity.
- Exploration of heterogeneity of OTU TVAP patterns.
- Verification of the framework through the analysis of multiple real-life datasets (varying taxonomic levels, abundance levels, from different host environments, etc.) and simulated data.

5.2 Results

We processed several datasets—both real-life and simulated—through the pipeline and visualised patterns in the temporal variation of abundance profiles (TVAP patterns). Then we analysed the visualisations to illustrate how we can use them to gain insights into the microbial community dynamics. We present the results that illustrate some of the key arguments in this section. First, let us look at an example visualisation (Figure 5.1) to clearly understand the underlying meaning.

5.2.1 Non-Conformity Among the Communities of the Same OTU in Different Host Environments

We explore the temporal variation of abundance patterns of four OTUs—*Bacilli, Actinobacteria, Clostridia,* and *Gammaproteobacteria*—in the neonatal infant gut data set [La Rosa et al., 2014]. These particular OTUs were selected as they form the intersection of the ten highest abundant (averaged over time) OTUs in all the host environments (infants)—we shall call these the major OTUs (see Section 5.5.3).

The first observation we would like to draw attention to is how each OTU's TVAP is separated into different clusters, as seen in Figure 5.2. This clustering indicates no conformity to a typical pattern observed for an OTU across different communities. This heterogeneous behaviour can again be identified in the TVAP curves seen in Figure 5.3.

5.2.2 Conformity Among the Communities of the Same OTU

Secondly, we observe that the different communities neither subscribe to a typical behaviour nor completely sporadic. Let us consider the clusters shown in Figure 5.2.

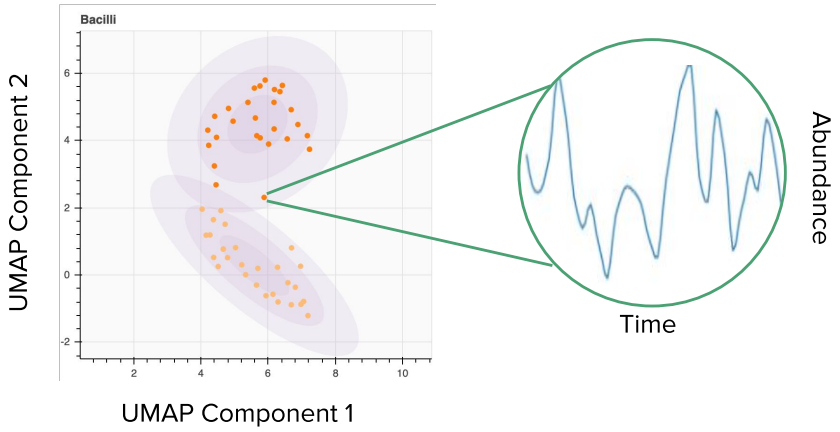

Figure 5.1: An example visualisation. In the cluster plots, each axis represents a dimensionally reduced component. Each dot in the cluster plot represents a specific environment's temporal variation of the abundance profile (TVAP) of *Bacilli* (in this example). Host environments where *Bacilli* show similar TVAPs are clustered together, while host environments where *Bacilli* show distinct TVAPs are placed further away from each other. Trios of co-centred circles represent Gaussian Mixture Model clusters, where the varying opacity indicate the likelihood of a datapoint belonging to that cluster.

Especially the clusters of *Gammaproteobacteria* are tightly knit together. Although we have clustered the *Gammaproteobacteria* communities into two clusters according to the silhouette value (see Section 5.5), we observe a distinctly identifiable subcluster within one of the main clusters. In Figure 5.3, we have increased the cluster numbers to closely observe how the temporal variation patterns differed. It indicates that although the primary separation is based on falling–rising behaviours, the subclusters differ on when the rise happens. The communities belonging to the smaller of the three clusters all show an initial rise in *Gammaproteobacteria* abundance. This cluster, along with the cluster in the third column of Figure 5.3, where a general rising behaviour is observed, formed the larger cluster of Figure 5.2 (indicated in deep red). Interestingly this conformance to a specific behaviour happens in certain subsets of communities.

Observing the TVAP plots in Figure 5.3, we also note that the deviancy from the cluster's median is different for each cluster. Observation of this heterogeneous behaviour is also of interest.

However, as a summary, we can say that there are three distinctly identifiable TVAP patterns amongst *Gammaproteobacteria* communities in the infant gut. The first is a gradual reduction of relative abundance; the second is an initial rise in relative abundance and subsequent maintenance. The third can be categorised as a late rise in relative abundance. In our original clustering of Figure 5.2, the two clusters represented a rise and a fall in relative abundance. In the finer clustering of Figure 5.3, the rise was characterised into two different rising patterns. Although this behaviour is demonstrated in the *Gammaproteobacteria* communities, sub-cluster separation can

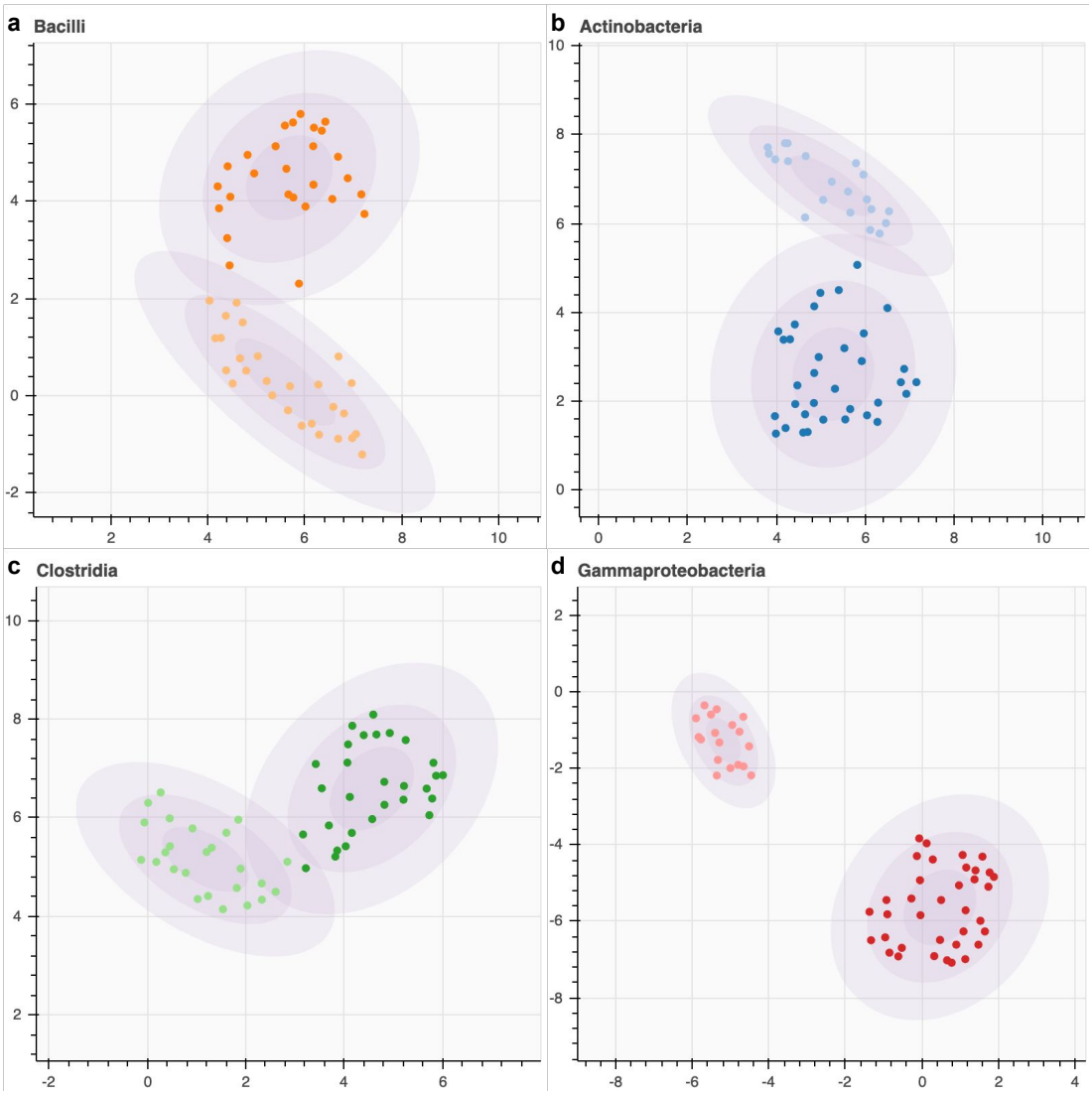

Figure 5.2: Clustering of TVAP of the major OTUs in the gut microbiome of preterm infants while in a neonatal ICU La Rosa et al. [2014] is shown in the figure. Distances between TVAP were calculated using dynamic time warp (DTW) distance and visualised with UMAP (Uniform Manifold Approximation and Projection). Colours are local to each figure and represent clusters identified through Gaussian mixture models (GMM) clustering, where the highest silhouette score determined the number of clusters. All four major OTUs (a. *Bacilli*, b. *Actinobacteria*, c. *Clostridia*, and d. *Gammaproteobacteria*) show clear cluster separation. *Gammaproteobacteria* shows the most explicit separation. Although the optimal cluster number, according to silhouette score, is two, we observe sub-cluster separations in the dark red cluster. **The axes in these plots are: UMAP Component 1 (x) and UMAP Component 2 (y).**

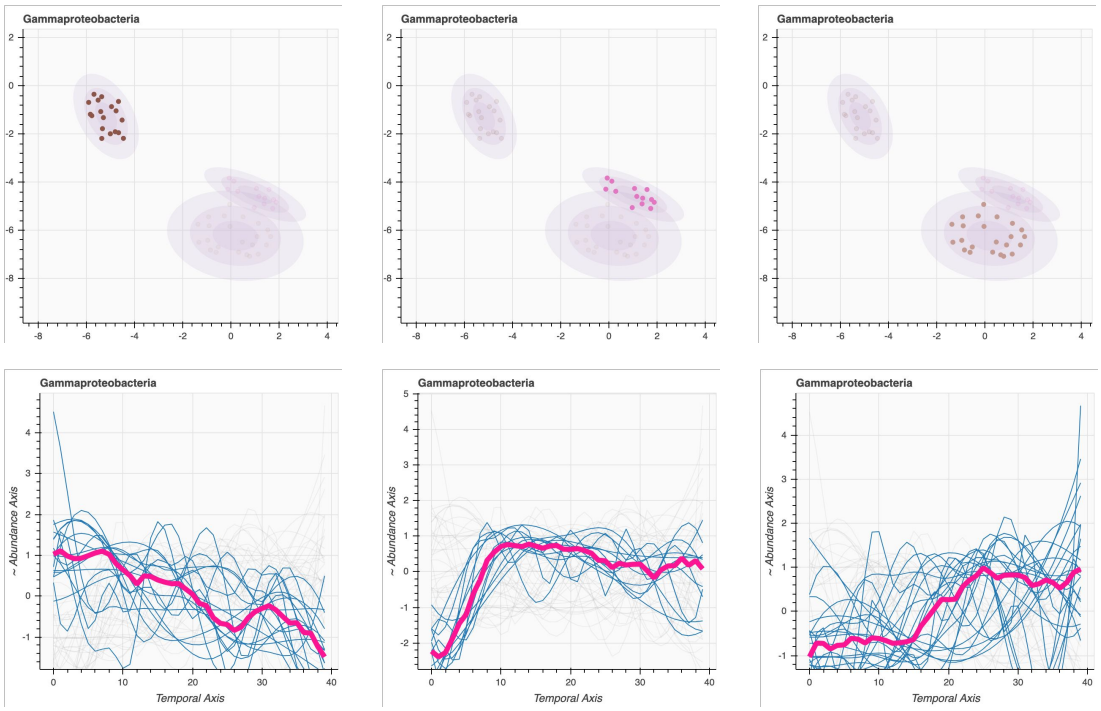

Figure 5.3: The figure shows distinctly identifiable TVAP patterns of *Gammaproteobacteria* in the La Rosa et al. [2014] dataset. Each subfigure in the top row highlights a separable cluster of subjects, and corresponding subfigures in the bottom row show the TVAP of the highlighted subjects (blue lines). The thick pink line represents the median. The highlighted cluster in the third column shows more variation than the other two. It can be observed that the cluster from the first column encapsulates subjects in which the *Gammaproteobacteria*'s relative abundance drops from birth to discharge. In the second column, the relative abundance rises steeply soon after birth and maintains an abundance level thereafter. The third column is harder to classify clearly but shows a general trend of rising and falling while favouring a final rise. **The axes in the top row plots are: UMAP Component 1 (x) and UMAP Component 2 (y).**

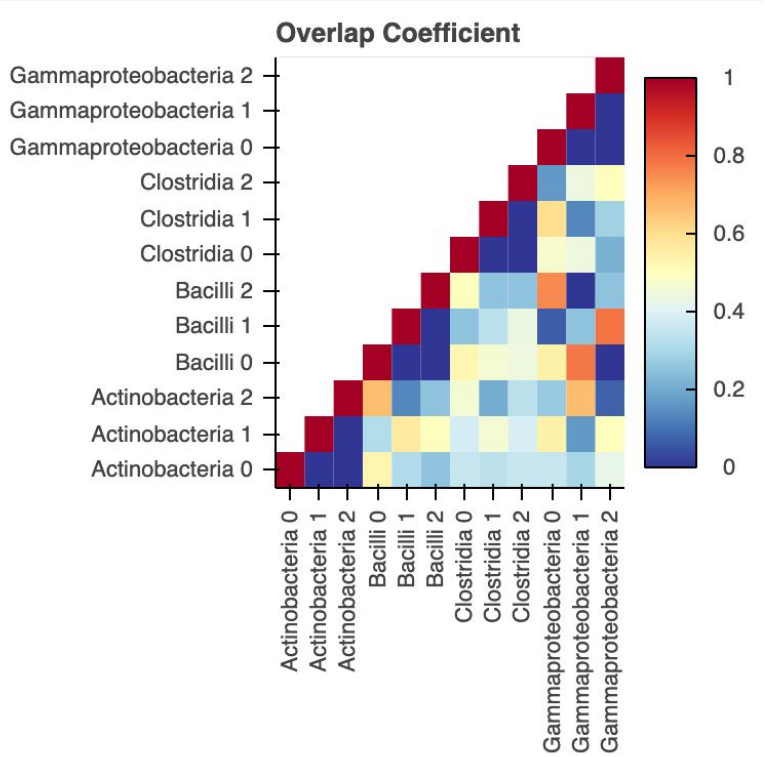

Figure 5.4: This figure shows the distribution of the overlap coefficient between the different clusters in the analysis of the La Rosa et al. [2014] dataset. *Gammaproteobacteria*, *Clostridia*, *Bacilli*, and *Actinobacteria* were clustered into three. Highest overlaps are observed in *Gammaproteobacteria* and *Bacilli* clusters.

be identified in other communities as well. We can discern the existence of possibly separable sub-clusters by observing silhouette scores.

5.2.3 Agreement of Clusters of Different OTUs

Next, we investigated the connection between the clusters of different OTUs. To quantify this, we looked at the overlap coefficient (see Section 5.5) of cluster membership. The cluster membership distribution calculated for the same dataset can be seen in Figure 5.4. Firstly, it gives us a quantitative sense of the overlaps, and secondly, it provides us with information that is hard to discern visually.

From Figure 5.4, we identified the overlaps of *Gammaproteobacteria* and *Bacilli*. This association between the three clusters were statistically significant, $\chi^2 = (4, N = 54) = 43.6075$, $p = 7.40076E-9$ (Fisher's exact test $= 1E-8$) (< 0.01). Furthermore, the relationship between the two main clusters of Gammaproteobacteria and Bacilli were statistically significant, $\chi^2 = (1, N = 54) = 10.7628$, $p = 0.001036 (< 0.01)$. We further explored the corresponding behaviour in Figure 5.5. In the left half of the figure, we observe that *Gammaproteobacteria*'s TVAP rises initially and maintains that level, while *Bacilli*'s corresponding TVAP falls and maintains very low. Interestingly

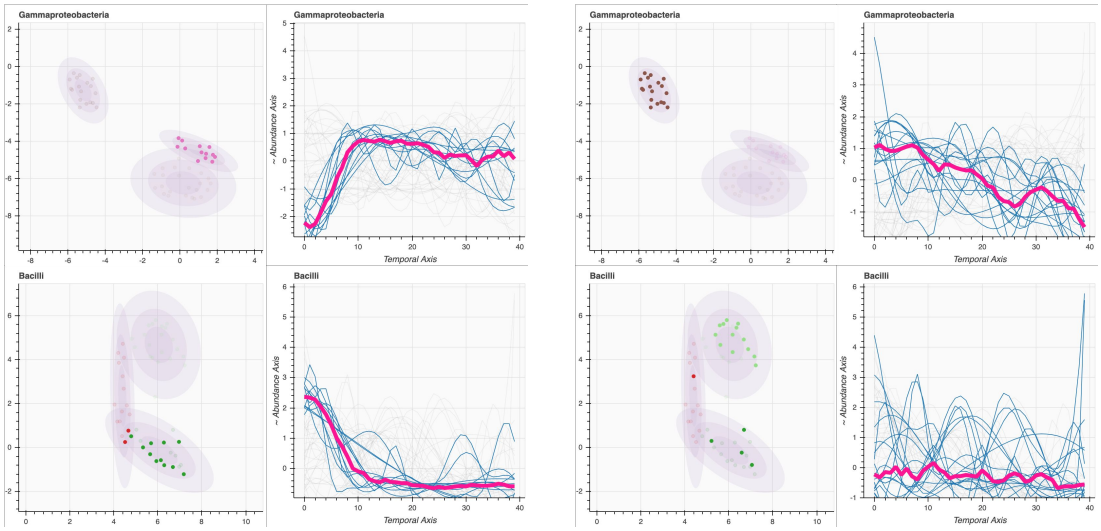

Figure 5.5: The figure shows the corresponding behaviour of *Gammaproteobacteria* (top row) clusters with *Bacilli* (bottom row) clusters. Subfigures show TVAP clustering and the TVAP curves corresponding to the highlighted host environments (infants), with the median curve for the subset of highlighted host environments shown as the thick pink line. Clusters on the left show that the initial rise in Gammaproteobacteria corresponds with an initial fall in Bacilli (The overlap coefficient for these two clusters are 71% with a Jaccard Index of 48%). Clusters on the right show that decreasing relative abundance of Gammaproteobacteria corresponds with the rising and falling behaviour of Bacilli. Observations seen in the clusters on the left are in line with the observations of La Rosa et al. [2014]. Still, while they generalise this observation for the whole community, we highlight that this is only one of the three distinctly identifiable trends observable in this dataset. **The axes in the cluster plots are: UMAP Component 1 (x) and UMAP Component 2 (y).**

while this exchange of prominence in the bacterial community is clear, we cannot find a clear explanation for the corresponding behaviour in the other two clusters from the median TVAP patterns.

5.2.4 Clusters and External Factors

We were interested in investigating whether the clustering of host environments had a connection to the external factors. Some datasets include clinical factors and other environmental variables, which could broaden the value of our analysis. For example, in our analysis of the neonatal gut microbial dataset [La Rosa et al., 2014], a connection between the infants' delivery method seemed to correlate to the clustering of *Bacilli* and *Gammaproteobacteria* TVAPs. However, a chi-square test of independence showed no significant association between the TVAP clustering and delivery method. $\chi^2 = (1, N = 54) = 0.8105$, $p > 0.1$

5.2.5 Common Themes in Multiple Real Life Datasets

We explore a second dataset, which consists of gut, nasal and throat microbiomes of infants Grier et al. [2018]. The primary aim of this exploration is to observe common themes which can be identified in real-life datasets.

In Figure 5.6, we note that six major OTUs are identified in the Grier et al. [2018] gut dataset. Among the six are the four identified in the La Rosa et al. [2014] study's analysis. The separation of clusters also shows similarity, with *Gammaproteobacteria* communities arguably separating well, although not as well as in the previous study. *Actinobacteria* is also notably separated into three different clusters. However, overall, in all the OTU *clusterings*, we can identify many clear subclusters. We also observe connected clusters in *Bacteroidia* and *Coriobacteriia* communities.

In Figure 5.7, we explore the TVAP patterns in the nasal microbiome from the Grier et al. [2018] study. We observe five major OTUs identified, of which *Clostridia*'s separation and *Betaproteobacteria*'s separation are clearly identifiable. Most noteworthy out of all is *Bacilli*'s TVAP patterns, which show a set of triangularly interconnected clusters. This cluster layout is an excellent example of the gradual behavioural change in the communities where a balance between individuality and conformity may exist. *Bacteriodia* and *Actinobacteria* show connected clusters as well.

In Figure 5.8, we explore the TVAP patterns of the throat microbiome of Grier et al. [2018], where six major OTUs have been identified. We observe clearly disjoint clusters in *Fusobacteria* and *Betaproteobacteria*. Throat microbiome is the second type of community where *Betaproteobacteria* shows good separation, and the third, when we consider *Proteobacteria* as a whole (Figures 5.6, 5.7 and 5.8). The cluster shapes and directions again prove interesting, with *Clostridia* showing a tree-like structure. This structure suggests that each cluster shows gradual changes in three different directions, deviating from a central pattern

5.2.6 Differentiating Disjoint Clusters and Connected Clusters

A secondary observation we can make from Figures 5.2 and 5.3 is that in some OTUs, the clusters are distinct and disjoint, while in other OTUs, the clusters are connected. The same behaviour is highlighted in Figure 5.9. Here we can identify another peculiar behaviour of TVAP patterns: Disjoint clusters represent TVAP patterns which correspond to a set behaviour, as opposed to connected clusters, we can see a gradual change of behaviour in TVAP.

5.2.7 Analysis Across Taxonomic Resolutions

In Figure 5.10, starting from the phylum (L2) taxonomic level, moving up to genus (L6) taxonomic level, we have demonstrated that the traits of individuality and conformity are present at varying taxonomic levels—in addition to the observations at the class (L3) level. Also, we observe the TVAP pattern clusters change as we traverse through the taxonomic hierarchy. Also noteworthy is the conserved structure of the taxonomic hierarchy of Phylum *Firmicutes*, Class *Bacilli*, Order *Lactobacillales*, Family

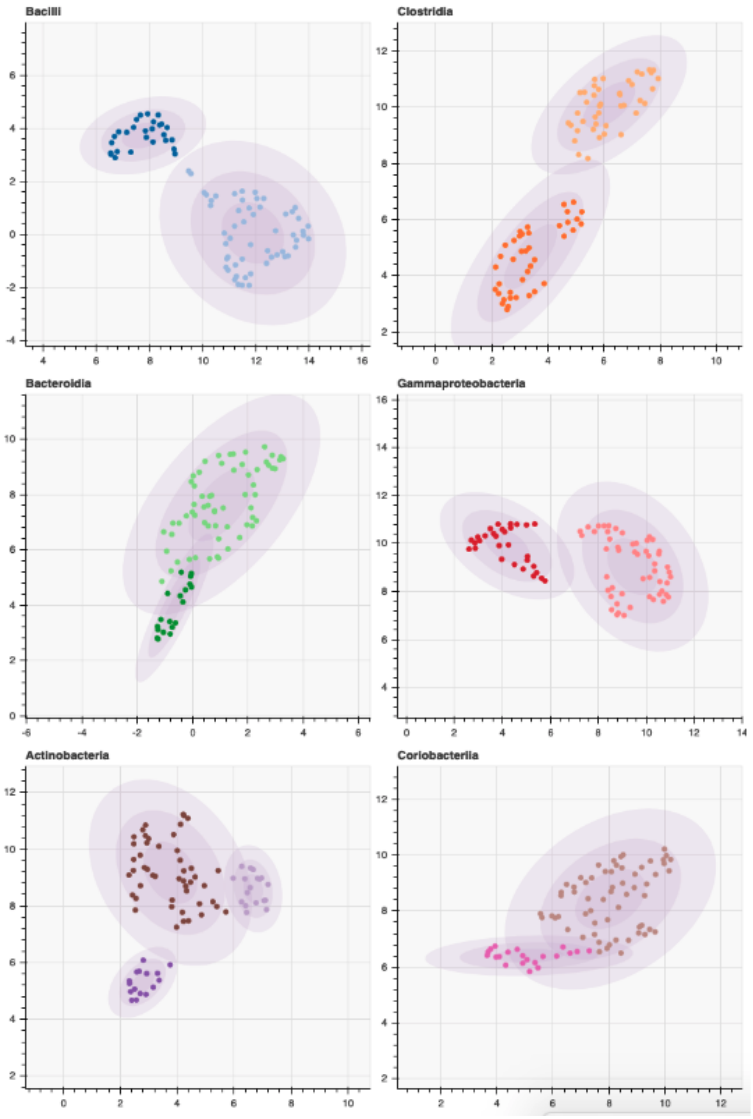

Figure 5.6: Clustering of TVAP of the major OTUs in the gut microbiome of infants from Grier et al. [2018] study are shown in the figure. Distances between TVAP were calculated using DTW, and and visualised with UMAP. Colours are local to each figure and represent clusters identified through GMM clustering, where the highest silhouette score determined the number of clusters. All four major OTUs (*Bacilli*, *Gammaproteobacteria*, *Clostridia*, and *Actinobacteria*) found in the La Rosa et al. [2014] dataset are also found to be major OTUs in this dataset, with the addition of *Corriobacterlia* and *Bacteroidia*. Similar to the La Rosa et al. [2014] study, *Gammaproteobacteria*, *Bacilli*, and *Clostridia* show clear separation into two clusters. **The axes in these plots are: UMAP Component 1 (x) and UMAP Component 2 (y).**

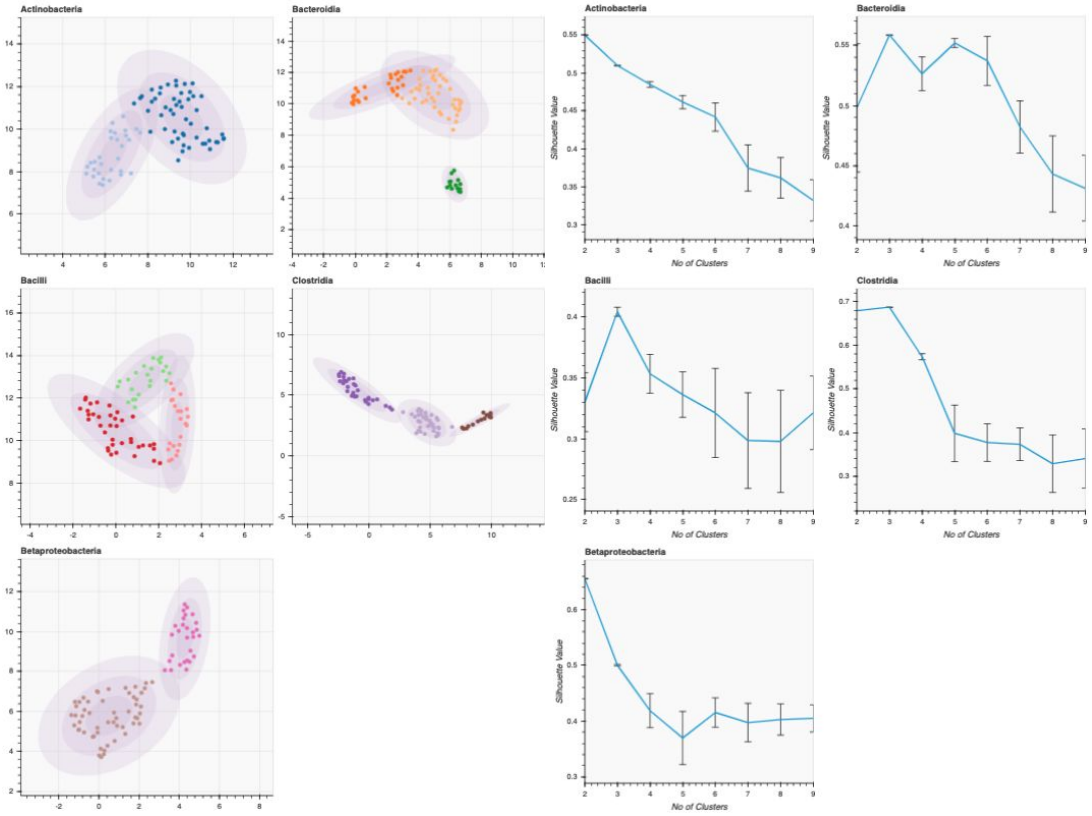

Figure 5.7: Clustering of TVAP of the major OTUs in the nasal microbiome of infants from Grier et al. [2018] are shown in the left-hand figure. Distances between TVAP were calculated using DTW and visualised with UMAP. Silhouette scores for different cluster numbers for each of the major OTUs are plotted in the right-hand figure. Colours are local to each sub-figure in the left-hand figure and represent clusters identified through GMM clustering. The number of clusters was determined by the highest silhouette scores, as shown in the right-hand figure. Three major OTUs (*Bacilli*, *Clostridia*, and *Actinobacteria*) found in the La Rosa et al. [2014] dataset are also found to be major OTUs in this dataset. The absence of *Gammaproteobacteria* could be explained by the aerobic environment of the nasal cavity. While some OTUs have silhouette scores indicative of a superior number of clusters, others have closely competing cluster numbers. **The axes in the cluster plots are: UMAP Component 1 (x) and UMAP Component 2 (y).**

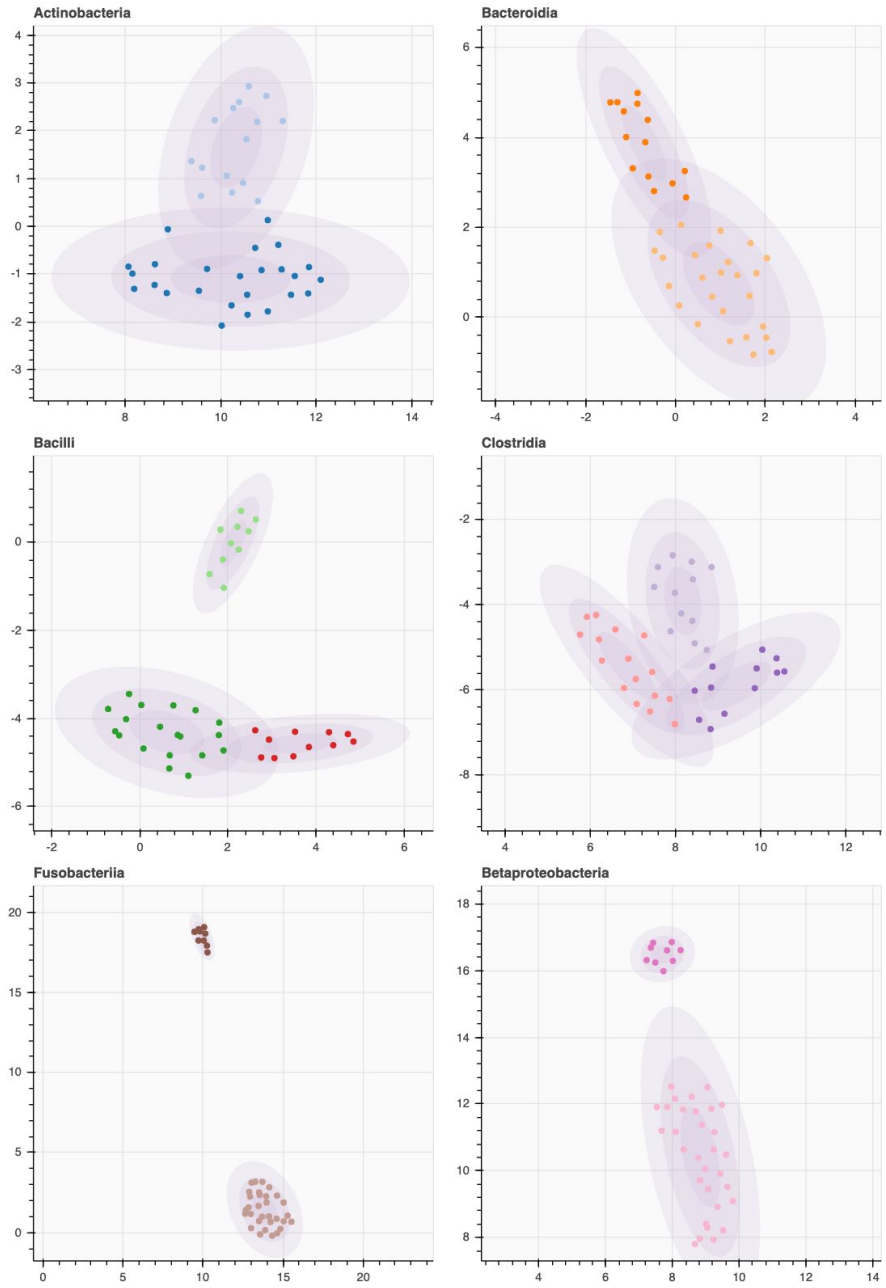

Figure 5.8: Clustering of TVAP of the major OTUs in the throat microbiome of infants from Grier et al. [2018] study are shown in the figure. Distances between TVAP were calculated using DTW, and visualised using UMAP. Colours are local to each figure and represent clusters identified through GMM clustering, where the highest silhouette score determined the number of clusters. Three major OTUs (*Bacilli*, *Clostridia*, and *Actinobacteria*) found in the La Rosa et al. [2014] dataset are also found to be major OTUs in this dataset. The absence of *Gammaproteobacteria* could again be explained by the aerobic environment of the throat cavity. While *Bacilli*, *Betaproteobacteria* and *Fusobacteria* show clear separation, other OTUs show gradual changes in TVAP. **The axes in the cluster plots are: UMAP Component 1 (x) and UMAP Component 2 (y).**

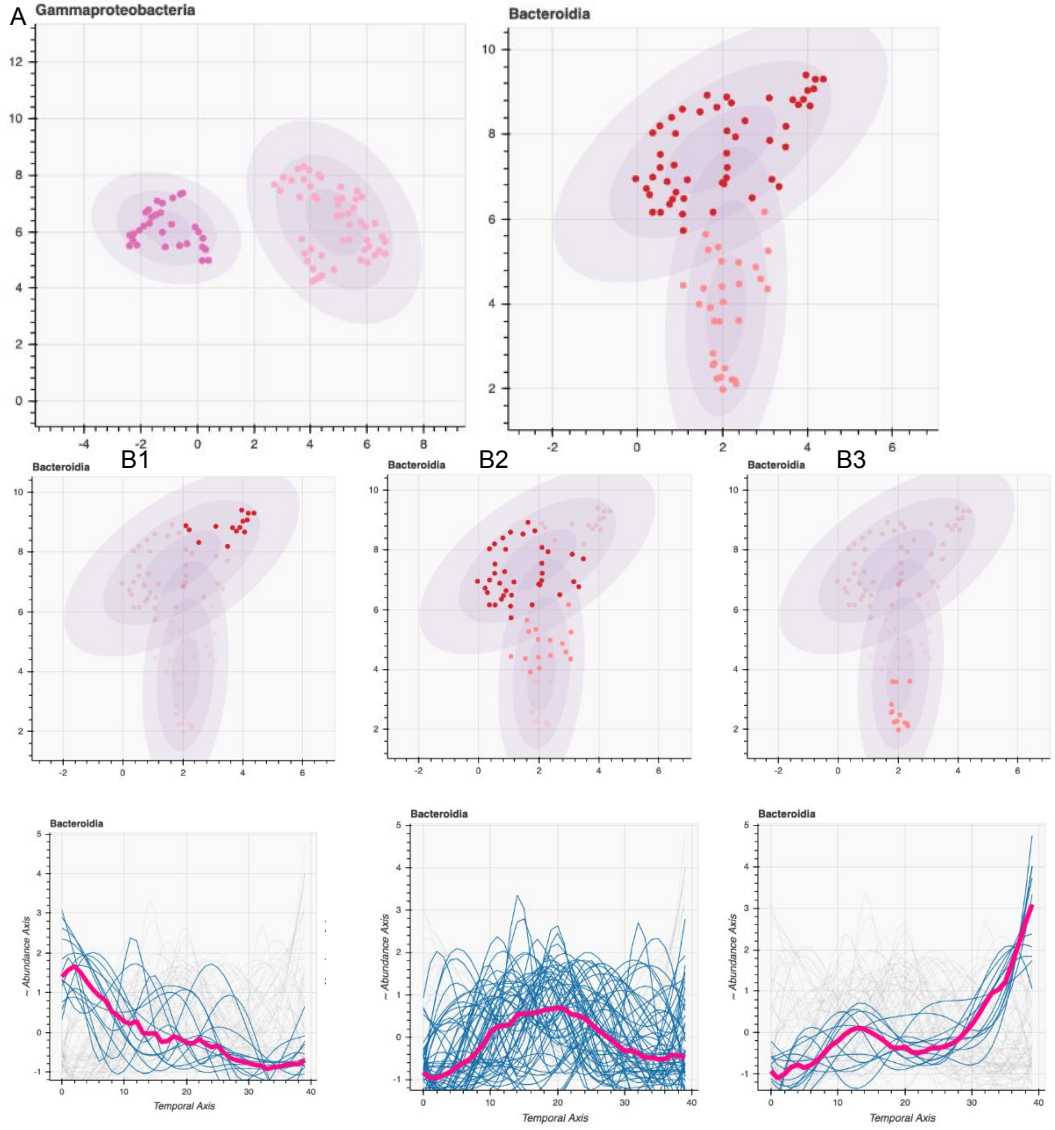

Figure 5.9: The figure shows clustering of the major OTUs in infants' gut microbiome from the Grier et al. [2018] study. Distances between TVAP were calculated using DTW and visualised using UMAP. Colours are local to each figure and represent clusters identified through Gaussian Mixture Models (GMM) clustering, where the highest silhouette score determined the number of clusters. The TVAP patterns of *Gammaproteobacteria* show separated two main clusters and subclusters within. However, the TVAP of *Bacteriodia* shows gradual change across the clusters, as shown in the subfigures B1 - B3. **The axes in the cluster plots are: UMAP Component 1 (x) and UMAP Component 2 (y).**

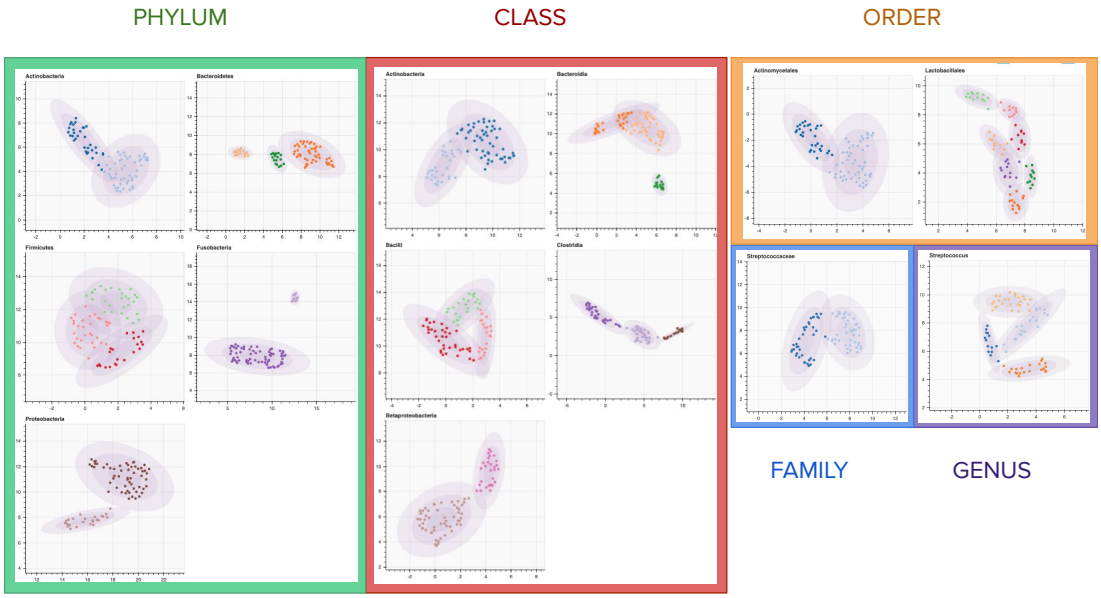

Figure 5.10: This figure shows the TVAP clustering of the infants' nasal microbiome from Grier et al. [2018] at various taxonomic levels. Colours are local to each figure and represent clusters identified through GMM clustering, where the highest silhouette score determined the number of clusters. The number of major OTUs reduces with the increased taxonomic resolution. However, at each resolution, we observe separated clusters. **The axes in the cluster plots are: UMAP Component 1 (x) and UMAP Component 2 (y).**

Streptococcaceae, and Genus *Streptococcus* which is featured in Figure 5.11. Also, at each taxonomic level, there are both disjoint and connected clusters present.

5.2.8 Major OTUs and Secondary OTUs

We mainly focused on the major OTUs, which were defined as the most abundant OTUs common to all the host environments. However, we are also interested in observing the nature of secondary OTUs' TVAP patterns. Figure 5.12 shows a data set from the throat microbiome again, but we have chosen the second most abundant taxa instead of the major ones. We notice that even the secondary OTUs show attributes we discussed above. We especially take note in *TM7-3* and *Flavobacteria* which show clearly disjoint clusters. Although non-major OTUs are discarded in some studies from the analysis [Shaw et al., 2016], we suggest that CoPR can successfully give meaningful visualisations for those.

5.2.9 Silhouette Scores and the Number of Clusters

In using any kind of clustering, deciding the optimal number of clusters is essential. We used the silhouette scores for this purpose. In Figure 5.7, we show the silhouette

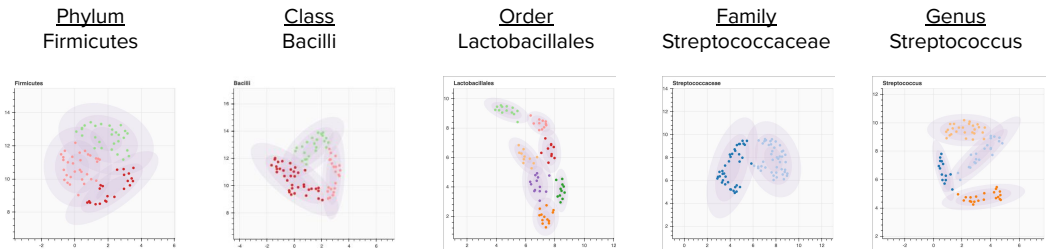

Figure 5.11: A selection along a branch of the taxonomic tree from Figure 5.10. We note change of the cluster structures across varying taxonomic levels. **The axes in the cluster plots are: UMAP Component 1 (x) and UMAP Component 2 (y).**

scores next to the cluster visualisations. We observe that silhouette score graphs can also bring invaluable information about the microbial community activity analysis. As an example, we will look at *Clostridia* and *Betaproteobacteria*. Each of these has clearly separated clusters, and clearly prominent peaks of silhouette scores at their respective optimal cluster numbers show that the clustering is robust. We also observe the standard deviation (error) bands at the respective optimal cluster numbers are small for these OTUs.

Another example is *Bacteroidia*, whose silhouette score peaks at three, but it also indicates that smaller cluster numbers can also provide "good enough" silhouette values. However, the standard deviation bands reconfirm that the clustering into three is the most robust and consistent option, regardless of the initial points selected for the clustering.

5.2.10 Simulated Data

After examining two real-life datasets, we look at a simulated dataset. The simulated dataset is created to approximate a known grouping with the clustering. Although ground-truth cluster labels are impossible to find in practice, we carry out the simulation to test the limits of CoPR in uncovering the known truths. In creating the simulated data, we faced several challenges. Foremost, it is currently impossible to create a dataset where the shape or the pattern of the TVAP is directly linked to the interaction parameters while preserving randomness. Hence we created multiple stencils for TVAP patterns and approximated them with known functions with randomness in parameters and noise. The experiment was designed to achieve the original stencil patterns as the median TVAP of each cluster.

The objective of the simulation was to approximate a typical longitudinal abundance dataset. The simulation has 100 subjects in total, with 20 OTUs in the microbial community. This number was much lower than what one would find in a typical microbial community. However, as most OTUs in a typical community are rare OTUs, we were satisfied with the simulation data generation process's lower number. From

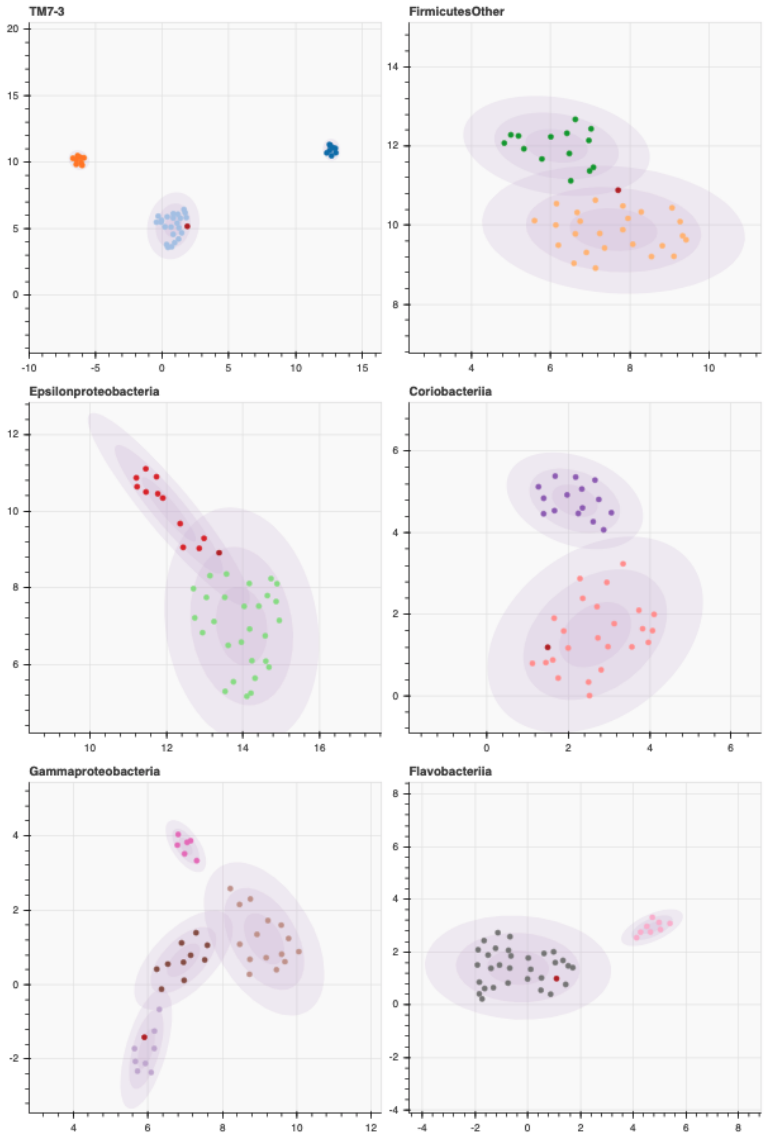

Figure 5.12: TVAP clustering of the secondary OTUs. Again, we are visualising the infant throat microbiome from Grier et al. [2018] at Family (L3) taxonomic level, but with non-major OTUs; Colours are local to each figure and represent clusters identified through GMM clustering, where the highest silhouette score determined the number of clusters. Secondary OTUs also show similar behavioural patterns to those of major OTUs. **The axes in the cluster plots are: UMAP Component 1 (x) and UMAP Component 2 (y).**

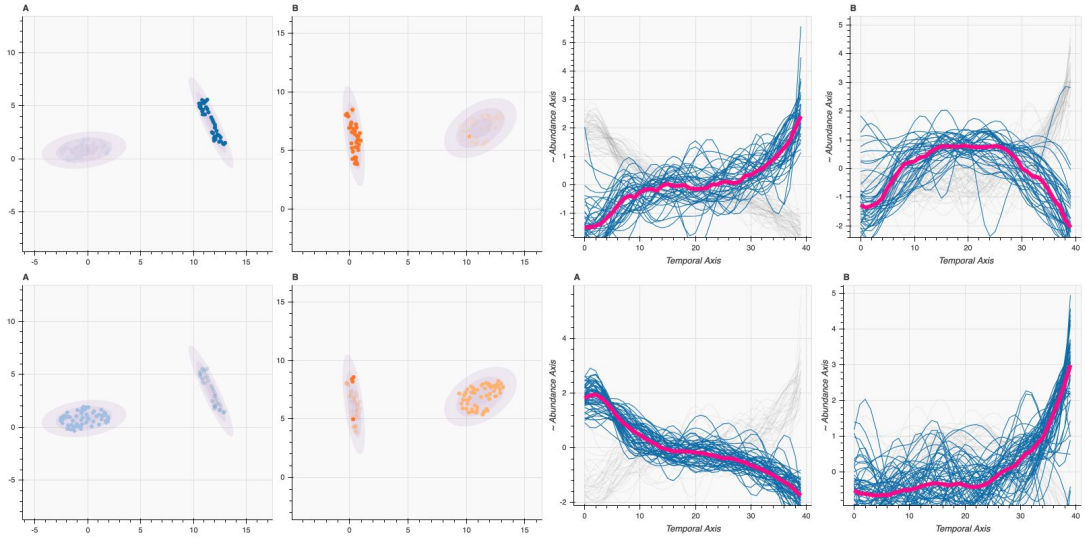

Figure 5.13: This figure shows the visualisation of the simulated OTUs A and B, with two corresponding clusters selected. The left-most half contains the cluster placement in the reduced dimension scatter plot, and the right-most half shows the TVAP of the selected clusters and the median for each cluster shown in thick pink. **The axes in the cluster plots are: UMAP Component 1 (x) and UMAP Component 2 (y).**

the 20 OTUs, four were considered to be major OTUs, while the other 16 were of secondary abundance. Approximately a major OTU's abundance was ten times that of a minor OTU. Out of the major OTUs, two—OTU-A and OTU-B—had two TVAP patterns each (Figure 5.13). Grouping of TVAP patterns in OTU-A and OTU-B corresponded to each other, except for randomly generated outliers. These outliers are highlighted in the TVAP cluster subplot of OTU-B corresponding to rising behaviour (Figure 5.13). OTU-C and OTU-D had three distinct behavioural patterns, which had no connection to each other or to that of OTU-A or OTU-B.

Table 5.1: Summary of the Figures of Chapter 5.

	Data	Summary of Observations	Identified Individualities	Identified Conformities
Figure 5.2	La Rosa et al. [2014] Gut (L3)	• Clear separation of clusters in all four OTUs. • Some sub-clusters observed.	• Existence of multiple clusters. • Observation of sub-clusters.	• Clustering behaviour in itself.
Figure 5.3	″	• TVAP patterns correspond to clusters	• Micro-variations within clusters	• Macro-view of clustered TVAP patterns
Figure 5.5	″	• Correspondence of clusters exists between OTUs	• Overlap is not 100%. Deviating individuals observed.	• Majority of the community show conforming behaviour.
Figure 5.6	Grier et al. [2018] Gut (L3)	*Similar to Figure 5.2*		
Figure 5.7	Grier et al. [2018] Nasal (L3)	• Number of clusters & silhouette values	*Silhouette value is a measurement of cohesiveness **and** separation*	
Figure 5.8	Grier et al. [2018] Throat (L3)	• Different structures of the clusters	• Heterogeneity in cluster structures	•
Figure 5.9	Grier et al. [2018] Gut (L3)	• Disjoint and joint clusters. • Gradation of TVAP.	• Changes in the TVAP.	• Observed changes are smooth.
Figure 5.10	Grier et al. [2018] Gut (L2–L6)	*Similar observations to the above—across different taxonomic resolutions*		
Figure 5.11	″	• Cluster structures change along a branch in the taxonomic tree	• Not all taxonomic levels show similar structures	• Some preserved structure across taxonomic resolutions
Figure 5.12	Grier et al. [2018] Throat (L3) Secondary OTUs	*Similar observations to the above—in the secondary OTUs*		
Figure 5.13	Simulated Data	• Simulated stencils re-identified through CoPR	• Planned individualities identified through CoPR	• Planned conformities identified through CoPR

5.3 Discussion

In this section, we further discuss the concepts of individuality and conformity, analyse the visualisations obtained by CoPR, and discuss the significance of the key findings.

5.3.1 Individuality versus Conformity

Individuality is a quality, trait or behaviour that separates something from the rest. In contrast, conformity is the opposite, where something's qualities, traits and behaviours are according to a set norm or standard and non-deviating from the rest. This concept has been applied in social science to describe human populations [Mughal, 2014; Wilson, 2009]. These concepts have emerged briefly in the literature of ecology [Hull, 1980] and made a comeback only recently in the field of microbiology [Montassier et al., 2018]. Inspired by these ideas, we would like to define the individuality and conformity of OTU activity in the scope of this work.

Let us define individuality as the tendency of the same OTU in similar environments to show different temporal variations in their abundance profiles. Conversely, let us define conformity as the tendency of the same OTU in similar environments to show similar variations in their abundance profiles. In the scope of this work, we define the traits of individuality and conformity to co-exist. We propose a fuzzy interpretation of the concepts, where each OTU community show a membership towards individuality and conformity."

Individuality and conformity are not phenomena limited to the microbial world. We observe this in human society, animal and plant kingdoms and many other natural and human-made systems. In most of these systems, we encounter generalisations to be helpful to an extent. However, generalisations have to be considered, coupled with the correct assumptions of circumstance. While generalisations are helpful, generalisation beyond reasonable assumptions is not. As an example, take plant care. While it can be assumed that plants of the same species need similar care in most cases, there may be individual plants that require a different kind of approach. We hypothesise that reasoning similar to this is valid for the microbial world as well. When we develop a generalised model for microbial community dynamics, it is essential that we are aware of the singularities of each community. Our visualisations provide qualitative insight into this balance.

Identifying these common tendencies or conformities will help us build better models to simulate microbial communities in general. They will help us understand better the links between different OTU communities and different types of host environments. After identifying the conformities, we can also identify the individualities for further analysis. Together with the conformities, the individualities give us specialised information about the specific issues related to a single OTU community. Information on both levels will assist us in obtaining a more practical idea of the microbial community dynamics.

Firstly, we simply do not see everything gathered in a single cluster in the reduced

dimension scatter plots. If that were the case, it would mean every community of the same OTU behaves in the same way. The OTUs being scattered around signify that there is no set norm for OTU communities' activity in a specific host environment—it would be incorrect to assume that, for example, TVAP of *Clostridia* will always show a particular tendency in a human gut environment.

Conversely, we do observe clusters rather than completely scattered points in the reduced dimension scatter plots. This observation means that subsets of OTU communities do show similar behaviour. When an OTU shows multiple prominent clusters, we consider that there may be multiple likely ways for this particular OTU communities to behave in this particular host environment. Hence, the existence of clusters is a degree of conformity we observe in the microbial communities.

We further hypothesise a connection between individualistic traits and external factors. Each host environment has specific environmental, clinical or other external factors. These external factors certainly affect the microbiome and its dynamics. The concepts of individuality and conformity may very well indicate the communities' reaction to their environment. More individuality than conformity in a particular microbial community's behaviour may indicate the community's sensitivity to the external environment. Although CoPR visualisation does not grant knowledge qualified to make a statement about the causality or correlation of the clinical factors to the microbial dynamics, we can conclude that these correlations can be identified.

We summarise the observations from the results, together with the identified individualities and conformities in Table 5.1. In this table, we detail how specific observations can indicate individuality, and others, conformity. We would like to draw attention to concurrent observations that suggest both individualist traits and conformist traits.

5.3.2 Visualisation

In this subsection, we will discuss the visualisations available from the analysis.

5.3.2.1 GMM Clusters

The main visualisation of the analysis is the Gaussian Mixed Model (GMM) Clustering. GMM is capable of identifying non-circular clusters as well. The EM algorithm that determines the cluster membership results in clusters of TVAP suitable for further interpretation in the context of microbiology. Each OTU is plotted separately, with each host environment represented by a dot plotted in the UMAP reduced dimensions. The UMAP dimension reduction is based on the Dynamic Time Warp distance between the TVAP curves. The plots also indicate cluster membership at three different membership probabilities. This visualisation primarily gives us information on how many distinct TVAP patterns could be identified for the specific OTU and how separated they are from each other. Secondarily we can identify cluster placement and shapes, which can provide us with information about the microbial dynamics.

5.3.2.2 Median Plot

The median plot of the TVAPs is straightforward in terms of mathematics associated but is very useful in comparing and contrasting the temporal variation across different clusters. The median line also helps us visualise the deviation of the member OTU's patterns from the typical pattern. We may pass on the median TVAP lines from all the OTUs to a microbial interaction inference algorithm to obtain interaction parameters.

5.3.2.3 Silhouette Index

Silhouette Index, plotted against the number of clusters, provides insight into the nature of the clusters. The silhouette index provides a quantitative measure of the clusters' cohesion and separation, reflecting the microbial communities' individuality and conformity. This quantification complements the qualitative idea we gather from the cluster plots. The error bars in the silhouette index plots indicate how consistent the clustering is.

5.3.2.4 Jaccard and Overlap Indices

Jaccard and Overlap indices are suitable metrics to confirm the visual from the cluster plots as they are quantitative measures of the agreement between the clusters of different OTUs. These plots complement the cluster plots.

5.3.3 Assumptions Involved

In any method involving microbial dynamics, there are several assumptions involved. These assumptions are due to the unobservable nature of the microbial dynamics and the poor understanding of various internal and external factors that affect the microbial communities. We also take the liberty to involve some assumptions in our application pipeline.

The principal assumption is that the studies conducted have sampled the microbiome within a time duration of high interest. For example, La Rosa et al. [2014] are interested in looking purely at the infant gut microbiome from birth to when they are ready to be discharged from the ICU. We assume that the scientific interest purely lies in the period between the starting event (sampling point) and the ending event (sampling point) and not in the actual clock/calendar time. As a result of this, our method analyses the TVAP patterns within the time duration of interest. We pass the burden of responsibility to users to use data captured within a duration of interest.

Some of the auxiliary assumptions are:

- that the TVAP is uniform between the sampling points;
- that the microbial variation patterns are independent of/minimally affected by external influences;

- that different host environments' microbial communities may have time lags and slower or faster dynamics;

- that OTU TVAP patterns are meaningful when considered independent of the other OTUs.

However, we do not involve some assumptions usually taken with microbial community dynamics analysis, including the assumption of a particular pattern for an OTU behaviour.

5.3.4 Knowledge from precision medicine

Precision medicine is where an individual's specialised needs are considered before prescribing the medicine. Although this seems trivial, not all individualities are considered in medicine, especially when microbial individualities are concerned. There is a growing interest in gut treatments to consider the composition and dynamics of an individual's gut microbial ecosystem before prescribing medicine. Furthermore, research suggests that other ailments also could be treated through the gut microbiome targeted precision medicine therapy [Cammarota et al., 2020]. Use of the same antibiotics and probiotics as a generalised therapy is undesirable if the patients' gut microbial compositions are entirely different, as we can reasonably expect that the reaction to anti/probiotics would vary for different microbial communities. Hence, an analysis that exposes the generalisation level applicable to individuals is a requirement for precision medicine.

5.3.5 Intra-cluster variations / sub-clusters

In some cluster configurations, we also observe sub-clusters. This observation, we propose to indicate that even in the apparent intra-cluster conformity, some individualistic traits prevail. These subclusters, especially when visualised as abundance variation patterns, can show us minute idiosyncrasies of the microbial dynamics. We believe that we would explain these subclusters in the future with enough clinical and environmental information.

5.3.6 *Gammaproteobacteria* & *Betaproteobacteria* Clusters

We identified *Gammaproteobacteria* as a major OTU in the gut and a secondary OTU in the respiratory microbiome. *Betaproteobacteria* was identified as a major OTU in the respiratory microbiome. The behaviour of the *Proteobacteria* in the visualisation was intriguing. Summarily, they almost always showed clearly separated and tight TVAP pattern clusters. Both *Gammaproteobacteria* and *Betaproteobacteria* were often the most clearly separated in many datasets, which was consistent in the datasets we examined.

5.3.7 Separation of Clusters at Different Taxonomic Resolutions

We observe the separation of clusters and indications of individuality and conformity at all taxonomic resolutions. We propose that this indicates that our visualisation is not necessarily a good technique only at the class level but also at other taxonomic levels. We observe the preserved structure in Figure 5.11 across the taxonomic hierarchy of Phylum *Firmicutes*, Class *Bacilli*, Order *Lactobacillales*, Family *Streptococcaceae*, and Genus *Streptococcus*. We propose that we interpret this preservation of structure as a trait (related to microbial dynamics) that is similarly observed in closely related microbial species.

5.3.8 Heterogeneity and Complexity

We argued earlier that time-series microbial datasets are complex and heterogeneous. The CoPR visualisations confirm that fact. We observe that the underlying structures of the microbial abundances are not homogeneous even in similar host environments. They are also different across different OTUs in the same host environment. We also see connections, such as cluster agreement which appear across OTUs, and across host environments. These observations reinforce our argument of the complexity and heterogeneity of microbial abundance data.

5.3.9 Distinctions from Other Collective Pattern Recognition Approaches

One of the main pitfalls we identify is the assumption that a typical TVAP pattern exists. When a method strives to achieve that typical pattern, it results in a loss of information. However, with our visualisation, we have shown that such a common trait does not exist, and that assumption should be invalid.

Secondly, another distinction in our approach is that it prevents the loss of individuality. Like other approaches, we also identify common patterns. However, that conformist approach is not at the expense of loss of individuality. The dominant pattern is preserved in the other approaches while forcing other patterns to transform into it [Lugo-Martinez et al., 2019]. While we agree that using the DTW distance can be considered a transform, its use is always pairwise—hence, it does not give prominence to a single TVAP pattern.

5.3.10 Future Work

We have identified several future research directions made possible through the CoPR visualisations.

Firstly, we are interested in exploring the subclusters and the intricacies involved in their separation. In the future, with more clinical and environmental data, we believe this could be quite an intriguing research direction to take.

Secondly, we can exploit the median TVAP curves in IMPARO to obtain MINs for each cluster combination. Because each OTU has multiple clusters, there would be multiple MINs inferred. However, this is in agreement with the idea presented in

IMPARO [Vidanaarachchi et al., 2020] that we cannot infer a single solution for MIN by examining NGS data. Hence, observations from the CoPR pipeline and observations from IMPARO directly complement each other. By identifying cluster overlaps that are statistically significant through the CoPR pipeline, we can improve the interpretation of MINs acquired through IMPARO. Isolation of overlapping clusters will help identify the behavioural/interaction patterns of OTUs in a homogeneous subset of environments, assisting in developing a global picture of OTU interactions where ambiguity is minimised. These findings may shine a light on separating environmental factors from MINs as well. We identify and this task as an exciting future research direction.

Thirdly we propose it would be interesting to observe and characterise the meaningful differences in the TVAP clustering patterns of major, secondary and rare OTUs. Especially if we observe different host environments, we might be able to find whether OTUs have different temporal dynamics when they are a major OTU or not.

Fourthly, we propose to investigate a connection between the notion of dynamic microbial interaction networks and CoPR, where we consider time-windows of a lengthy dataset (such as Caporaso et al. [2011]) to be a different host environment. Thus, we can apply collective pattern recognition techniques to a single abundance profile and analyse the temporal dynamics of microbial interactions through the clusters. Primarily, this could help us explore the repeated MINs we discussed in Chapter 4.

5.4 Conclusion

The TVAP patterns show that microbial community activity is heterogeneous and complex. We conclude that the behaviours of different OTUs across host environments vary and is best explored on a case by case basis. As per our discussion, there is a balance of conformity and individuality in the TVAP patterns. We propose that this behaviour can be an informative characterisation of OTU communities. We presented CoPR, a visualisation framework for collective pattern recognition for microbial data. Through unsupervised clustering of the data, our visualisation approach provided an exciting insight into the microbial communities. We believe this kind of analysis would be ideal for analysing new datasets. We also raise the question that the TVAP patterns may be connected with clinical factors. This question, however, remains to be fully answered in the future.

5.5 Methods

In this section, we first introduce the datasets we processed. Then we explain the pipeline in detail, including the techniques used and the terminology used throughout the paper.

5.5.1 Datasets

The following two datasets were used in our analysis.

5.5.1.1 Neonatal Infant Gut Microbial Dataset

The first dataset we process is an Infant Gut Microbial dataset collected by La Rosa et al. [2014]. This longitudinal dataset consists of 58 subjects, with an average of 16-time points each. Each subject is an infant in an intensive care unit. Stool samples were collected from each infant during their stay, and we have access to the abundance profiles generated through 16S rRNA sequencing and several clinical information about the infants, such as milk consumption, post-conception age, and delivery method. In our analysis of our dataset, we try to observe whether there is a connection between the clinical factors and the Temporal Variation of Abundance Profile (TVAP) patterns.

5.5.1.2 Infant Gut and Respiratory Microbial Dataset

The second dataset we look at in this chapter is another Infant Microbial dataset collected by Grier et al. [2018]. This longitudinal abundance data set has data from 82 infants, of whom 38 are pre-term and 44 are full term. We also have data from multiple body sites. Communities from the nasal, throat and gut microbiome are analysed, contrasted with and compared to each other.

5.5.1.3 Data Simulation

After testing the CoPR pipeline with real-life datasets, we used simulated data to verify our analysis. In this section, we will look at how I simulated the data. The dataset is created to approximate a known group with the clustering, as we identified correlations in the real-life datasets. I used a stencil-based approach, where a stencil is defined through a mathematical function. Twenty such stencils were defined. These included ten behavioural functions that defined major OTU behaviour after typical temporal behaviours examined in high abundant OTUs. A further ten reflected rare OTUs' temporal behaviour for minor/secondary OTUs. In simulating the data, each OTU population were probabilistically assigned stencils to follow. For example, each OTU–A's population had a probability of 0.4 to follow behaviour no. 1, 0.59 for behaviour no. 2, and 0.001 for behaviour nos. 11 to 20. TVAP were calculated according to these pre-defined behavioural stencils, to which uniform noise was added to complete the simulation.

5.5.2 Application Pipeline

In this subsection, we discuss the application pipeline we use in our work. The pipeline's input is microbial abundance data, and the output is the visualisation. Figure 5.14 illustrates the different parts of this pipeline. We then discuss each of the techniques we have used in the pipeline.

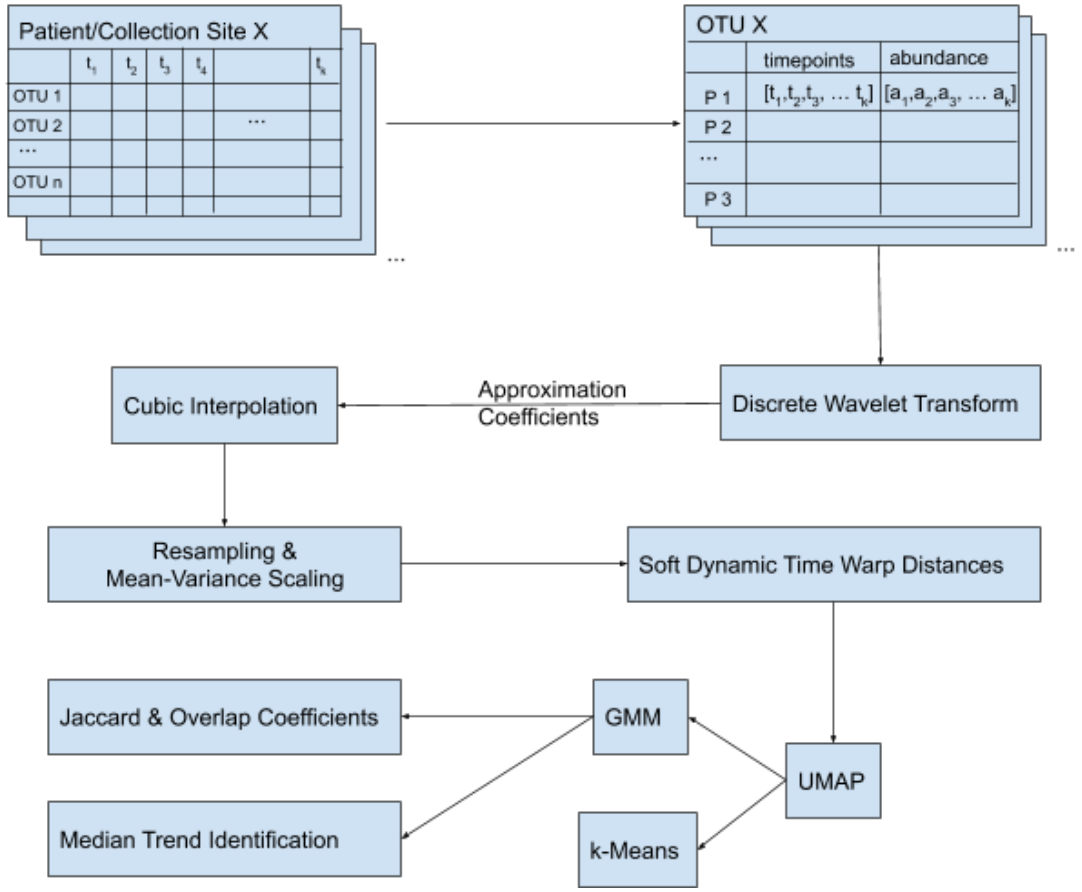

Figure 5.14: The first step is to convert the abundance data in a host-environment-wise into an OTU-wise configuration. For each host environment in the original data, we have an OTU abundance tabled against time points. We have a two-column table for each OTU in the new data where each row represents a patient/collection site. The two columns correspond to the sampling time points and the corresponding OTU abundance. The next step is to send the time-series abundance samples through a symmetrical Discrete Wavelet Transform (DWT). We then take the Approximation Coefficients and perform a cubic interpolation. After the interpolation, all the time-series are resampled to the same timeline and re-centred with respect to mean and variance. After these pre-processing steps, pair-wise soft dynamic time warp distances are calculated. These distances are used in place of the dimensions for the UMAP dimension reduction. The dimension reduced data is clustered with a Gaussian Mixed Model clustering, where the optimal cluster number has been identified through a silhouette score calculation. We also provide a k-Means clustering visualisation in the interest of comparison. As auxiliary visualisations, we provide the median trend for each cluster and selection and Jaccard and Overlap Coefficients of clusters across the major OTUs.

The first step is to convert the abundance data in a host-environment-wise configuration into an OTU-wise configuration. For each host environment in the original data, we have an OTU abundance tabled against time points. We have a two-column table for each OTU in the new data where each row represents a patient/collection site. The two columns correspond to the sampling time points and the corresponding OTU abundance. The next step is to send the time-series abundance samples through a symmetrical Discrete Wavelet Transform (DWT). We then take the Approximation Coefficients and perform a cubic interpolation. After the interpolation, all the time-series are resampled to the same timeline and re-centred with respect to mean and variance. After these pre-processing steps, pair-wise soft dynamic time warp distances are calculated. These distances are used in place of the dimensions for the UMAP dimension reduction. The dimension reduced data is clustered with a Gaussian Mixed Model clustering, where the optimal cluster number has been identified through a silhouette score calculation. We also provide a k-Means clustering visualisation in the interest of comparison. As auxiliary visualisations, we provide the median trend for each cluster and selection and Jaccard and Overlap Coefficients of clusters across the major OTUs.

5.5.2.1 Discrete Wavelet Transform

Discrete Wavelet Transform(DWT) is the first pre-processing step. DWTs conserve both frequency and location (time in temporal data) information; hence it is more suitable for our task than a Fourier Transform. We use a symmetrical kernel filter in our DWT step. A wavelet transform can act as a high-pass and low-pass filter. We use this quality to remove the noise and small fluctuations, to obtain the general trend we are interested in. Hence after the DWT, we discard the detailed coefficients and keep the approximation coefficients to represent the time-series data. The DWT also increases the frequency resolution of the data. We used the PyWavelets implementation of DWT in our application pipeline [Lee et al., 2019].

5.5.2.2 Cubic Interpolation

As the second pre-processing step, we use a 1D cubic interpolation. The interpolation aims to fill in the gaps between the sampling time points, as we require a continuous abundance variation pattern for comparison across host environments. Each host environment dataset is interpolated from its first time point to the last, with no extrapolation. Most other methods tend to use a spline interpolation; however, as our effort focuses on identifying an overall pattern, we consider a 1D interpolation to be more suitable in contrast with temporally localised patterns. We are assuming that the abundance pattern variation in between the sampling points is uniform.

5.5.2.3 Mean-Variance Scaling

As a third pre-processing step, we scale each interpolated time series to be centred around the mean, with a variance of 1 ($\mu = 0$, $\sigma = 1$). We aim to isolate each OTU

community's abundance pattern from the rest of the host environment by doing this preprocessing step. As we are merely interested in the temporal variation pattern of each OTU across multiple host environments, this allows direct comparison. The effect of this step is especially prominent in host environments, where there exist two dominant OTUs. This pre-processing step will hinder any quantification of microbial interactions. Hence, it is crucial to reiterate that we are not interested in the microbial interactions in this visualisation pipe-line, but rather the abundance variation pattern is our interest.

5.5.2.4 Resampling

The final step in our pre-processing approach is resampling. It is done to align the timelines of different host environments. We acknowledge that there are solid arguments for and against resampling the timelines across different datasets. The resampling will shift the timelines and change the temporal scale, which results in losing temporal information. However, when the sampling is done in a meaningful time scale, the resampling can help find a better overlap across samples. Hence the choice of resampling aligns with our choice of using dynamic time warp distance as well.

5.5.2.5 Dynamic Time Warp Distance

Dynamic Time Warp (DTW) Distance has been used in many time-series clustering-based methods in the literature. DTW is best explained as the distance between two-time series at their best temporal alignment. We seek a temporal alignment as different host environments could have delayed or temporally inconsistent behaviour, which can be identified to be based on similar variation patterns. By using the DTW distance, we can cluster similar variation patterns, despite temporal inconsistencies. We use the tslearn [Tavenard et al., 2020] implementation of the DTW distance in our application pipeline.

5.5.2.6 UMAP

Uniform Manifold Approximation and Projection (UMAP) [McInnes et al., 2018] is a manifold learning technique for dimension reduction. It is considered to have high visualisation quality and preserve more global structure than other dimension reduction techniques such as t-SNE [Van Der Maaten and Hinton, 2008]. We use UMAP to reduce the temporal dimensions and visualise each OTU's TVAP as a point in a 2-D plane. The neighbourhoods are determined by the DTW distance between each pair of datasets. The UMAP visualisation gives us an idea about the similarities and differences between time series data sets. We can observe that points that are clustered together correspond to similar TVAP patterns.

5.5.2.7 GMM Clustering

We cluster the data-points in the reduced dimension using a Gaussian Mixed Model (GMM) clustering. The number of clusters was determined by calculating the Silhouette Score for each cluster configuration. While we also consider k-Means clustering, GMM clustering results in superior identification of clusters. Because GMM considers the variation and the mean for its clustering, GMM more accurately identifies cluster membership of adjacent clusters of different sizes.

5.5.2.8 Silhouette Score

Silhouette score is a measure of how similar an object is to its own cluster and how different it is compared to the objects in other clusters (cohesion vs separation). The silhouette score graph we presented is the silhouette score as a function of the number of clusters. By examining the silhouette score graphs, we can understand how distinct the separation is and how similar the cohesion is at different cluster numbers. The higher the silhouette score is, the better the cohesion and separation. We calculate the silhouette score for the same number of clusters multiple times and take the average. This calculation also provides us with the standard deviation (shown with the error bar) for the silhouette scores. A narrower error bar means that the clustering is consistent at that number. A high silhouette score with a narrower error bar is the ideal cluster configuration we are looking for.

5.5.2.9 Overlap Coefficient & Jaccard Index

We use both the Overlap Coefficient (Szymkiewicz-Simpson Coefficient) (Equation 5.1) and the Jaccard Index (Equation 5.2) to examine the corresponding behaviour among the clusters. While the Jaccard Index indicates a bi-directional correspondence amongst two clusters, using the Overlap Coefficient allows us to identify unidirectional correspondence.

$$overlap(X, Y) = \frac{|X \cap Y|}{min(|X|, |Y|)} \tag{5.1}$$

$$Jaccard(X, Y) = \frac{|X \cap Y|}{|X \cup Y|} \tag{5.2}$$

5.5.2.10 Bokeh Visualisation Engine

Finally, we use the Bokeh Visualisation Engine [Bokeh Development Team, 2018] to provide interactivity to the visualisations, allowing a thorough manual exploration of the clusters and associated TVAP patterns.

5.5.3 Terminology

In this subsection, we will introduce some terminology which we have used throughout the chapter.

5.5.3.1 TVAP

Temporal Variation of abundance Profile. When we look at the abundance of a certain OTU as a function of time, we get the abundance profile's temporal variation. We have obtained a continuous graph by interpolating between the sampling points. This function, when visualised, will show certain tendencies. These tendencies are what we would regard to be identifiable patterns specific to each OTU community. As an example, we may notice rising abundance patterns, dropping abundance patterns, sudden peaks, etc. We argue in this paper that we can use these patterns to characterise an OTU community.

5.5.3.2 OTU community

For this chapter's scope, we characterise an OTU community as the organisms of a specific Operational Taxonomic Unit, which inhabit a particular host environment. To illustrate, we consider *Gammaproteobacteria* in a specific infant's gut microbiome as an OTU community. There could be several *Gammaproteobacteria* communities in the same infant, such as the gut *Gammaproteobacteria* community and the respiratory *Gammaproteobacteria* community. For the purpose of this paper, we consider them to be two separate and independent OTU communities.

5.5.3.3 Major OTU

In order to efficiently compute the CoPR analysis pipeline, we choose a subset of OTUs. In most analyses, the interest would be on the most abundant OTUs. Hence, consider the n highest abundant OTUs of each host environment in terms of average abundance. For this work's scope, let us define major OTUs as the intersection of the sets of n highly abundant OTUs in all the parallel host environments. Likewise, let us define secondary OTUs as the OTUs in the top $2n$, excluding the major OTUs.

The parameter n can be set according to the need of the analysis.

Chapter 6

Exploratory Study of Incremental Microbial Signatures

"Every reference-body has its own particular time"

—Albert Einstein,
Relativity: The Special and General Theory (1920)

This chapter is partially composed of material that appears in a manuscript titled "Incremental OTU Signatures" by Vidanaarachchi R., et al. that is being finalised for submission to a journal.

Summary

Background Microbial community dynamics may be different in similar but heterogeneous environments. We are interested in understanding and visualising these dynamics. However, being highly heterogeneous, it is hard to find a general representation of the temporal variation of microbial abundances. It has been argued that there is no unique signature that can represent the temporal variation of a microbial Operational Taxonomic Unit (OTU) community across multiple host environments. However, we still come across situations where a general representation is necessary.

Results We explore the use of incremental time microbial signatures. Our proposed signature pipeline provides a simple solution to the problem of a unified signature. We use time-warped distance metrics together with incremental dimensional reduction visualisation techniques to provide incremental signatures. We propound that these signatures are helpful to characterise the microbial dynamics in similar but heterogeneous environments.

Conclusions We successfully obtain unified incremental signatures for OTU communities in the human gut and respiratory environments. This exploratory study also highlights the importance of a signature that respects microbial communities' temporal and compositional heterogeneous nature.

6.1 Background

Understanding microbial life is crucial to our understanding of the world. Microbial community dynamics is an essential aspect of the said understanding. Throughout this thesis, we discussed our attempt to answer the question "What are they doing [in the microbial communities]" raised by Boon et al. [2013].

In Chapter 5, we discussed the use of collective pattern recognition techniques to enhance our understanding by augmenting available data. We also explored OTU communities' heterogeneous nature, with a balance of individuality and conformity existing in their temporal dynamics. We further illustrated how most earlier work on collective pattern recognition might have reduced an OTU community's temporal variation to a single pattern. However, we argued that this could lead to a loss of information, especially information pertaining to the non-prominent and micro behavioural patterns.

We argue that loss of information and generalisation of this nature is undesirable in characterising OTU community dynamics, leading to misinformed decisions and misrepresentation. Since microbial community dynamics are studied in sensitive settings such as clinical studies, we are wary of unnecessary generalisations. Hence, we propose a need for characterisation techniques that can represent the OTU communities as a whole, including the heterogeneous peculiarities—in terms of influence, abundance and temporal behaviour.

We draw inspiration from the study of microbial Community State Types (CSTs), utilised in many previous studies (Stewart et al. [2018a]; DiGiulio et al. [2015]; Grier et al. [2018]), which is a mode of characterising the microbial communities according to their static composition. The CSTs change across time as the compositional changes occur. Our interest also lies in characterising these communities, but with two distinctions. Firstly, we look at each OTU separately and examine its behaviour in multiple similar but heterogeneous environments (e.g. a collection of gut microbial environments, where not all guts are homogeneous). Secondly, we look at the temporal dynamics over the entire time duration of interest. Summarily, our characterisation is on the temporal dynamics, focused on a specific OTU.

To illustrate the utility of such a characterisation, let us consider the data-sets we explored in Chapter 5. We identified that *Gammaproteobacteria* community has potentially interesting behaviour in the neonatal infant gut (La Rosa et al. [2014] data-set). Further to this, we noted a correlation between the *Gammaproteobacteria* and the *Clostridia* community dynamics. We now propose two use cases. Firstly, we can compare the behaviour of *Gammaproteobacteria* with that of *Clostridia*. Secondly, we can compare *Gammaproteobacteria*'s behaviour in one set of environments to that

of another.

6.1.1 Related Work

Characterisation of the world around us has been instrumental in many fields in science. We explore the literature in two directions. Firstly, we will look into the various approaches of characterising the microbiome. Secondly, we will consider the characterisation of time-series data in general.

6.1.1.1 Characterising the Microbiome

Static microbial signatures have been used in many studies to identify the host environment or the state of the host environment uniquely. Desikan [2017] and Tridico et al. [2014] show instances where the humans could be identified through various bodily microbiome signatures. Banerjee et al. [2018] and Romero et al. [2014] recognise that various illnesses could be identified and categorised using the microbial signatures. In these works, microbial signatures refer to the static composition of a sample.

However, other studies have used dynamic signatures as well. Gerber et al. [2012] and Yang et al. [2019] identify instances where the dynamic microbial signatures have been successfully used to gain knowledge about the microbiome–host-environment relationships. Other studies propose the importance of creating customised signatures using knowledge of machine learning to pick important features of longitudinal microbial data [Knights et al., 2011].

Community State Types (CSTs) approach proposed in Ravel et al. [2011] has been used in many studies to characterise the microbiome. CSTs respect the heterogeneous nature of the microbial communities and have been used in time-varying settings.

6.1.1.2 Characterising Time-Series Data

Time-series data collection has a long history. Records from naval trade indicate that the importance of time-series data was known in the early days of human civilisation [Wilkinson et al., 2011]. In analysing data, different approaches of characterisation have been taken. Classical statistical methods were focused on finding linear relationships Zou et al. [2019]. With recent advances, using Bayesian models and feature selection approaches have gained more prominence. Particularly with multivariate data available in contemporary fields of study, which require non-linear modelling necessitates novel ways of characterising time-series data [Fulcher, 2017]. Considering the high-dimensional data we consider with longitudinal microbial abundance profiles, we have looked into dimensional reduction approaches for feature selection. Some of the suitable methods are parametric t-SNE [Van Der Maaten, 2009], parametric UMAP [Sainburg et al., 2020] and SONG [Senanayake et al., 2019].

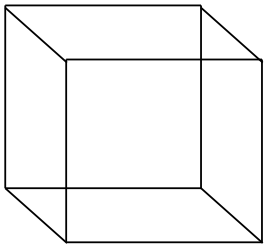

Figure 6.1: Representation of a 3D cube in 2D

6.1.2 Motivation and Contributions

From the literature, we can note the need to have a unified signature for comparison and other calculations. It is essential that this signature can be qualified both quantitatively and qualitatively. We note that in their work, Lugo-Martinez et al. [2019] uses pattern matched signatures in a Dynamic Bayesian Network to infer microbial interactions. Their signatures are successful in both visual perceptibility and quantifiability for further computational needs. However, we argue that pattern matching approaches result in a loss of heterogeneous information and micro-behavioural information.

While we acknowledge that our method is yet to be quantified successfully, and utility in further computations is limited, our focus is on information preservation—specifically the heterogeneous and micro-behaviours. Signatures generated by our method is suitable for visual and qualitative exploration of microbial communities.

We also note different dimension reduction practices from the literature and argue that some dimension reduction approaches are better suited for temporally dynamic data. We have successfully used dimension reduction in the previous chapter to represent the multi-dimensional temporal dynamics into an easily comprehensible two-dimensional plot. Our microbial signatures are also obtained by reducing the dimensions of multidimensional temporal variation of the abundance profile into two dimensions.

The significant challenge with our results is the ambiguity in the lower dimensional representations and lack of interpretability. However, we can provide an estimate in interpreting high dimensional data represented in a lower dimension. A simple example is a drawing of a 3D cube on a two-dimensional pane (Figure 6.1). Although the representation is in the two dimensions, we can perceive higher-dimensional information. Hence we propose that our lower-dimensional representations will retain a degree of interpretability in the correct contextualisation.

Dimensional reduction and averaging may appear to have similar effects. Hence, we would like to revisit the 3D cube example to clarify this. As we are interested in a 2D visualisation, if we are to average the 2D panes of the cube, we will obtain a nonsensical 2D representation. However, with the calculated 'dimensionally reduced' representation, we can still preserve information about the original structure.

Likewise, the microbial communities in heterogeneous environments we consider in this study cannot be successfully averaged—instead, we are tasked with devising a meaningful dimensionality reduction method.

Summarily, our motivations for the research based on the identified gaps are:

- A need for characterisation of temporally dynamic microbial community activity.
- Current characterisations may result in a loss of information and generalisation.
- Heterogeneity preserving visualisations through dimension reduction practices is a relatively less explored area.

In this chapter, we present an exploratory study of incremental microbial signatures and a preliminary application pipeline for heterogeneity preserving characterisations of the temporal dynamics of OTU communities. Ours is the first attempt to extract a signature to characterise OTU communities' temporal dynamics while respecting the microbial communities' heterogeneous temporal behaviours. We also discuss challenges associated with temporally dynamic microbial activity characterisation through a signature in terms of high-dimensionality, directionality, and interpretability. We present the signals generated for real-world gut microbial samples from two different studies and respiratory data sets.

Key contributions summarised:

- Characterisation of temporally dynamic OTU behaviour with reduced loss of information.
- Presentation of qualitative and visually comparable signatures.
- Discussion on challenges pertaining to high-dimensional temporal data visualisation.
- Obtaining and analysing signatures for real-life data.

6.2 Results

Before presenting the signatures we have obtained from various real-life microbial datasets, we explain how we obtain the signatures with a simple demonstration in Figure 6.2.

First, we used our application pipeline on real-life data from the gut microbiome in the Grier et al. [2018] dataset at the L3 Level. Out of these signatures, we have presented the signature of *Coriobacteriia* in Figure 6.3. We observe interesting behaviour in the figure and have included mark-up to clarify these visually on the signature itself.

Next, we overlapped the signatures from each of the major OTUs into the same plot, where we centred each signature's starting point at the origin. As individual

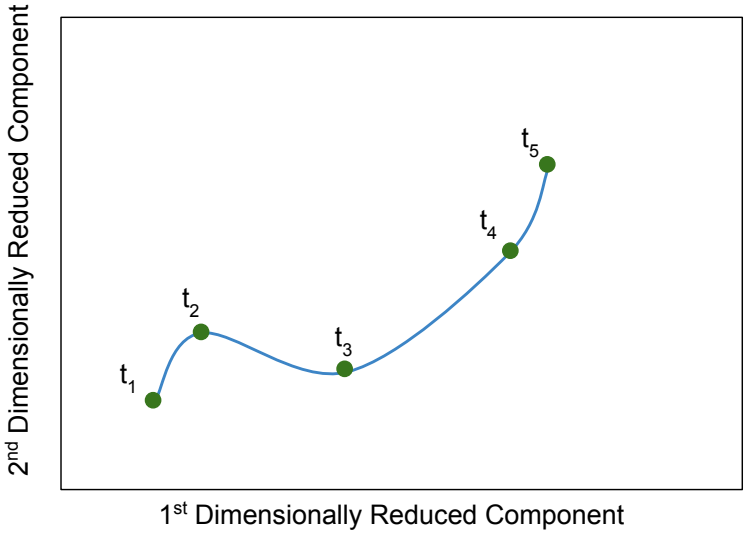

Figure 6.2: Explanation of incremental signatures. Each green dot is a dimensionally reduced representation of an OTU's abundance across a collection of similar but heterogeneous environments. These dots, marked sequentially as $t_1, t_2, ..., t_5$ were sequentially obtained through incremental dimensionality reduction techniques. The blue line connecting these dots was generated using a 1D Gaussian filter. As such, this line is representative of the change of composition of the OTU community across multiple environments. This line is our signature.

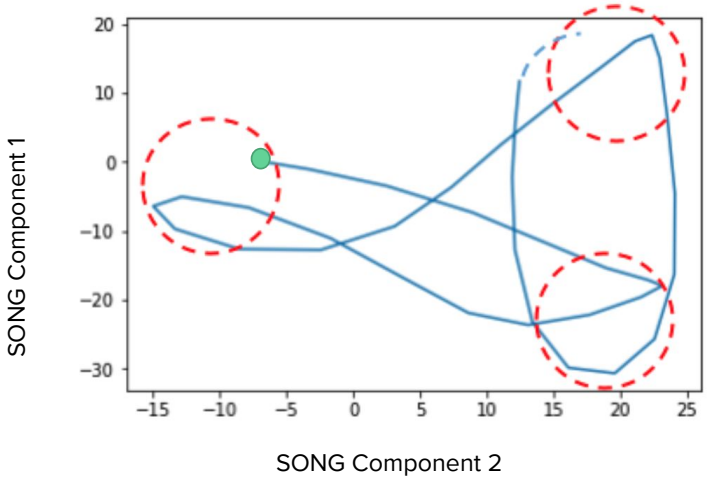

Figure 6.3: Signature of *Coriobacteriia* in a collection of 54 infant gut environments (Grier et al. [2018]). The signature starts at the green circle. Then, the signature goes to the lower right corner, and comes back towards the upper-left, then traverses to the upper-right, back to lower-right and again to the upper-right forming a loop. Although we are not yet armed with the complete knowledge on how to interpret these signatures, we can suggest that it is of interest to see returning behaviour. Namely, the regions of upper-left, lower-right and upper-right seem to be congregating areas. The continuous blue line is the signature. The broken blue line is the projection of the signature. Red circles are the regions of interest. The plot is a time series dimension reduction using the SONG (Self-Organising Nebulous Growths) algorithm, and as such, the axes are the first and second components of the dimension reduction.

signatures are independent of each other, this shift does not affect the information represented in the signatures. However, it facilitates the comparison between different signatures. These visualisations are presented in Figures 6.4, 6.5 and 6.6.

Then, we generated the signatures for the microbial communities from La Rosa et al. [2014] dataset at the L3 Level. The centred signatures are presented in Figure 6.7. We were further interested in visualising the signatures in a three-dimensional view—presented in Figure 6.8.

6.3 Discussion

In this section, we will discuss our results and the ideas behind the characterisation of time-series OTU behaviour. First, we discuss the data required for our methods and scenarios where our methods are applicable. Next, we will further discuss the results we presented in Section 6.2. We then discuss the signatures in terms of their temporal nature, heterogeneity, visual comprehensibility, dimensionality, and other qualities. Finally, we will discuss the future directions for this research and our methods' applicability in different contexts.

6.3.1 Contextualising the OTU Signatures

As an extension of the previous chapter's idea, we consider similar data in this chapter. Summarily, our data is a collection of time-series microbial abundance datasets that have been collected from similar host environments. However, this chapter's aim is significantly different from the previous one; rather than exploring the microbial dynamics, we are interested in characterising an OTU community's microbial dynamics over a collection of heterogeneous environments. As we explained in the background section, the interest in this stems from a need to provide a single representation akin to other methods, however, without losing heterogeneous information. I highlight that, as we cannot adequately separate the influence of the external factors across heterogeneous environments, a primary purpose in this characterisation is to compare OTU communities across distinct ecosystem or host environment groups.

6.3.2 Discussion of Results and Interpretability of Signatures

We are now faced with the challenge of interpreting the incremental signatures. As we discussed with the cube analogy earlier, with more contextual knowledge, we can better estimate the interpretation of certain aspects of the visualisation, which is the aim of this exploratory study. Some of the interesting features of signatures we can theorise are,

- the tendency of the trajectory to return to the vicinity of the origin;
- stretched versus compact trajectories;
- signatures showcasing (regular/irregular) looping behaviour;

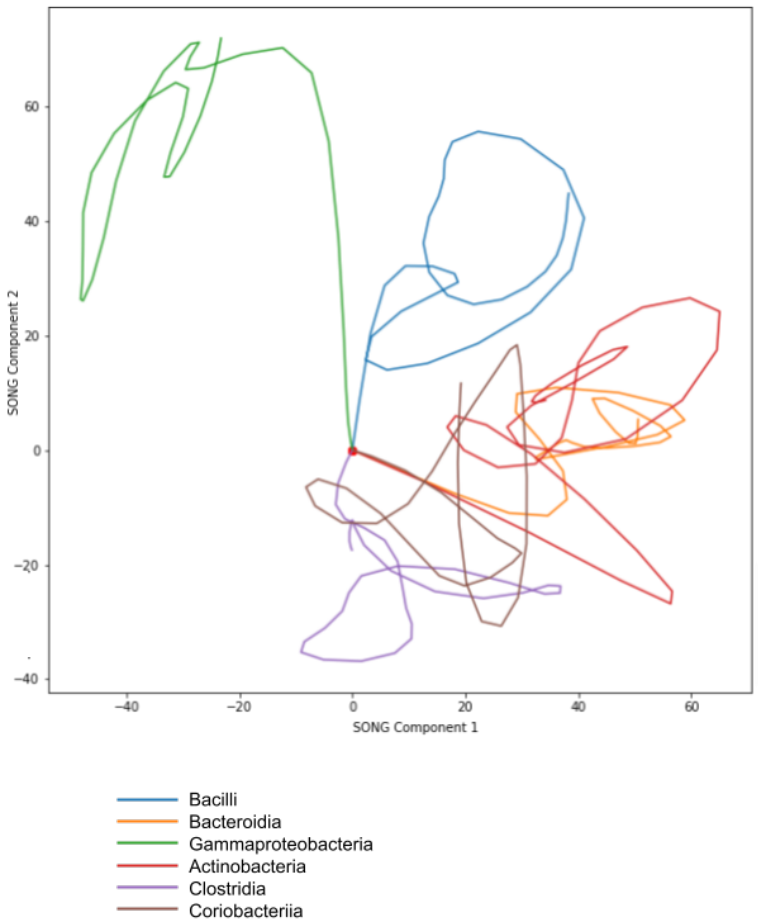

Figure 6.4: Signatures of the major OTUs at the Class (L3) taxonomic level in a collection of 82 infant gut environments (Grier et al. [2018]). The OTU classes represented are *Bacilli*, *Bacteroidia*, *Gammaproteobacteria*, *Actinobacteria*, *Clostridia* and *Coriobacteriia*. Each line's starting position is centred to the origin, as the positioning of OTUs is independent of each other in the reduced dimension. The colours are as described in the legend. We note the following observations that are of interest. *Gammaproteobacteria* and *Bacilli* tend to move in a similar direction, while *Actinobacteria*, *Coriobacteriia*, *Bacteriodia* move in another direction, leaving *Clostridia* in its own direction in the beginning. Arguably, *Bacteroidia* and *Actinobacteria* show roughly similar behaviour in their signatures throughout as well.

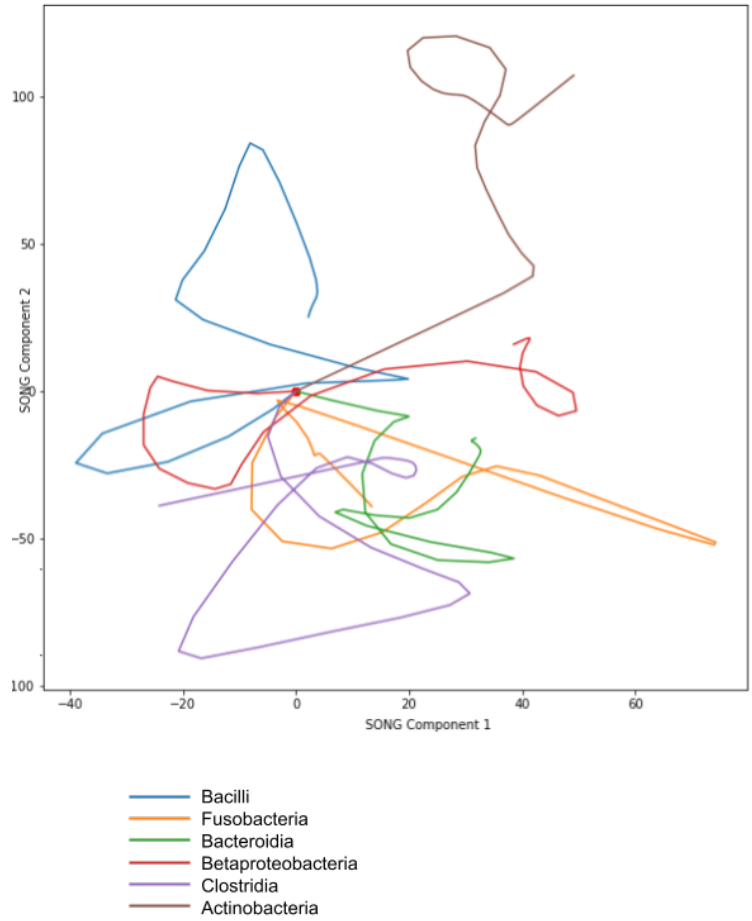

Figure 6.5: Signatures of the major OTUs at the Class (L3) taxonomic level in a collection of 82 infant throat environments (Grier et al. [2018]). The OTUs *Bacilli*, *Fusobacteria*, *Bacteroidia*, *Betaproteobacteria*, *Clostridia*, and *Actinobacteria* are shown here. Each line's starting position is centred to the origin, as the positioning of OTUs is independent of each other in the reduced dimension. The colours are as described in the legend. Again, in this figure, we can observe some OTUs moving in a similar direction at the beginning and continuing dispersedly later. We also note that 3 out of the 6 OTUs' signatures return to the origin's vicinity at certain points.

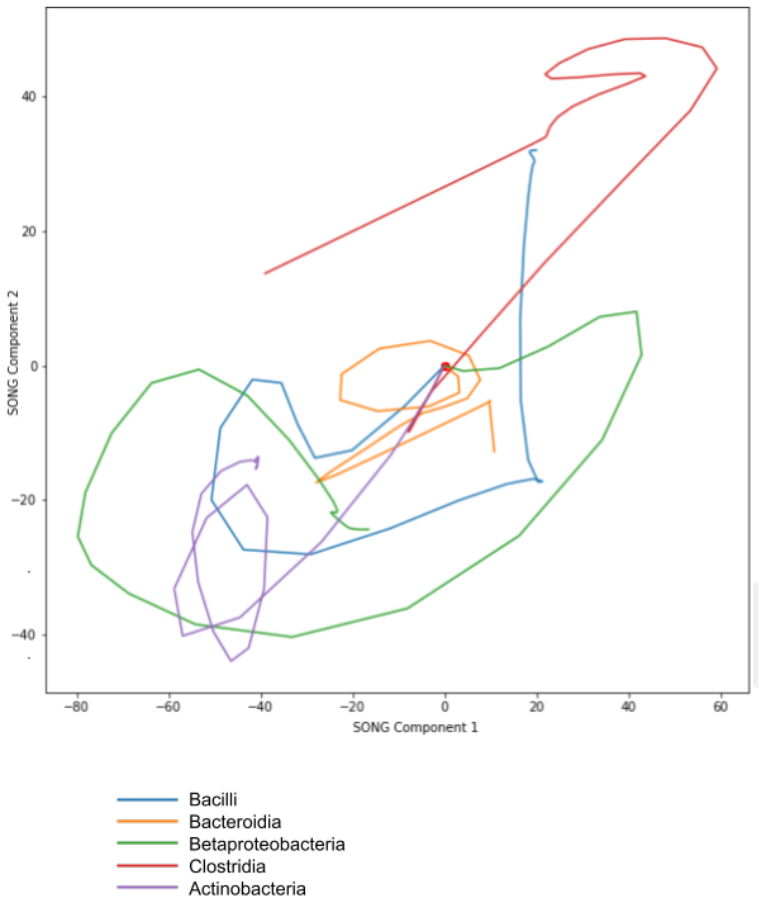

Figure 6.6: Signatures of the major OTUs at the Class (L3) taxonomic level in a collection of 82 infant nasal environments (Grier et al. [2018]). The OTUs *Bacilli*, *Bacteroidia*, *Betaproteobacteria*, *Clostridia*, and *Actinobacteria* are shown here. Each line's starting position is centred to the origin, as the positioning of OTUs is independent of each other in the reduced dimension. The colours are as described in the legend. The most interesting observation in this graph is the tendency of multiple signatures to form loops, significantly more than in the other graphs we have examined. Potentially this aligns with the idea of recurring dynamics, which we explored in Chapter 4. Apart from this, we notice that *Bacteroidia*'s signature stays around the origin while forming loops, while *Actinobacteria*'s signature forms a loop after travelling away from the origin. *Bacilli* and *Bacteroidia* also show a tendency of looping, albeit more stretched. Only *Clostridia* shows linear-like behaviour, though it has once returned to the origin.

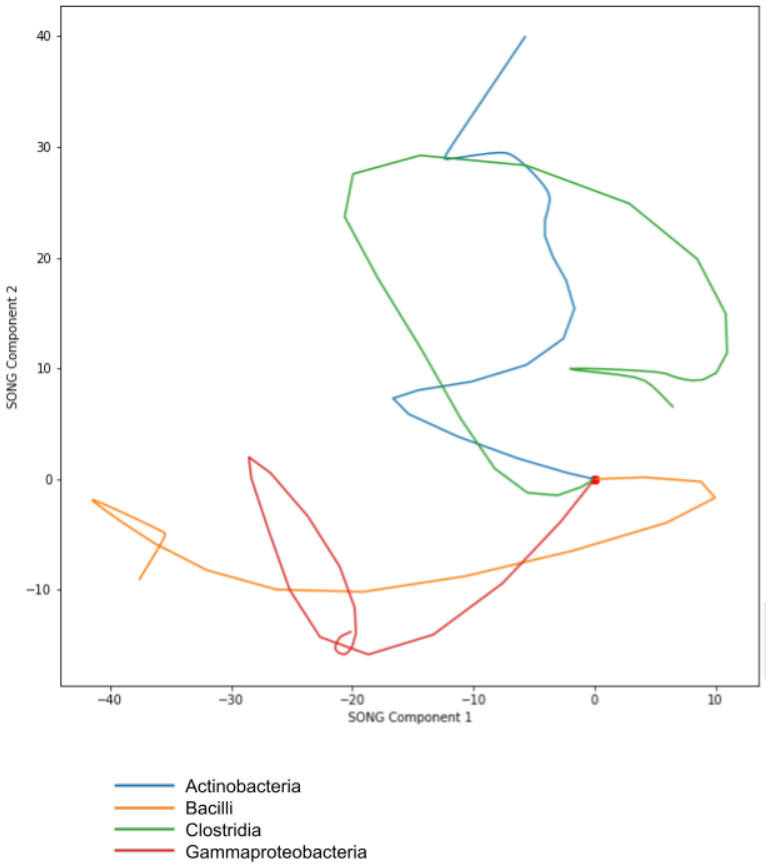

Figure 6.7: Signatures of the major OTUs at the Class (L3) taxonomic level in a collection of 54 infant gut environments during the infants' stay at a neonatal ICU [La Rosa et al., 2014]. The OTUs *Actinobacteria*, *Bacilli*, *Clostridia*, and *Gammaproteobacteria* are shown here. Each line's starting position in centred to the origin, as the position of OTUs is independent of each other in the reduced dimension. The colours are as described in the legend. In this figure we observe the *Actinobacteria* and *Clostridia* signatures showing a different tendency to those of *Bacilli* and *Gammaproteobacteria*. Also we see looping behaviour in *Clostridia* and *Gammaproteobacteria*.

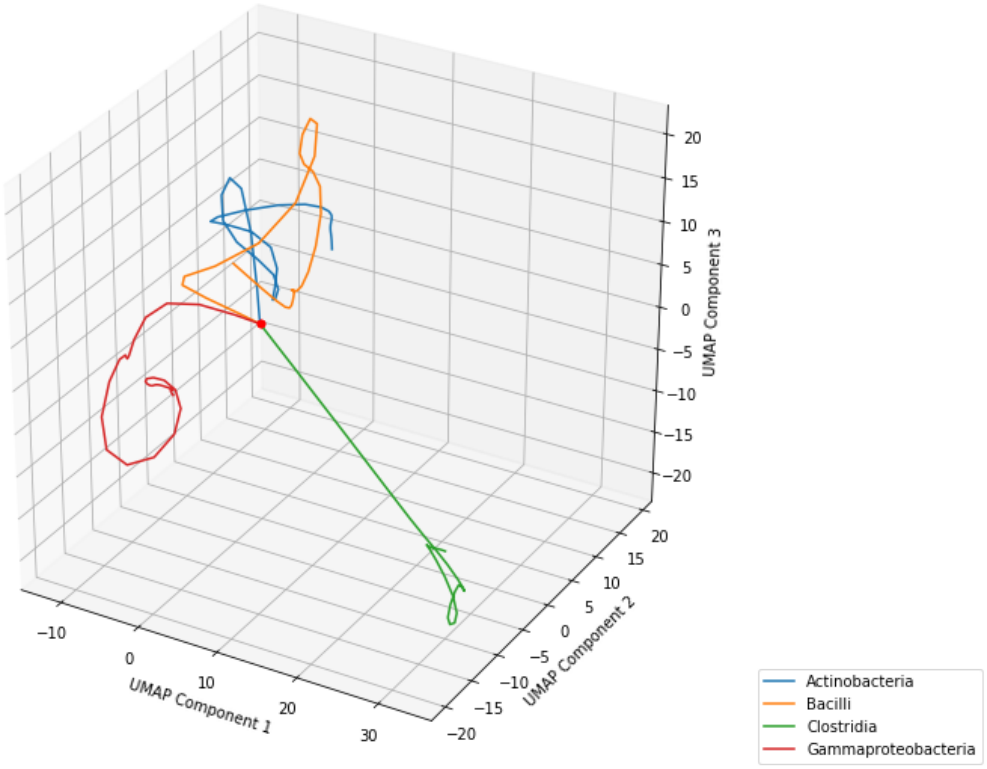

Figure 6.8: Signatures of the major OTUs at the Class (L3) taxonomic level in a collection of 54 infant gut environments during the infants' stay at a neonatal ICU [La Rosa et al., 2014]. The OTUs *Actinobacteria*, *Bacilli*, *Clostridia*, and *Gammaproteobacteria* are shown here. Each line's starting position is centred to the origin, as the position of OTUs is independent of each other in the reduced dimension. The colours are as described in the legend. We observe differences in this three-dimensional visualisation when compared to the two-dimensional visualisation. Specifically, we note the similarity in the signatures of *Gammaproteobacteria* and *Bacilli*, only at the beginning. However, we observe more looping behaviour in this visualisation.

- straight-lined trajectories versus curved trajectories;
- direction of the trajectory (in relation to each other);
- similar but scaled trajectories;
- trajectories that stay near the origin versus trajectories that drift away;
- overlapping trajectories; and
- smooth versus chaotic trajectories.

Some of these features are present in our visualisations, but, at this point in the study, we are unable to interpret these features of the signatures further. However, we hypothesise the features to be representative of certain qualities of the dataset.

Comprehension is arguably the most crucial aspect of any characterisation. Compared to existing methods, our method is not as easy to visually comprehend. However, it remains visually comparable. Nevertheless, we propound that pending visual interpretation, analytically, it better represents the OTU communities in all the host environments we consider.

6.3.3 Challenges in Characterisation due to Temporal Dynamics

Characterisation of OTU dynamics is challenging simply due to their temporal nature. Time-series data are directional, which we can define as ordered data containing probabilistic properties which depend on the direction of the time. In other words, they are irreversible. Hence, our characterisation methods should reflect the order and attempt to capture the directional probability. SONG [Senanayake et al., 2019] which we use for dimensionality reduction, is suitable for continuously growing (time-series) data.

The temporal dynamics of microbial communities are also complex. This complexity stems from various causes. The data is riddled with noise; the data collection is usually not uniform or consistent; the number of sampling points is different from dataset to dataset; the data is sparse; the data contains an underlying structure, mostly unknown to us, etc. Additionally, the structure of the data itself is assumed to be dynamic, resulting in the inability to accurately model the temporal dynamics in a system with dynamic interaction parameters, with a set of first-order differential equations such as the Lotka-Volterra equations (as discussed in Chapter 4). When we characterise data with these complex temporal dynamics in the form of a unified signature, the said complexity is reflected in it.

6.3.4 Quantitative and Qualitative Representations

Quantitative data is measurable and numerically informative. The signatures we generate are not quantitative. Qualitative data is descriptive and describes phenomena that are observable but not measurable. The signatures that we generate qualify as

qualitative; however, as we mentioned above, they are better visually comprehended comparatively than on their own.

Stemming from this, we discuss in Section 6.3.9 that there is a potential in further developing the characterisation into –

1. A qualitative interpretation where a visual indication of the qualities is better comprehensible.

2. A quantitative interpretation suitable for analytical methods for further processing.

6.3.5 Heterogeneity

Various forms of heterogeneity are present in temporal microbial abundance data. In this section, we discuss the ideas of heterogeneity and its manifestation in the OTU signatures.

6.3.5.1 Heterogeneous Composition

The first and the most apparent heterogeneity is that microbial communities consist of distinct taxa. We discussed in Chapter 3 that these taxa could be classified as high abundance (>1%), low abundance (<1% and >0.1%), and rare (<0.1%) depending on their average abundance. Furthermore, in considering the collation of datasets, we introduced the concept of major OTUs, which are prominently present in all such datasets we collate, coming from a similar host environment. These are some heterogeneous aspects that we can identify as the heterogeneity of composition.

6.3.5.2 Individuality versus Conformity

As we discussed in the previous chapter, the second heterogeneity we discuss in the microbial communities is that there is a balance between individuality and conformity in the temporal variation. This heterogeneity is usually omitted when considering a unified pattern or a signature in current literature. In our method, we do not simply represent the most prominent temporal abundance variation pattern as the signature of the OTU community. Our signatures are a representation of all the individually shown patterns with the temporal information intact.

6.3.5.3 Preservation of Representation

In a discussion of time-series OTU signatures, it is essential to consider whether a signature represents all the communities of a specific OTU. However, it should also characterise the whole collection of the OTU communities fairly, in a unified manner. We achieve this in our signature by obtaining the signature as a lower-dimensional mapping of a high-dimensional signature. In the higher dimensionality, our signature is simply the relative composition of each OTU mapped against each other

and time. The two-dimensional mapping is a neighbourhood preserving embedding extracted from a higher dimension.

6.3.5.4 Heterogeneity of Environments

As microbial data is acquired from heterogeneous environments, I believe a discussion on the signatures across heterogeneous environments would be helpful. To briefly compare the signatures obtained across two heterogeneous environments, we look at Figures 6.4 and 6.7. The first shows signatures of a collection of 82 infant gut environments from Grier et al. [2018]. The second shows signatures of a collection of 54 infant gut environments from La Rosa et al. [2014]. Both these communities share the major OTUs *Actinobacteria, Bacilli, Clostridia,* and *Gammaproteobacteria*. Out of the four OTUs' signatures, we can identify *Gammaproteobacteria*'s signatures to show similar characteristics. *Actinobacteria*'s signatures are marginally similar, while *Bacilli* and *Clostridia* signatures in the two datasets are starkly different. Again, I highlight that we do not have the necessary information to interpret the findings entirely. However, I propose that the findings show promise and carry valuable information.

6.3.6 OTU Signatures Compared to Community State Types

As we noted in Section 6.1.1, Community State Types (CSTs) is an idea that is closely related to our approach. In CST based analysis, both heterogeneity and temporal dynamics are considered. However, the determination of the CSTs itself is static as CST is based on the composition of a microbial community at a particular time point. Thus, CSTs change with time, resulting in them being a representation of heterogeneity and temporal dynamics. Similarly, when we consider multiple similar host environments, each environment can exist in a different CST at a given time point. Our signature, however, characterises a specific OTU community rather than the entire microbial community in the host environment. Furthermore, our signature captures the OTU community's temporal variation into a signature in the lower dimension. Hence, the two concepts' underlying idea differ, and these are distinct characterisations based on different aspects of the data.

6.3.7 Dimension Reduction

We achieve the unified signature through the dimension reduction using SONG [Senanayake et al., 2019]. However, we have ensured the following during the dimension reduction process. Firstly, the reduced dimension is representative of all the dimensions (different host environments). Hence it does not result in a complete loss of information. Secondly, we can visualise the change with time, preserve the temporal directionality, and retain the ability to visualise the signature as time-bound segments. However, we do lose the linearity of time. Our justification for this lies in considering that multiple host environments may have different notions of time and that time-warping is nevertheless included as a preprocessing step.

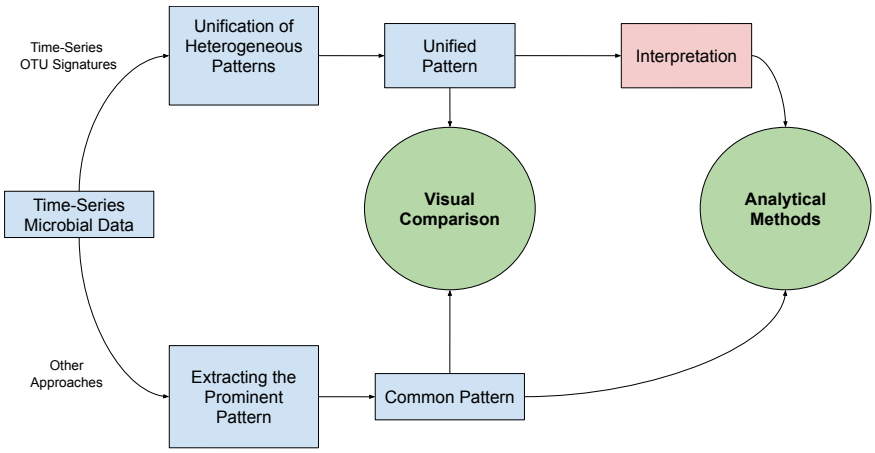

Figure 6.9: Time-series OTU Signatures (our method) contrasted with a generalisation of other approaches. Our method unifies heterogeneous information into a unified pattern, while the other methods obtain the most prominent pattern as a common pattern. While both are suitable for visual comparisons, signatures obtained through our method require an extra interpretation step before being used in analytical methods.

6.3.8 How are our signatures different?

The main difference between our signature constructing approach and other approaches is that ours provide a unified characterisation of an OTU's behaviour across multiple host environments. However, as shown in Figure 6.9, even though both are sufficient for visual comparison, ours will require an interpretation technique to be used directly in analytical methods such as interaction network inference approaches. The signature extracting approach is also different, as other methods focus on extracting in the prominent pattern, while ours focus on unifying all the different patterns observed in the dataset.

6.3.9 Future work

Following this exploratory work, we can identify several future work directions. We will briefly discuss some of the future research directions in this Sub Section.

6.3.9.1 Further Interpretations of Signatures

The main future research direction we could identify and whose importance is highlighted throughout our discussion is the further interpretation of the signatures. The features of interest we have identified in our discussion can be further explored. In improving the interpretability, controlling the degree of randomisation of the dimension reduction stage can also be helpful.

6.3.9.2 Actively Modified Signatures and Comparative Analysis

Another future research direction is to create a mechanism to modify the signatures. Let us illustrate with an example. As a signature represents multiple dimensions, say A is the set of dimensions representing the set of host environments. Let us consider B and C as disjoint subsets of A s.t. $A = B \cup C$ and $B \cap C = \emptyset$. Consider the signatures obtained for the dimensions of B and C, which are, in essence, split signatures of the original signature of A. Comparing the signatures of subsets of host environments would help analyse and represent the environmental factors' effect on microbial behavioural dynamics. Likewise, we can split the data along any axis. Splitting along the temporal axis—for example, following important events in the environment— will provide a comparison helpful in analysing event-specific changes in microbial community behaviour. Likewise, we can separate the OTU actors in order to obtain distinct groupings of signatures.

When the described subsets are the more homogeneous clusters as identified through CoPR (Chapter 5), we can identify a TVAP with lesser deviations. This TVAP can be used with IMPARO and its extensions to quantify microbial interactions. Conversely, the signatures generated for the subset is expected to be a general characterisation of the OTU in this homogeneous environment. This future research direction will be valuable to connect quantified interactions to unique characterisations provided through the incremental signatures.

6.3.9.3 Signatures for Other Applications

Lastly, we suggest that the idea of time-series signatures could be applied to other similar systems as well. Especially gene expression data are closely related to microbial abundance data, and there are existing explorations into the time-variant nature of the gene interactions [Song et al., 2009a]. Macrobial ecosystems and their analysis could also benefit from time-series signatures. For a non-biological application, we had already discussed the similarities of microbial abundance datasets to stock market data [Tan et al., 2016]. With fast-changing temporal behaviour and a plethora of data available, stock markets would be an exciting system to generate time-series signatures.

6.4 Conclusions

The signatures derived from our application pipeline characterise the temporal dynamics of a specific OTU community across several host environments. It is superior to other representations in terms of inclusiveness and information preservation and is complementary to the information provided in different methods. The signatures are possibly informative as a single signature. However, they can be used subjectively in visual comparison with other signatures across taxa and host environment groups. The application to real-life data sets shows the signature's effectiveness in bringing out underlying structural information about the microbial community dy-

namics. There needs to be more work done in quantifying the information contained in the signature, which would be valuable to be used in further computational tasks, and in objectively interpreting the signature, which is currently complicated due to the dimension reduction approach used in its generation. Overall, we believe this is a helpful tool in comparative studies of microbial communities and their temporally dynamic behaviour.

6.5 Methods

In this section, we first introduce the data sets we processed. Then we explain the pipeline in detail, including the specific techniques used in the pipeline and the terminology used throughout the paper.

6.5.1 Datasets

We use the same datasets used in Chapter 5 for this analysis. However, a brief description is provided below for the sake of completeness.

6.5.1.1 Neonatal Infant Gut Microbial Dataset (La Rosa et al. [2014])

The first dataset we process is the La Rosa et al. [2014] dataset. This longitudinal microbial dataset consists of 58 subjects, with an average of 16-time points each. Each subject is an infant in an Intensive Care Unit (ICU). Stool samples were collected from each infant during their stay in the ICU, and we have access to the abundance profiles generated through 16S rRNA sequencing and several clinical information about the infants, such as milk consumption, post-conception age, and delivery method.

6.5.1.2 Infant Gut and Respiratory Microbial Dataset (Grier et al. [2018])

The second dataset we look at in this chapter is the Grier et al. [2018] data set. This longitudinal abundance data set has data from 82 infants, of whom 38 are pre-term and 44 are full term. We also have data from multiple body sites, namely the respiratory tract (nose and throat) and the gut.

6.5.2 Application Pipeline

As this chapter is an extension of the ideas presented in CoPR (Chapter 5), we use the first part of the CoPR pipeline as is. The latter part is distinct, functioning specifically for signature generation. After the signatures are obtained, we have centred the starting points of each signature to the origin by performing a simple coordinate shift.

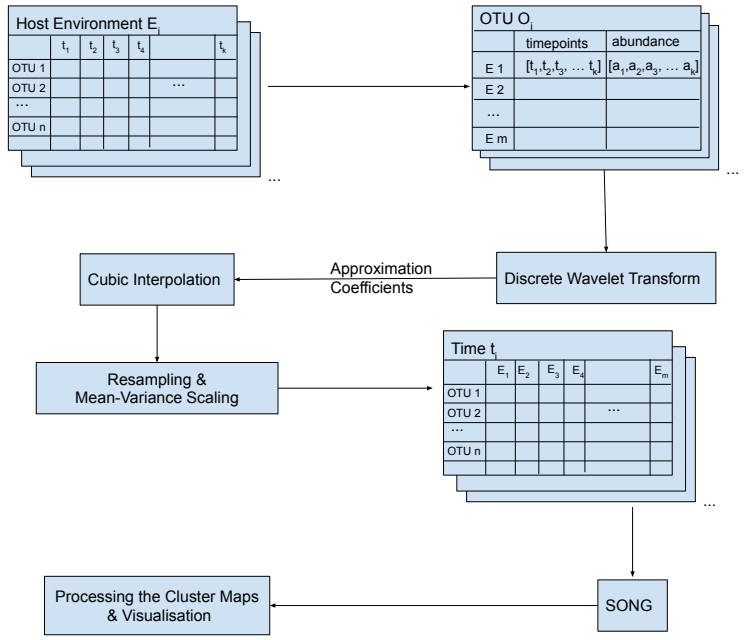

Figure 6.10: Application pipeline for generating signatures. The data collected in the form of OTU abundance tables are tabulated as one table per one host environment, with the abundance of each OTU against the sample collection timepoint is transformed to a per OTU format. These are then pre-processed with a discrete wavelet transform, whose approximation coefficients are forwarded into a cubic interpolation step. Lastly, this data is scaled with a mean-variance correction and resampled into a common timeline. At this point of the CoPR pipeline, we re-tabulate the data with one table per time point. Each OTU's abundance at that particular time point is tabulated against the host environment in each table. In summary, for each row of data representative of an OTU, there are m dimensions, where m is the number of host environments included in the dataset. This is then sent into the SONG algorithm to be reduced dimension-wise. After the dimension reduction, the cluster maps are analysed to identify the movement of the data points. These paths are the incremental signatures in the visualisation.

6.5.2.1 CoPR Pipeline

As we have discussed the CoPR pipeline methods in great detail in the previous chapter, we will not repeat the details here. However, a summary of the CoPR pipeline is provided with Figure 6.10.

6.5.2.2 Self Organising Nebulous Growths (SONG)

Self Organising Nebulous Growths [Senanayake et al., 2019] is a non-parametric dimension reduction algorithm that supports mapping continuously growing data in reduced dimensions. It preserves the structure of the data as the data grows, which is perfect for our use case of time-series microbial data coming from multiple datasets. SONG is also better suited (compared to t-SNE [Van Der Maaten and Hinton, 2008] and UMAP [McInnes et al., 2018]) for datasets with high variance and noise; which is also vital for microbial data.

Chapter 7

Conclusion

In this thesis, I sought answers to the question of "What are they [microbial OTUs] doing [in their communities]?" presented by Boon et al. [2013]. Specifically, I have applied data analysis methods based on systems biology's foundations to analyse longitudinal microbial abundance datasets. I propose that my methods and discussions shine a new light on the process of analysing longitudinal microbial datasets to understand microbial community dynamics.

My research was focused on analysing complex, heterogeneous and time-series data. From a systems biology point of view, my research focused on modelling microbial communities' temporal dynamics.

Single Microbial Environments In the initial part of the thesis, I analysed microbial abundance data pertaining to a specific environment.

Inference of Static Microbial Interactions Firstly, in Chapter 3, I looked at the question of inferring static microbial interaction networks through analysing microbial abundance profiles from a single environment. I aimed to improve model-based approaches in parametrising microbial interactions. I treated this as a large parameter optimisation problem and introduced IMPARO, a genetic algorithm-based solution. My attempt was successful in obtaining quantitatively improved results. Further to this, I argued for the existence of multiple solutions given the nature of this optimisation problem. I used systems biology principles to fine-tune the genetic algorithm resulting in quantitatively and qualitatively improved microbial interaction network parameters.

Temporal Dynamics of Microbial Interactions Secondly, in Chapter 4, I explored the temporally dynamic nature of the microbial interaction networks themselves. I base my argument on the basis that in a complex ecological dynamics model, the static nature of the pairwise interactions is the assumption. I challenge this assumption based on the improbability for interspecies interactions remaining static in complex environments where everything else is dynamic. Here, I critically evaluated the use of Lotka-Volterra equations for modelling microbial community ecosystems. Exploring further, I discussed the necessity of using a system of higher-order differ-

ential equations to model the microbial interactions successfully. I presented proof for the interactions' dynamic nature and argued that patterns are recognisable in the interaction parameters themselves.

Heterogeneous Collections of Similar Environments In the latter part of the thesis, I capitalised on the availability of data collections from similar environments to overcome the challenges pertaining to the lack of data that I identified earlier. However, the aim of this part was not to infer the interactions quantitatively but to qualitatively characterise the temporal behaviour of OTUs.

Collective Pattern Recognition Thirdly, in Chapter 5, I discussed the practical issues in inferring microbial interactions through the limited availability of longitudinal abundance profiles. I highlighted that longitudinal abundance profiles are usually not numerous enough and that their sampling frequencies are usually low and non-consistent to achieve a reasonable quality inference. I explored the idea behind collective pattern recognition as a solution to this problem. I highlighted the importance of considering the heterogeneity of the OTUs and finding a balance between their individuality and conformity. I also presented a novel visualisation tool that is capable of successfully identifying TVAP patterns. I argued that it would provide invaluable insight into any collection of longitudinal microbial datasets.

Incremental Signatures for Microbial Community Activity Fourthly, in Chapter 6, I explored the fascinating idea of characterising high-dimensional temporal microbial data with unified low-dimensional signatures. In this exploration, I contrasted the pros and cons of existing methods of obtaining signatures with dimensionally reduced signatures. I also explained the challenges and prospects of interpreting the signatures and the ability to be used in analytical pipelines. This chapter was an exploration which tested the proverbial water and provided insights to future study.

In this thesis, I have attempted to shed light on the myriad of techniques available to us to better understand what [microbial organisms] are doing [in microbial communities]. I have approached this problem through two facades. Firstly, through inferring interactions within a microbial community, I focussed on quantifying the driving factors behind the temporal variations. I also discussed how microbial interactions are better modelled as time-varying parameters. Secondly, I looked at collections of bacterial communities and explored temporal variation of individual microbial OTUs. This second approach included visualisations that qualitatively informed patterns of microbial behaviour as well as an exploration into unique characterisations of temporal microbial behaviour. To summarise the whole idea behind these two broader sections, I would like to revisit the thesis's title - "Pattern Recognition

for Complex Heterogeneous Time-Series Data: An Analysis of Microbial Community Dynamics". My research was an exercise in understanding a complex system of heterogeneous components with temporal dynamicity. I identified that the second question by Boon et al. [2013], 'What are they doing?', encompasses these three subtleties within it. Although I probed the microbial community dynamics through two facades, the explorations were interconnected. Throughout the thesis, I explored links between the two broader sections. Chapter 3 laid the foundation for Chapter 4 and Chapter 5. The quantification of temporal dynamics in Chapter 4 led to the visual exploration of temporal variation patterns presented in Chapter 5. The clustering of temporal behaviours offered a possible explanation for the multiple solutions discussed in Chapter 3 and laid the groundwork to characterise the microbial communities in Chapter 6. Though not a part of this thesis, future research directions exist in using this characterisation in understanding the MINs.

In conclusion, I have compiled this thesis with research outcomes within the scope of 'What are they [microbial organisms] doing [in their communities]?'. I drew inspiration from the answers to the challenging question of 'Who is there?', and I believe that I have laid a path towards better answering the challenging question of 'How will they respond [to external stimuli]?'.

7.1 Future Work

I also acknowledge that in this thesis, sometimes I have uncovered limitations of my methods where extensions are possible, and sometimes I had to limit the scientific exploration due to externally enforced constraints—such as time, not due to lack of interest. At other times, I encountered issues that have been left unresolved to be addressed later. In this section, I will discuss some of these and frame them as future research questions in their own right. I will also discuss exciting prospects of future work, for which I laid the foundations in this thesis.

7.1.1 Ranking Multiple Solutions from IMPARO

In Chapter 3 we inferred multiple solutions for microbial interaction networks with IMPARO. However, we stopped short of ranking the solutions according to their likelihood of being the best solution. The ranking is a particularly challenging task as the multiple solutions are viewed equally given analytical metrics. There need to be subjective biological considerations for the ranking, which is carried out ideally by microbiology researchers.

7.1.2 Using Autoencoders for Interaction Inference

Autoencoders have an interesting usage in identifying underlying properties of a dataset. When an autoencoder recreates a dataset, the latent space is a representation of the said dataset. Long-Short Term Memory (LSTM) autoencoders could be successfully applied to recreate longitudinal abundance profiles. Hence we assume

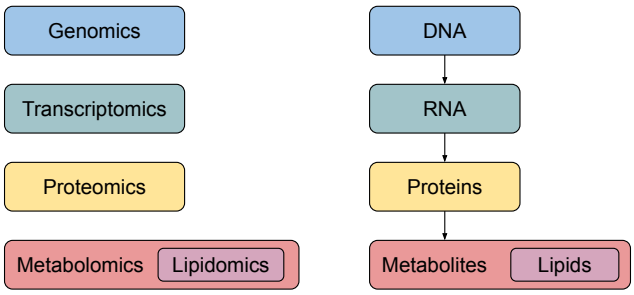

Figure 7.1: Central Dogma of Molecular Biology

that the latent space would carry a fair representation of the dynamics involved in the abundance profile generation. We tried this in an attempt to capture the nature of temporally dynamic microbial behaviour described in Chapter 4. However, we were unsuccessful in interpreting the latent variables.

Our approach was to restrict the latent representation. In our first attempt, we modelled the latent representation after the LV equations. We also tried a relaxed representation of the LV equations where a set of second-order differential equations—where the original LV parameters were modelled to vary with time—was used for the latent space. However, the models we trained suffered from posterior collapse, where the model learns to ignore the latent variables and the posterior mimics the prior. This space remains an exciting area to investigate, especially with fine-tuning the latent representation restrictions and giving it more freedom. However, that approach will require new models with more degrees of freedom than LV models to explain the microbial dynamics.

If successful, such an approach would not be limited in the area of inferring microbial interactions but could be used in many other instances where temporally dynamic systems are modelled.

7.1.3 Augmentation with other Omics Data

In this thesis, we looked at longitudinal microbial abundance profiles derived from 16S rRNA sequencing data. While this metagenomic data has much informative value, other omics data is likely to be widely available in the future. When considering the connections between the different omics data, as seen in Figure 7.1, we can consider a future research direction where we exploit this and augment our datasets.

As technology develops, the costs associated with data collections will reduce, allowing scientists to collect multiple omics data from a single study, which can significantly expand the opportunities to use data analysis methods that take advantage of all the available data.

7.1.4 Using Collective Pattern Recognition to Understand Repeating Patterns of Microbial Dynamics

In Chapters 4 and 5 we looked at dynamics of microbial interactions, and collective pattern recognition respectively. We note that we can separate the dataset into time windows in analysing longer longitudinal datasets and use visualising techniques from collective pattern recognition. These techniques would allow us to examine the microbial behavioural dynamics at different time points—such as important life events—and acknowledge any periodic or repeating behaviour.

7.1.5 Answers to 'How Will They Respond?'

As we explored the question 'What are they doing?', we are finally ready to the advice from Boon et al. [2013] and shift our focus to understand 'How will they respond?'. We can integrate background and clinical data into our analysis pipelines to investigate the potential effects of external factors on the microbiome. As we already know, diurnal cycles, seasonal changes, temperature variations, etc., has an observable effect on the microbiome; we can try to isolate them. A simple approach to kick-start this research direction would be to assign external factors to a 'dummy' variable in the LV modelling.

7.1.6 Exploring Effects of Climate Change on the Microbiome

Extending from the above research direction, we can use the data analysis methods we discussed in this thesis to study climate change's effects on the microbiome. We have published a manuscript along this idea and presented our work at an IEEExplore indexed conference as mentioned in the Section 1.3. Our approach is simple. We look at environments where significant changes are observable in the macrobial ecology due to climate change-related events and analyse available microbial datasets from such environments. We again look at habitually receptive (for macrobial life) environments and analyse datasets from these. Comparing our analysis, we get a preliminary idea of the effects on the microbiome we can expect to observe. Some interesting results were initially obtained by comparing the microbiome of sites with high UV exposure with that of regular sites.

7.1.7 Testing on Diverse Microbial Datasets

The methods introduced in this thesis were mainly tested on the datasets introduced in Caporaso et al. [2011], La Rosa et al. [2014], and Grier et al. [2018]. Although they cover different body sites, all are related to the human microbiome. In Section 2.3.1 we summarised the availability of datasets from a variety of environments. Applying these methods to a variety of microbial communities from floral, faunal, and environmental settings is requisite.

Bibliography

Aach, J. and Church, G. M., 2001. Aligning gene expression time series with time warping algorithms. *Bioinformatics*, 17, 6 (2001), 495–508. doi:10.1093/bioinformatics/17.6.495. https://pubmed.ncbi.nlm.nih.gov/11395426/. (cited on pages xvi, 21, and 63)

Abel, G. J. and Sander, N., 2014. Quantifying global international migration flows. *Science (New York, N.Y.)*, 343, 6178 (3 2014), 1520–2. doi:10.1126/science.1248676. http://www.ncbi.nlm.nih.gov/pubmed/24675962. (cited on page 51)

Alshawaqfeh, M.; Serpedin, E.; and Younes, A. B., 2017. Inferring microbial interaction networks from metagenomic data using SgLV-EKF algorithm. *BMC Genomics*, 18, 3 (2017), 2–16. doi:10.1186/s12864-017-3605-x. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5374605/pdf/12864_2017_Article_3605.pdf. (cited on pages 13, 31, and 32)

Amann, R. and Rosselló-Móra, R., 2016. After All, Only Millions? *mBio*, 7, 4 (9 2016), 00999–16. doi:10.1128/MBIO.00999-16. http://www.ncbi.nlm.nih.gov/pubmed/27381294http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4958260. (cited on pages 30 and 48)

Ankerst, M.; Breunig, M. M.; Kriegel, H. P.; and Sander, J., 1999. OPTICS: Ordering Points to Identify the Clustering Structure. *SIGMOD Record (ACM Special Interest Group on Management of Data)*, 28, 2 (6 1999), 49–60. doi:10.1145/304181.304187. https://dl.acm.org/doi/10.1145/304181.304187. (cited on page 23)

Apprill, A., 2017. Marine Animal Microbiomes: Toward Understanding Host–Microbiome Interactions in a Changing Ocean. *Frontiers in Marine Science*, 4 (7 2017), 222. doi:10.3389/fmars.2017.00222. http://journal.frontiersin.org/article/10.3389/fmars.2017.00222/full. (cited on pages 30 and 48)

Bach, L. L.; Ram, A.; Ijaz, U. Z.; Evans, T. J.; and Lindström, J., 2021. A Longitudinal Study of the Human Oropharynx Microbiota Over Time Reveals a Common Core and Significant Variations With Self-Reported Disease. *Frontiers in Microbiology*, 11 (1 2021), 3545. doi:10.3389/fmicb.2020.573969. (cited on page 18)

Banerjee, S.; Tian, T.; Wei, Z.; Shih, N.; Feldman, M. D.; Peck, K. N.; DeMichele, A. M.; Alwine, J. C.; and Robertson, E. S., 2018. Distinct microbial signatures associated with different breast cancer types. *Frontiers in Microbiology*, 9, MAY (5 2018), 951. doi:10.3389/fmicb.2018.00951. http://journal.frontiersin.org/article/10.3389/fmicb.2018.00951/full. (cited on pages 25 and 97)

Bar-Joseph, Z.; Gerber, G. K.; Gifford, D. K.; Jaakkola, T. S.; and Simon, I., 2003. Continuous Representations of Time-Series Gene Expression Data. In *Journal of Computational Biology*, vol. 10, 341–356. Mary Ann Liebert, Inc. doi:10.1089/10665270360688057. https://www.liebertpub.com/doi/abs/10.1089/10665270360688057. (cited on pages xvi, 21, and 63)

Bar-Joseph, Z.; Gitter, A.; and Simon, I., 2012. Studying and modelling dynamic biological processes using time-series gene expression data. doi:10.1038/nrg3244. www.nature.com/reviews/genetics. (cited on pages xvi and 63)

Barberán, A.; Bates, S. T.; Casamayor, E. O.; and Fierer, N., 2012. Using network analysis to explore co-occurrence patterns in soil microbial communities. *ISME Journal*, 6, 2 (9 2012), 343–351. doi:10.1038/ismej.2011.119. https://www.nature.com/articles/ismej2011119. (cited on pages 12 and 33)

Bilmes, J. A., 1998. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *International Computer Science Institute*, 4, 510 (1998), 126. (cited on page 23)

Blaser, M. J.; Cardon, Z. G.; Cho, M. K.; Dangl, J. L.; Donohue, T. J.; Green, J. L.; Knight, R.; Maxon, M. E.; Northen, T. R.; Pollard, K. S.; and Brodie, E. L., 2016. Toward a Predictive Understanding of Earth's Microbiomes to Address 21st Century Challenges. (2016). doi:10.1128/mBio.00714-16. https://commons.wikimedia.org/. (cited on pages 30 and 48)

Bokeh Development Team, 2018. Bokeh: Python library for interactive visualization. https://bokeh.pydata.org/en/latest/. (cited on page 92)

Boon, E.; Meehan, C. J.; Whidden, C.; H-J Wong, D.; Langille, M. G.; and Beiko, R. G., 2013. Interactions in the microbiome: communities of organisms and communities of genes. (2013). doi:10.1111/1574-6976.12035. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4298764/pdf/fmr0038-0090.pdf. (cited on pages 1, 2, 30, 48, 60, 96, 117, 119, and 121)

Bradley, E. and Kantz, H., 2015. Nonlinear time-series analysis revisited. *Chaos*, 25, 9 (4 2015), 097610. doi:10.1063/1.4917289. http://aip.scitation.org/doi/10.1063/1.4917289. (cited on pages 26 and 27)

Bray, J. R. and Curtis, J. T., 1957. An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecological Monographs*, 27, 4 (2 1957), 325–349. doi:10.2307/1942268. (cited on pages 42 and 58)

Brooks, J. P.; Buck, G. A.; Chen, G.; Diao, L.; Edwards, D. J.; Fettweis, J. M.; Huzurbazar, S.; Rakitin, A.; Satten, G. A.; Smirnova, E.; Waks, Z.; Wright, M. L.; Yanover, C.; and Zhou, Y.-H., 2017. Changes in vaginal community state types reflect major shifts in the microbiome. *Microbial Ecology in Health and Disease*, 28, 1 (1 2017), 1303265. doi:10.1080/16512235.2017.1303265. https:

//www.tandfonline.com/doi/full/10.1080/16512235.2017.1303265. (cited on page 25)

Bryan, N. C.; Christner, B. C.; Guzik, T. G.; Granger, D. J.; and Stewart, M. F., 2019. Abundance and survival of microbial aerosols in the troposphere and stratosphere. *The ISME Journal*, (7 2019), 1–11. doi:10.1038/s41396-019-0474-0. http://www.nature.com/articles/s41396-019-0474-0. (cited on page 48)

Cammarota, G.; Ianiro, G.; Ahern, A.; Carbone, C.; Temko, A.; Claesson, M. J.; Gasbarrini, A.; and Tortora, G., 2020. Gut microbiome, big data and machine learning to promote precision medicine for cancer. *Nature Reviews Gastroenterology and Hepatology*, 17, 10 (10 2020), 635–648. doi:10.1038/s41575-020-0327-3. www.nature.com/nrgastro. (cited on pages 24, 64, and 85)

Cao, H. T.; Gibson, T. E.; Bashan, A.; and Liu, Y. Y., 2017. Inferring human microbial dynamics from temporal metagenomics data: Pitfalls and lessons. *BioEssays*, 39, 2 (2017), 1–12. doi:10.1002/bies.201600188. (cited on pages 34 and 56)

Caporaso, J. G.; Lauber, C. L.; Costello, E. K.; Berg-Lyons, D.; Gonzalez, A.; Stombaugh, J.; Knights, D.; Gajer, P.; Ravel, J.; Fierer, N.; Gordon, J. I.; and Knight, R., 2011. Moving pictures of the human microbiome. Technical report. doi:10.1186/gb-2011-12-5-r50. http://genomebiology.com/2011/12/5/R50. (cited on pages 17, 35, 37, 51, 53, 55, 57, 60, 87, and 121)

Castiglioni, P.; Piccini, L.; and Di Rienzo, M., 2003. Interpolation technique for extracting features from ECG signals sampled at low sampling rates. In *Computers in Cardiology*, vol. 30, 481–484. IEEE Computer Society. doi:10.1109/cic.2003.1291197. (cited on page 16)

Cavicchioli, R.; Ripple, W. J.; Timmis, K. N.; Azam, F.; Bakken, L. R.; Baylis, M.; Behrenfeld, M. J.; Boetius, A.; Boyd, P. W.; Classen, A. T.; Crowther, T. W.; Danovaro, R.; Foreman, C. M.; Huisman, J.; Hutchins, D. A.; Jansson, J. K.; Karl, D. M.; Koskella, B.; Mark Welch, D. B.; Martiny, J. B. H.; Moran, M. A.; Orphan, V. J.; Reay, D. S.; Remais, J. V.; Rich, V. I.; Singh, B. K.; Stein, L. Y.; Stewart, F. J.; Sullivan, M. B.; van Oppen, M. J. H.; Weaver, S. C.; Webb, E. A.; and Webster, N. S., 2019. Scientists' warning to humanity: microorganisms and climate change. *Nature Reviews Microbiology*, (6 2019), 1–18. doi:10.1038/s41579-019-0222-5. http://www.nature.com/articles/s41579-019-0222-5. (cited on page 48)

Chandereng, T. and Gitter, A., 2020. Lag penalized weighted correlation for time series clustering. *BMC Bioinformatics*, 21, 1 (1 2020). doi:10.1186/s12859-019-3324-1. https://pubmed.ncbi.nlm.nih.gov/31948388/. (cited on pages xvi, 20, 21, and 64)

Chen, I.; Kelkar, Y. D.; Gu, Y.; Zhou, J.; Qiu, X.; and Wu, H., 2017. High-dimensional linear state space models for dynamic microbial interaction networks. *PLOS ONE*, 12, 11 (11 2017), e0187822. doi:10.1371/journal.pone.0187822. https://dx.plos.org/10.1371/journal.pone.0187822. (cited on page 43)

Cho, I. and Blaser, M. J., 2012. Applications of Next-Generation Sequencing: The human microbiome: at the interface of health and disease. *Nature Publishing Group*, 13 (2012). doi:10.1038/nrg3182. www.nature.com/reviews/genetics. (cited on pages 30 and 48)

Ciechanowski, M.; Zajac, T.; Biłas, A.; and Dunajski, R., 2007. Spatiotemporal variation in activity of bat species differing in hunting tactics: Effects of weather, moonlight, food abundance, and structural clutter. *Canadian Journal of Zoology*, 85, 12 (12 2007), 1249–1263. doi:10.1139/Z07-090. https://cdnsciencepub.com/doi/abs/10.1139/Z07-090. (cited on page 48)

Clarke, S. F.; Murphy, E. F.; O'Sullivan, O.; Ross, R. P.; O'Toole, P. W.; Shanahan, F.; and Cotter, P. D., 2013. Targeting the Microbiota to Address Diet-Induced Obesity: A Time Dependent Challenge. *PLoS ONE*, 8, 6 (6 2013), e65790. doi:10.1371/journal.pone.0065790. http://dx.plos.org/10.1371/journal.pone.0065790. (cited on page 39)

Claussen, J. C.; Skiecevičienė, J.; Wang, J.; Rausch, P.; Karlsen, T. H.; Lieb, W.; Baines, J. F.; Franke, A.; and Hütt, M. T., 2017. Boolean analysis reveals systematic interactions among low-abundance species in the human gut microbiome. *PLoS Computational Biology*, 13, 6 (2017). doi:10.1371/journal.pcbi.1005361. (cited on pages 13, 31, and 32)

Clemente, J. C.; Ursell, L. K.; Parfrey, L. W.; and Knight, R., 2012. The Impact of the Gut Microbiota on Human Health: An Integrative VIew. *Cell*, , 148 (2012), 1258 – 1270. doi:10.1016/j.cell.2012.01.035. https://ac.els-cdn.com/S0092867412001043/1-s2.0-S0092867412001043-main.pdf?_tid=27e5c9b0-8ff0-493d-8987-ffef2d89aa69&acdnat=1542067127_d6d117a9fb79254d219d2474be932e34. (cited on pages 30 and 48)

Clooney, A. G.; Eckenberger, J.; Laserna-Mendieta, E.; Sexton, K. A.; Bernstein, M. T.; Vagianos, K.; Sargent, M.; Ryan, F. J.; Moran, C.; Sheehan, D.; Sleator, R. D.; Targownik, L. E.; Bernstein, C. N.; Shanahan, F.; and Claesson, M. J., 2021. Ranking microbiome variance in inflammatory bowel disease: A large longitudinal intercontinental study. *Gut*, 70, 3 (3 2021), 499–510. doi:10.1136/gutjnl-2020-321106. https://gut.bmj.com/content/70/3/499https://gut.bmj.com/content/70/3/499.abstract. (cited on page 20)

Comaniciu, D. and Meer, P., 2002. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 24, 5 (2002), 603–619. (cited on page 23)

Criel, J. and Tsiporkova, E., 2006. Gene time eχpression warper: A tool for alignment, template matching and visualization of gene expression time series. *Bioinformatics*, 22, 2 (1 2006), 251–252. doi:10.1093/bioinformatics/bti787. https://pubmed.ncbi.nlm.nih.gov/16293669/. (cited on pages xvi, 21, and 64)

Dam, P.; Fonseca, L. L.; Konstantinidis, K. T.; and Voit, E. O., 2016. Dynamic models of the complex microbial metapopulation of lake mendota. *npj Systems Biology and Applications*, 2 (2016). doi:10.1038/npjsba.2016.7. http://www.bst.bme.gatech.edu/research.php. (cited on pages 33 and 49)

Dawson, W.; Hör, J.; Egert, M.; van Kleunen, M.; and Pester, M., 2017. A Small Number of Low-abundance Bacteria Dominate Plant Species-specific Responses during Rhizosphere Colonization. *Frontiers in Microbiology*, 8 (5 2017), 975. doi:10.3389/fmicb.2017.00975. http://journal.frontiersin.org/article/10.3389/fmicb.2017.00975/full. (cited on page 42)

De Boor, C., 2001. *A Practical Guide to Splines; Rev. Ed.* Applied mathematical sciences. Springer, Berlin. https://cds.cern.ch/record/1428148. (cited on page 22)

De Filippis, F.; Vitaglione, P.; Cuomo, R.; Canani, R. B.; and Ercolini, D., 2018. Dietary interventions to modulate the gut microbiome-how far away are we from precision medicine. doi:10.1093/ibd/izy080. https://academic.oup.com/ibdjournal/article/24/10/2142/4970097. (cited on pages 24 and 64)

De Seta, F.; Campisciano, G.; Zanotta, N.; Ricci, G.; and Comar, M., 2019. The vaginal community state types microbiome-immune network as key factor for bacterial vaginosis and aerobic vaginitis. *Frontiers in Microbiology*, 10, OCT (10 2019), 2451. doi:10.3389/fmicb.2019.02451. https://www.frontiersin.org/article/10.3389/fmicb.2019.02451/full. (cited on page 25)

DeSalle, R. and Perkins, S. L., 2015. *Welcome to the microbiome: getting to know the trillions of bacteria and other microbes in, on, and around you*. Yale University Press. (cited on page 1)

Desikan, P., 2017. Our microbial signatures. doi:10.4103/ijmm.IJMM{_}17{_}250. https://linkinghub.elsevier.com/retrieve/pii/S0255085720303789. (cited on pages 25 and 97)

DiGiulio, D. B.; Callahan, B. J.; McMurdie, P. J.; Costello, E. K.; Lyell, D. J.; Robaczewska, A.; Sun, C. L.; Goltsman, D. S.; Wong, R. J.; Shawa, G.; Stevenson, D. K.; Holmes, S. P.; and Relman, D. A., 2015. Temporal and spatial variation of the human microbiota during pregnancy. *Proceedings of the National Academy of Sciences of the United States of America*, 112, 35 (9 2015), 11060–11065. doi:10.1073/pnas.1502875112. https://www.pnas.org/content/112/35/11060https://www.pnas.org/content/112/35/11060.abstract. (cited on pages 19, 24, 25, 62, 64, 65, and 96)

Dodge, Y., 2008. Kolmogorov–Smirnov Test. In *The Concise Encyclopedia of Statistics*, 283–287. Springer New York, New York, NY. ISBN 978-0-387-32833-1. doi:10.1007/978-0-387-32833-1{_}214. https://doi.org/10.1007/978-0-387-32833-1_214. (cited on pages 43 and 45)

Dong, F.; He, Y.; Wang, T.; Han, D.; Lu, H.; and Zhao, H., 2020. Predicting viral exposure response from modeling the changes of co-expression networks using time series gene expression data. *BMC Bioinformatics*, 21, 1 (8 2020). doi:10.1186/s12859-020-03705-0. https://pubmed.ncbi.nlm.nih.gov/32842958/. (cited on pages xvi, 21, and 64)

Durán, C.; Ciucci, S.; Palladini, A.; Ijaz, U. Z.; Zippo, A. G.; Sterbini, F. P.; Masucci, L.; Cammarota, G.; Ianiro, G.; Spuul, P.; Schroeder, M.; Grill, S. W.; Parsons, B. N.; Pritchard, D. M.; Posteraro, B.; Sanguinetti, M.; Gasbarrini, G.; Gasbarrini, A.; and Cannistraci, C. V., 2021. Nonlinear machine learning pattern recognition and bacteria-metabolite multilayer network analysis of perturbed gastric microbiome. *Nature Communications*, 12, 1 (3 2021), 1–22. doi:10.1038/s41467-021-22135-x. https://www.nature.com/articles/s41467-021-22135-x. (cited on page 11)

Ester, M.; Kriegel, H.-P.; Sander, J.; and Xu, X., 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, 226–231. www.aaai.org. (cited on page 23)

Faust, K.; Lahti, L.; Gonze, D.; de Vos, W. M.; and Raes, J., 2015. Metagenomics meets time series analysis: Unraveling microbial community dynamics. doi:10.1016/j.mib.2015.04.004. https://ac.els-cdn.com/S1369527415000478/1-s2.0-S1369527415000478-main.pdf?_tid=3033124e-c52e-11e7-8608-00000aab0f01&acdnat=1510218962_587eb4a1c04ff7b1a6a465a3e02616eb. (cited on pages 15 and 48)

Faust, K. and Raes, J., 2012. Microbial interactions: from networks to models. (2012). doi:10.1038/nrmicro2832. www.nature.com/reviews/micro. (cited on page 30)

Fenchel, T.; Whitfield, M.; Meadows, P.; and Huisman, J. Microbial Ecology on Land and Sea [and Discussion]. doi:10.2307/55757. https://www.jstor.org/stable/55757. (cited on page 48)

Finkel, O. M.; Castrillo, G.; Herrera Paredes, S.; Salas González, I.; and Dangl, J. L., 2017. Understanding and exploiting plant beneficial microbes. *Current Opinion in Plant Biology*, 38 (8 2017), 155–163. doi:10.1016/j.pbi.2017.04.018. https://linkinghub.elsevier.com/retrieve/pii/S1369526617300158. (cited on pages 30 and 48)

Fisher, C. K. and Mehta, P., 2014. Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression. *PLoS ONE*, 9, 7 (2014), 1–10. doi:10.1371/journal.pone.0102451. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0102451. (cited on pages 11, 12, and 31)

Fitzpatrick, C. R.; Copeland, J.; Wang, P. W.; Guttman, D. S.; Kotanen, P. M.; and Johnson, M. T., 2018. Assembly and ecological function of the root microbiome

across angiosperm plant species. *Proceedings of the National Academy of Sciences of the United States of America*, 115, 6 (2 2018), E1157–E1165. doi:10.1073/pnas.1717617115. (cited on pages 30 and 48)

Freire, M.; Moustafa, A.; Harkins, D. M.; Torralba, M. G.; Zhang, Y.; Leong, P.; Saffery, R.; Bockmann, M.; Kuelbs, C.; Hughes, T.; Craig, J. M.; and Nelson, K. E., 2020. Longitudinal Study of Oral Microbiome Variation in Twins. *Scientific Reports*, 10, 1 (5 2020), 1–10. doi:10.1038/s41598-020-64747-1. https://www.nature.com/articles/s41598-020-64747-1. (cited on page 19)

Friedman, J. and Alm, E. J., 2012. Inferring Correlation Networks from Genomic Survey Data. *PLoS Computational Biology*, 8, 9 (9 2012), e1002687. doi:10.1371/journal.pcbi.1002687. https://dx.plos.org/10.1371/journal.pcbi.1002687. (cited on pages 12, 31, 32, and 62)

Fulcher, B. D., 2017. Feature-based time-series analysis. doi:10.1201/9781315181080-4. http://arxiv.org/abs/1709.08055. (cited on pages 26, 27, and 97)

Funchain, P. and Eng, C., 2013. Hunting for cancer in the microbial jungle. *Genome Medicine*, 5, 42 (2013). doi:10.1186/gm446. http://genomemedicine.com/content/5/5/42. (cited on pages 30 and 48)

Gajer, P.; Brotman, R. M.; Bai, G.; Sakamoto, J.; Schütte, U. M.; Zhong, X.; Koenig, S. S.; Fu, L.; Ma, Z.; Zhou, X.; Abdo, Z.; Forney, L. J.; and Ravel, J., 2012. Temporal dynamics of the human vaginal microbiota. *Science Translational Medicine*, 4, 132 (5 2012), 132ra52. doi:10.1126/scitranslmed.3003605. /pmc/articles/PMC3722878/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3722878/. (cited on pages 19 and 62)

Gao, X.; Huynh, B.-T.; Guillemot, D.; Glaser, P.; and Opatowski, L., 2018. Inference of Significant Microbial Interactions From Longitudinal Metagenomics Data. *Frontiers in microbiology*, 9 (2018), 2319. doi:10.3389/fmicb.2018.02319. http://www.ncbi.nlm.nih.gov/pubmed/30386306http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6198172. (cited on pages 13, 32, 50, and 56)

Gerber, G. K.; Onderdonk, A. B.; and Bry, L., 2012. Inferring Dynamic Signatures of Microbes in Complex Host Ecosystems. *PLoS Computational Biology*, 8, 8 (8 2012), e1002624. doi:10.1371/journal.pcbi.1002624. https://dx.plos.org/10.1371/journal.pcbi.1002624. (cited on pages 25 and 97)

Gibson, T. E.; Bashan, A.; Cao, H.-T.; Weiss, S. T.; and Liu, Y.-Y., 2016. On the Origins and Control of Community Types in the Human Microbiome. *PLOS Computational Biology | Liu Y-Y*, 12, 2 (2016), 1004688. doi:10.1371/journal.pcbi.1004688. http://qiita.ucsd.edu. (cited on pages 34, 35, and 42)

Gilpin, M. E. and Ayala, F. J., 1973. Global models of growth and competition. *Proceedings of the National Academy of Sciences of the United States of America*, 70,

12 (I) (12 1973), 3590–3593. doi:10.1073/pnas.70.12.3590. https://www.pnas.org/content/70/12/3590https://www.pnas.org/content/70/12/3590.abstract. (cited on page 50)

Gong, J.; Qing, Y.; Zou, S.; Fu, R.; Su, L.; Zhang, X.; and Zhang, Q., 2016. Protist-bacteria associations: Gammaproteobacteria and Alphaproteobacteria are prevalent as digestion-resistant bacteria in ciliated protozoa. *Frontiers in Microbiology*, 7, APR (4 2016), 498. doi:10.3389/fmicb.2016.00498. http://journal.frontiersin.org/Article/10.3389/fmicb.2016.00498/abstract. (cited on pages 24 and 64)

González-Cabaleiro, R.; Mitchell, A. M.; Smith, W.; Wipat, A.; and Ofiteru, I. D., 2017. Heterogeneity in pure microbial systems: Experimental measurements and modeling. doi:10.3389/fmicb.2017.01813. www.frontiersin.org. (cited on page 4)

Grier, A.; McDavid, A.; Wang, B.; Qiu, X.; Java, J.; Bandyopadhyay, S.; Yang, H.; Holden-Wiltse, J.; Kessler, H. A.; Gill, A. L.; Huyck, H.; Falsey, A. R.; Topham, D. J.; Scheible, K. M.; Caserta, M. T.; Pryhuber, G. S.; and Gill, S. R., 2018. Neonatal gut and respiratory microbiota: Coordinated development through time and space. *Microbiome*, 6, 1 (10 2018), 193. doi:10.1186/s40168-018-0566-5. https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-018-0566-5. (cited on pages xviii, xxi, 19, 24, 62, 64, 72, 73, 74, 75, 76, 77, 79, 81, 88, 96, 99, 101, 103, 104, 105, 110, 113, and 121)

Hanke, J. E. and Wichern, D. W., 2013. *Business Forecasting*. Pearson, 9 edn. ISBN 9781292023007. https://www.pearson.com/us/higher-education/program/Hanke-Business-Forecasting-9th-Edition/PGM58861.html. (cited on page 16)

Hasegawa, M. and Inohara, N., 2014. Regulation of the gut microbiota by the mucosal immune system in mice. *International Immunology*, 26, 9 (9 2014), 481–487. doi:10.1093/intimm/dxu049. https://academic.oup.com/intimm/article-lookup/doi/10.1093/intimm/dxu049. (cited on page 39)

Hibberd, M. L., 2013. Microbial genomics: an increasingly revealing interface in human health and disease. *Genome Medicine*, 5, 31 (2013). doi:10.1186/gm435. http://genomemedicine.com/content/5/4/31. (cited on pages 30 and 48)

Hiergeist, A.; Gläsner, J.; Reischl, U.; and Gessner, A., 2015. Analyses of Intestinal Microbiota: Culture versus Sequencing. (2015). doi:10.1093/ilar/ilv017. https://academic.oup.com/ilarjournal/article-abstract/56/2/228/650795. (cited on pages 30 and 48)

Hiltemann, S. D.; Boers, S. A.; Van Der Spek, P. J.; Jansen, R.; Hays, J. P.; and Stubbs, A. P., 2019. Galaxy mothur Toolset (GmT): A user-friendly application for 16S rRNA gene sequencing analysis using mothur. *GigaScience*, 8, 2 (2 2019), 1–5. doi:10.1093/gigascience/giy166. https://academic.oup.com/gigascience/article/8/2/giy166/5266305. (cited on page 2)

Holland, J. H., 1992. Genetic Algorithms. *Scientific American*, 267, 1 (1992), 66–73. http://www.jstor.org/stable/24939139. (cited on page 45)

Hosoda, S.; Fukunaga, T.; and Hamada, M., 2021. Umibato: estimation of time-varying microbial interaction using continuous-time regression hidden Markov model. *bioRxiv*, (3 2021), 2021.01.28.428580. doi:10.1101/2021.01.28.428580. https://doi.org/10.1101/2021.01.28.428580. (cited on page 15)

Hull, D. L., 1980. Individuality and Selection. *Annual Review of Ecology and Systematics*, 11 (1980), 311–332. http://www.jstor.org/stable/2096911. (cited on pages 24 and 82)

Janda, J. M. and Abbott, S. L., 2007. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *Journal of clinical microbiology*, 45, 9 (9 2007), 2761–4. doi:10.1128/JCM.01228-07. http://www.ncbi.nlm.nih.gov/pubmed/17626177http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2045242. (cited on page 48)

Jiang, P.; Chamberlain, C. S.; Vanderby, R.; Thomson, J. A.; and Stewart, R., 2020. TimeMeter assesses temporal gene expression similarity and identifies differentially progressing genes. *Nucleic acids research*, 48, 9 (5 2020), e51. doi:10.1093/nar/gkaa142. https://pubmed.ncbi.nlm.nih.gov/32123905/. (cited on pages xvi, 21, and 64)

Khanna, S. and Tosh, P. K., 2014. A clinician's primer on the role of the microbiome in human health and disease. doi:10.1016/j.mayocp.2013.10.011. (cited on pages 30 and 48)

Kim, M. and Or, D., 2017. Hydration status and diurnal trophic interactions shape microbial community function in desert biocrusts. *Biogeosciences*, 14, 23 (12 2017), 5403–5424. doi:10.5194/bg-14-5403-2017. (cited on page 48)

Kingsbury, B., 1997. The concept of compliance as a function of competing conceptions of international law. *Mich. J. Int'l L.*, 19 (1997), 345. (cited on page 23)

Knights, D.; Parfrey, L. W.; Zaneveld, J.; Lozupone, C.; and Knight, R., 2011. Human-associated microbial signatures: Examining their predictive value. doi:10.1016/j.chom.2011.09.003. /pmc/articles/PMC3879110//pmc/articles/PMC3879110/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3879110/. (cited on pages 25 and 97)

Koenig, J. E.; Spor, A.; Scalfone, N.; Fricker, A. D.; Stombaugh, J.; Knight, R.; Angenent, L. T.; and Ley, R. E., 2011. Succession of microbial consortia in the developing infant gut microbiome. *Proceedings of the National Academy of Sciences of the United States of America*, 108 Suppl, Suppl 1 (3 2011), 4578–85. doi:10.1073/pnas.1000081107. http://www.ncbi.nlm.nih.gov/pubmed/20668239http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3063592. (cited on pages 17 and 55)

Kolar, M.; Song, L.; Ahmed, A.; and Xing, E. P., 2012. Estimating time-varying networks. *Annals of Applied Statistics*, 6, 1 (3 2012), 94–123. doi:10.1214/09-AOAS308. https://projecteuclid.org/journals/annals-of-applied-statistics/volume-4/issue-1/Estimating-time-varying-networks/10.1214/09-AOAS308.full. (cited on page 14)

Kumar, A. and Chordia, N., 2017. Role of Microbes in Human Health. *Applied Microbiology: Open Access*, 03, 02 (2017). doi:10.4172/2471-9315.1000131. (cited on pages 30 and 48)

Kurtz, Z. D.; Müller, C. L.; Miraldi, E. R.; Littman, D. R.; Blaser, M. J.; and Bonneau, R. A., 2015. Sparse and Compositionally Robust Inference of Microbial Ecological Networks. *PLoS Computational Biology*, 11, 5 (2015), 1–25. doi:10.1371/journal.pcbi.1004226. (cited on pages 12 and 31)

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., 2014. Patterned progression of bacterial populations in the premature infant gut. *Proceedings of the National Academy of Sciences of the United States of America*, 111, 34 (8 2014), 12522–12527. doi:10.1073/pnas.1409497111. (cited on pages xviii, xxi, 18, 61, 62, 66, 68, 69, 70, 71, 72, 73, 74, 75, 81, 84, 88, 96, 102, 106, 107, 110, 113, and 121)

Layeghifard, M.; Hwang, D. M.; and Guttman, D. S., 2017. Disentangling Interactions in the Microbiome: A Network Perspective. doi:10.1016/j.tim.2016.11.008. (cited on page 15)

Lee, G.; Gommers, R.; Waselewski, F.; Wohlfahrt, K.; and O'Leary, A., 2019. PyWavelets: A Python package for wavelet analysis. *Journal of Open Source Software*, 4, 36 (4 2019), 1237. doi:10.21105/joss.01237. http://arxiv.org/abs/1812.11214. (cited on page 90)

Lee, S. T.; Davy, S. K.; Tang, S. L.; Fan, T. Y.; and Kench, P. S., 2015. Successive shifts in the microbial community of the surface mucus layer and tissues of the coral Acropora muricata under thermal stress. *FEMS microbiology ecology*, 91, 12 (2015), 1–11. doi:10.1093/femsec/fiv142. (cited on page 48)

Ligon, B. L., 2004. Penicillin: Its Discovery and Early Development. *Seminars in Pediatric Infectious Diseases*, 15, 1 (1 2004), 52–57. doi:10.1053/j.spid.2004.02.001. (cited on page 4)

Locey, K. J. and Lennon, J. T., 2016. Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences of the United States of America*, 113, 21 (5 2016), 5970–5. doi:10.1073/pnas.1521291113. http://www.ncbi.nlm.nih.gov/pubmed/27140646http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4889364. (cited on pages 30 and 48)

Lovett, G. M.; Tear, T. H.; Evers, D. C.; Findlay, S. E.; Cosby, B. J.; Dunscomb, J. K.; Driscoll, C. T.; and Weathers, K. C., 2009. Effects of air pollution on ecosystems and biological diversity in the eastern United States. doi:10.1111/j.1749-6632.2009.04153.x. (cited on page 48)

Lugo-Martinez, J.; Ruiz-Perez, D.; Narasimhan, G.; and Bar-Joseph, Z., 2019. Dynamic interaction network inference from longitudinal microbiome data. *Microbiome*, 7, 1 (12 2019), 54. doi:10.1186/s40168-019-0660-3. https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-019-0660-3. (cited on pages xvi, 21, 56, 61, 62, 63, 65, 86, and 98)

Ma, Z. S. and Li, L., 2017. Quantifying the human vaginal community state types (CSTs) with the species specificity index. *PeerJ*, 2017, 6 (6 2017), e3366. doi:10.7717/peerj.3366. https://peerj.com/articles/3366. (cited on page 25)

Marino, S.; Baxter, N. T.; Huffnagle, G. B.; Petrosino, J. F.; and Schloss, P. D., 2014. Mathematical modeling of primary succession of murine intestinal microbiota. *PNAS*, 111, 1 (2014), 439–444. doi:10.1073/pnas.1311322111. http://www.pnas.org/content/pnas/111/1/439.full.pdf. (cited on pages 17, 33, 49, and 55)

Martins, B. M. and Locke, J. C., 2015. Microbial individuality: how single-cell heterogeneity enables population level strategies. doi:10.1016/j.mib.2015.01.003. (cited on pages 24 and 64)

McDonald, D.; Hyde, E.; Debelius, J. W.; Morton, J. T.; Gonzalez, A.; Ackermann, G.; Aksenov, A. A.; Behsaz, B.; Brennan, C.; Chen, Y.; DeRight Goldasich, L.; Dorrestein, P. C.; Dunn, R. R.; Fahimipour, A. K.; Gaffney, J.; Gilbert, J. A.; Gogul, G.; Green, J. L.; Hugenholtz, P.; Humphrey, G.; Huttenhower, C.; Jackson, M. A.; Janssen, S.; Jeste, D. V.; Jiang, L.; Kelley, S. T.; Knights, D.; Kosciolek, T.; Ladau, J.; Leach, J.; Marotz, C.; Meleshko, D.; Melnik, A. V.; Metcalf, J. L.; Mohimani, H.; Montassier, E.; Navas-Molina, J.; Nguyen, T. T.; Peddada, S.; Pevzner, P.; Pollard, K. S.; Rahnavard, G.; Robbins-Pianka, A.; Sangwan, N.; Shorenstein, J.; Smarr, L.; Song, S. J.; Spector, T.; Swafford, A. D.; Thackray, V. G.; Thompson, L. R.; Tripathi, A.; Vázquez-Baeza, Y.; Vrbanac, A.; Wischmeyer, P.; Wolfe, E.; Zhu, Q.; Knight, R.; Mann, A. E.; Amir, A.; Frazier, A.; Martino, C.; Lebrilla, C.; Lozupone, C.; Lewis, C. M.; Raison, C.; Zhang, C.; Lauber, C. L.; Warinner, C.; Lowry, C. A.; Callewaert, C.; Bloss, C.; Willner, D.; Galzerani, D. D.; Gonzalez, D. J.; Mills, D. A.; Chopra, D.; Gevers, D.; Berg-Lyons, D.; Sears, D. D.; Wendel, D.; Lovelace, E.; Pierce, E.; TerAvest, E.; Bolyen, E.; Bushman, F. D.; Wu, G. D.; Church, G. M.; Saxe, G.; Holscher, H. D.; Ugrina, I.; German, J. B.; Caporaso, J. G.; Wozniak, J. M.; Kerr, J.; Ravel, J.; Lewis, J. D.; Suchodolski, J. S.; Jansson, J. K.; Hampton-Marcell, J. T.; Bobe, J.; Raes, J.; Chase, J. H.; Eisen, J. A.; Monk, J.; Clemente, J. C.; Petrosino, J.; Goodrich, J.; Gauglitz, J.; Jacobs, J.; Zengler, K.; Swanson, K. S.; Lewis, K.; Mayer, K.; Bittinger, K.; Dillon, L.; Zaramela, L. S.;

Schriml, L. M.; Dominguez-Bello, M. G.; Jankowska, M. M.; Blaser, M.; Pirrung, M.; Minson, M.; Kurisu, M.; Ajami, N.; Gottel, N. R.; Chia, N.; Fierer, N.; White, O.; Cani, P. D.; Gajer, P.; Strandwitz, P.; Kashyap, P.; Dutton, R.; Park, R. S.; Xavier, R. J.; Mills, R. H.; Krajmalnik-Brown, R.; Ley, R.; Owens, S. M.; Klemmer, S.; Matamoros, S.; Mirarab, S.; Moorman, S.; Holmes, S.; Schwartz, T.; Eshoo-Anton, T. W.; Vigers, T.; Pandey, V.; Treuren, W. V.; Fang, X.; Zech Xu, Z.; Jarmusch, A.; Geier, J.; Reeve, N.; Silva, R.; Kopylova, E.; Nguyen, D.; Sanders, K.; Salido Benitez, R. A.; Heale, A. C.; Abramson, M.; Waldispühl, J.; Butyaev, A.; Drogaris, C.; Nazarova, E.; Ball, M.; and Gunderson, B., 2018. American Gut: an Open Platform for Citizen Science Microbiome Research. *mSystems*, 3, 3 (5 2018), 00031–18. doi:10.1128/mSystems.00031-18. http://msystems.asm.org/lookup/doi/10.1128/mSystems.00031-18. (cited on pages 17, 31, and 55)

McInnes, L.; Healy, J.; and Melville, J., 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv*, (2 2018). http://arxiv.org/abs/1802.03426. (cited on pages 27, 52, 58, 91, and 115)

Metropolis, N. and Ulam, S., 1949. The Monte Carlo Method. *Journal of the American Statistical Association*, 44, 247 (1949), 335–341. doi:10.1080/01621459.1949.10483310. (cited on page 34)

Minich, J. J.; Morris, M. M.; Brown, M.; Doane, M.; Edwards, M. S.; Michael, T. P.; and Dinsdale, E. A., 2018. Elevated temperature drives kelp microbiome dysbiosis, while elevated carbon dioxide induces water microbiome disruption. (2018). doi:10.1371/journal.pone.0192772. https://doi.org/10.1371/journal.pone.0192772. (cited on pages 30 and 48)

Mitra, A.; MacIntyre, D. A.; Lee, Y. S.; Smith, A.; Marchesi, J. R.; Lehne, B.; Bhatia, R.; Lyons, D.; Paraskevaidis, E.; Li, J. V.; Holmes, E.; Nicholson, J. K.; Bennett, P. R.; and Kyrgiou, M., 2015. Cervical intraepithelial neoplasia disease progression is associated with increased vaginal microbiome diversity. *Scientific Reports*, 5, 1 (11 2015), 1–11. doi:10.1038/srep16865. www.nature.com/scientificreports. (cited on page 25)

Momeni, B.; Xie, L.; and Shou, W., 2017. Lotka-Volterra pairwise modeling fails to capture diverse pairwise microbial interactions. *eLife*, 6 (3 2017). doi:10.7554/eLife.25051.001. (cited on page 50)

Montassier, E.; Al-Ghalith, G. A.; Hillmann, B.; Viskocil, K.; Kabage, A. J.; McKinlay, C. E.; Sadowsky, M. J.; Khoruts, A.; and Knights, D., 2018. CLOUD: A non-parametric detection test for microbiome outliers. *Microbiome*, 6, 1 (8 2018), 137. doi:10.1186/s40168-018-0514-4. https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-018-0514-4. (cited on pages 24 and 82)

Mounier, J.; Monnet, C.; Vallaeys, T.; Arditi, R.; Sarthou, A. S.; Hélias, A.; and Irlinger, F., 2008. Microbial interactions within a cheese microbial community.

Applied and Environmental Microbiology, 74, 1 (2008), 172–181. doi:10.1128/AEM.01338-07. http://aem.asm.org/content/74/1/172.full.pdf. (cited on pages 17, 33, 49, and 55)

Mueller, U. and Sachs, J., 2015. UC Riverside UC Riverside Previously Published Works Title Engineering Microbiomes to Improve Plant and Animal Health Publication Date. *Trends in Microbiology*, (2015). doi:10.1016/j.tim.2015.07.009. http://dx.doi.org/10.1016/j.tim.2015.07.009. (cited on pages 30 and 48)

Mughal, J. R. D. M. A., 2014. Individuality and Conformity. *SSRN Electronic Journal*, (9 2014). doi:10.2139/ssrn.2489676. https://papers.ssrn.com/abstract=2489676. (cited on pages 23 and 82)

Nason, G. P.; Powell, B.; Elliott, D.; and Smith, P. A., 2017. Should we sample a time series more frequently?: decision support via multirate spectrum estimation. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 180, 2 (2017), 353–407. doi:10.1111/rssa.12210. (cited on page 16)

O'Shaughnessy-Hunter, L. C.; Yu, A.; Rousseau, J. D.; Foster, R. A.; and Weese, J. S., 2021. Longitudinal study of the cutaneous microbiota of healthy horses. *Veterinary Dermatology*, (2021). doi:10.1111/vde.12983. https://onlinelibrary.wiley.com/doi/full/10.1111/vde.12983https://onlinelibrary.wiley.com/doi/abs/10.1111/vde.12983https://onlinelibrary.wiley.com/doi/10.1111/vde.12983. (cited on page 18)

Packard, N. H.; Crutchfield, J. P.; Farmer, J. D.; and Shaw, R. S., 1980. Geometry from a time series. *Physical Review Letters*, 45, 9 (9 1980), 712–716. doi:10.1103/PhysRevLett.45.712. https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.45.712. (cited on page 26)

Pecar, B., 2004. Visual recurrence analysis as an alternative framework for time series characterisation. *Proceedings of the Computational Finance, Bologna, 2004*, 38 (3 2004). www.witpress.com,. (cited on page 26)

Pepper, J. W., 2014. The evolution of bacterial social life: From the ivory tower to the front lines of public health. *Evolution, Medicine and Public Health*, 2014, 1 (2014), 65–68. doi:10.1093/emph/eou010. /pmc/articles/PMC3981165//pmc/articles/PMC3981165/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3981165/. (cited on page 2)

Quevauviller, P., 2004. Quality Assurance - Production of Reference Materials. In *Encyclopedia of Analytical Science: Second Edition*, 462–469. Elsevier Inc. ISBN 9780123693976. doi:10.1016/B0-12-369397-7/00510-0. (cited on page 3)

Ravel, J.; Gajer, P.; Abdo, Z.; Schneider, G. M.; Koenig, S. S.; McCulle, S. L.; Karlebach, S.; Gorle, R.; Russell, J.; Tacket, C. O.; Brotman, R. M.;

Davis, C. C.; Ault, K.; Peralta, L.; and Forney, L. J., 2011. Vaginal microbiome of reproductive-age women. *Proceedings of the National Academy of Sciences of the United States of America*, 108, SUPPL. 1 (3 2011), 4680–4687. doi:10.1073/pnas.1002611107. /pmc/articles/PMC3063603//pmc/articles/PMC3063603/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3063603/. (cited on pages 24, 25, 64, and 97)

Reis-Filho, J. S., 2009. Next-generation sequencing. *Breast cancer research*, 11, 3 (2009), 1–7. (cited on page 2)

Romero, R.; Hassan, S. S.; Gajer, P.; Tarca, A. L.; Fadrosh, D. W.; Nikita, L.; Galuppi, M.; Lamont, R. F.; Chaemsaithong, P.; Miranda, J.; Chaiworapongsa, T.; and Ravel, J., 2014. The composition and stability of the vaginal microbiota of normal pregnant women is different from that of non-pregnant women. *Microbiome*, 2, 1 (2 2014), 1–19. doi:10.1186/2049-2618-2-4. http://www.microbiomejournal.com/content/2/1/10. (cited on pages 25 and 97)

Sainburg, T.; McInnes, L.; and Gentner, T. Q., 2020. Parametric UMAP: learning embeddings with deep neural networks for representation and semi-supervised learning. http://arxiv.org/abs/2009.12981. (cited on pages 27 and 97)

Sanguinetti, E.; Guzzardi, M. A.; Tripodi, M.; Panetta, D.; Selma-Royo, M.; Zega, A.; Telleschi, M.; Collado, M. C.; and Iozzo, P., 2019. Microbiota signatures relating to reduced memory and exploratory behaviour in the offspring of overweight mothers in a murine model. *Scientific Reports*, 9, 1 (12 2019), 1–12. doi:10.1038/s41598-019-48090-8. https://doi.org/10.1038/s41598-019-48090-8. (cited on page 25)

Sastry, K.; Goldberg, D.; and Kendall, G., 2005. Genetic Algorithms. In *Search Methodologies*, 97–125. Springer US, Boston, MA. doi:10.1007/0-387-28356-0{_}4. http://link.springer.com/10.1007/0-387-28356-0_4. (cited on page 45)

Savitzky, A. and Golay, M. J., 1964. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, 36, 8 (1964), 1627–1639. doi:10.1021/ac60214a047. https://pubs.acs.org/sharingguidelines. (cited on pages 22 and 58)

Schubert, E.; Sander, J.; Ester, M.; Kriegel, H. P.; and Xu, X., 2017. DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems*, 42, 3 (7 2017), 1–21. doi:10.1145/3068335. https://dl.acm.org/doi/10.1145/3068335. (cited on page 23)

Senanayake, D.; Wang, W.; Naik, S. H.; and Halgamuge, S., 2019. Self Organizing Nebulous Growths for Robust and Incremental Data Visualization. (12 2019). http://arxiv.org/abs/1912.04896. (cited on pages 27, 97, 108, 110, and 115)

Shaw, G. T.-W.; Pao, Y.-Y.; Wang, D.; Tzun-Wen Shaw, G.; Pao, Y.-Y.; and Wang, D., 2016. MetaMIS: a metagenomic microbial interaction simulator

based on microbial community profiles. *BMC Bioinformatics*, 17, 1 (2016), 488. doi:10.1186/s12859-016-1359-0. http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1359-0https://docs.google.com/document/d/1I4Q0Zf50GxbXJOYqpY0Qr3ADNoLpDJckqlIJwoTpfK4/edit. (cited on pages 3, 13, 31, 33, 35, 37, 58, 62, and 77)

Shaw, M. E.; Sachdev, P. S.; Abhayaratna, W.; Anstey, K. J.; and Cherbuin, N., 2018. Body mass index is associated with cortical thinning with different patterns in mid- and late-life. *International Journal of Obesity*, 42, 3 (3 2018), 455–461. doi:10.1038/ijo.2017.254. (cited on page 5)

Shiu, J. H.; Keshavmurthy, S.; Chiang, P. W.; Chen, H. J.; Lou, S. P.; Tseng, C. H.; Justin Hsieh, H.; Allen Chen, C.; and Tang, S. L., 2017. Dynamics of coral-associated bacterial communities acclimated to temperature stress based on recent thermal history. *Scientific Reports*, 7, 1 (2017), 1–13. doi:10.1038/s41598-017-14927-3. http://dx.doi.org/10.1038/s41598-017-14927-3. (cited on page 48)

Singh, R. K.; Chang, H.-W.; Yan, D.; Lee, K. M.; Ucmak, D.; Wong, K.; Abrouk, M.; Farahnik, B.; Nakamura, M.; Zhu, T. H.; Bhutani, T.; and Liao, W., 2017. Influence of diet on the gut microbiome and implications for human health. *Journal of Translational Medicine*, 15 (2017), 73. doi:10.1186/s12967-017-1175-y. https://translational-medicine.biomedcentral.com/track/pdf/10.1186/s12967-017-1175-y. (cited on pages 30 and 48)

Smith, A. A.; Vollrath, A.; Bradfield, C. A.; and Craven, M., 2009. Clustered alignments of gene-expression time series data. In *Bioinformatics*, vol. 25, i119–i1127. Oxford Academic. doi:10.1093/bioinformatics/btp206. https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp206. (cited on pages xvi, 21, and 63)

Smits, S. A.; Leach, J.; Sonnenburg, E. D.; Gonzalez, C. G.; Lichtman, J. S.; Reid, G.; Knight, R.; Manjurano, A.; Changalucha, J.; Elias, J. E.; Dominguez-Bello, M. G.; and Sonnenburg, J. L., 2017. Seasonal cycling in the gut microbiome of the Hadza hunter-gatherers of Tanzania. *Science*, 357, 6353 (8 2017), 802–806. doi:10.1126/science.aan4834. http://www.ncbi.nlm.nih.gov/pubmed/28839072http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5891123http://www.sciencemag.org/lookup/doi/10.1126/science.aan4834. (cited on pages 30 and 48)

Somani, J.; Ramchandran, S.; and Lähdesmäki, H., 2020. A personalised approach for identifying disease-relevant pathways in heterogeneous diseases. *npj Systems Biology and Applications*, 6, 1 (12 2020). doi:10.1038/s41540-020-0130-3. https://pubmed.ncbi.nlm.nih.gov/32518234/. (cited on pages xvi, 21, and 64)

Song, L.; Kolar, M.; and Xing, E. P., 2009a. KELLER: Estimating time-varying interactions between genes. In *Bioinformatics*, vol. 25, i128–i136. Narnia. doi:10.1093/bioinformatics/btp192. https://academic.oup.com/bioinformatics/

article-lookup/doi/10.1093/bioinformatics/btp192. (cited on pages 14, 50, 52, and 112)

Song, L.; Kolar, M.; and Xing, E. P., 2009b. Time-varying dynamic Bayesian networks. Technical report. http://www.cs.cmu.edu/~epxing/papers/2009/song_kolar_xing_nips09.pdf. (cited on page 14)

Song, S. D.; Acharya, K. D.; Zhu, J. E.; Deveney, C. M.; Walther-Antonio, M. R. S.; Tetel, M. J.; and Chia, N., 2020. Daily Vaginal Microbiota Fluctuations Associated with Natural Hormonal Cycle, Contraceptives, Diet, and Exercise. *mSphere*, 5, 4 (7 2020). doi:10.1128/msphere.00593-20. http://msphere.asm.org/. (cited on page 60)

Stein, R. R.; Bucci, V.; Toussaint, N. C.; Buffie, C. G.; and Rätsch, G., 2013. Ecological Modeling from Time-Series Inference: Insight into Dynamics and Stability of Intestinal Microbiota. *PLoS Comput Biol*, 9, 12 (2013), 1003388. doi:10.1371/journal.pcbi.1003388. www.ploscompbiol.org. (cited on pages 33 and 49)

Steinway, S. N.; Biggs, M. B.; Loughran Jr, T. P.; Papin, J. A.; Albert, R.; and Jr, L. T., 2015. Inference of Network Dynamics and Metabolic Interactions in the Gut Microbiome. *PLOS Computational Biology | R*, 11, 6 (2015), 1004338. doi:10.1371/journal.pcbi.1004338. http://www.nigms.nih.gov/Research/. (cited on pages 12, 31, and 32)

Stewart, C. J.; Ajami, N. J.; O'Brien, J. L.; Hutchinson, D. S.; Smith, D. P.; Wong, M. C.; Ross, M. C.; Lloyd, R. E.; Doddapaneni, H. V.; Metcalf, G. A.; Muzny, D.; Gibbs, R. A.; Vatanen, T.; Huttenhower, C.; Xavier, R. J.; Rewers, M.; Hagopian, W.; Toppari, J.; Ziegler, A. G.; She, J. X.; Akolkar, B.; Lernmark, A.; Hyoty, H.; Vehik, K.; Krischer, J. P.; and Petrosino, J. F., 2018a. Temporal development of the gut microbiome in early childhood from the TEDDY study. *Nature*, 562, 7728 (10 2018), 583–588. doi:10.1038/s41586-018-0617-x. https://doi.org/10.1038/s41586-018-0617-x. (cited on page 96)

Stewart, C. M.; Kothari, P. D.; Mouliere, F.; Mair, R.; Somnay, S.; Benayed, R.; Zehir, A.; Weigelt, B.; Dawson, S. J.; Arcila, M. E.; Berger, M. F.; and Tsui, D. W., 2018b. The value of cell-free DNA for molecular pathology. doi:10.1002/path.5048. http://www.ncbi.nlm.nih.gov/pubmed/29380875http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6656375. (cited on page 25)

Takens, F., 1981. Detecting strange attractors in turbulence. In *Dynamical systems and turbulence, Warwick 1980*, 366–381. Springer. (cited on page 26)

Tan, L.; Chen, J.-J.; Zheng, B.; and Ouyang, Y., 2016. Exploring Market State and Stock Interactions on the Minute Timescale. (2016). doi:10.1371/journal.pone.0149648. http://www.wind.com.cn. (cited on pages 50, 52, and 112)

Tavenard, R.; Faouzi, J.; Vandewiele, G.; Divo, F.; Androz, G.; Holtz, C.; Payne, M.; Yurchak, R.; Russwurm, M.; Kolar, K.; and Woods, E., 2020. Tslearn, A

Machine Learning Toolkit for Time Series Data. *Journal of Machine Learning Research*, 21, 118 (2020), 1–6. http://jmlr.org/papers/v21/20-091.html. (cited on page 91)

Teknomo, K., 2006. K-means clustering tutorial. *Medicine*, 100, 4 (2006), 3. (cited on page 22)

Teodorescu, H. N., 2012. Characterization of nonlinear dynamic systems for engineering purposes - A partial review. doi:10.1080/03081079.2012.725538. https://www.tandfonline.com/doi/abs/10.1080/03081079.2012.725538. (cited on page 26)

Thaiss, C. A.; Zeevi, D.; Levy, M.; Zilberman-Schapira, G.; Suez, J.; Tengeler, A. C.; Abramson, L.; Katz, M. N.; Korem, T.; Zmora, N.; Kuperman, Y.; Biton, I.; Gilad, S.; Harmelin, A.; Shapiro, H.; Halpern, Z.; Segal, E.; and Elinav, E., 2014. Transkingdom control of microbiota diurnal oscillations promotes metabolic homeostasis. *Cell*, 159, 3 (10 2014), 514–529. doi:10.1016/j.cell.2014.09.048. (cited on pages 30 and 48)

Tiede, J.; Wemheuer, B.; Traugott, M.; Daniel, R.; Tscharntke, T.; Ebeling, A.; and Scherber, C., 2016. Trophic and Non-Trophic Interactions in a Biodiversity Experiment Assessed by Next-Generation Sequencing. *PLOS ONE*, 11, 2 (2 2016), e0148781. doi:10.1371/journal.pone.0148781. https://dx.plos.org/10.1371/journal.pone.0148781. (cited on page 48)

Toth, E., 2013. Catchment classification based on characterisation of streamflow and precipitation time series. *Hydrology and Earth System Sciences*, 17, 3 (2013), 1149–1159. doi:10.5194/hess-17-1149-2013. (cited on page 26)

Tridico, S. R.; Murray, D. C.; Addison, J.; Kirkbride, K. P.; and Bunce, M., 2014. Metagenomic analyses of bacteria on human hairs: A qualitative assessment for applications in forensic science. *Investigative Genetics*, 5, 1 (12 2014), 16. doi:10.1186/s13323-014-0016-5. http://www.investigativegenetics.com/content/5/1/16. (cited on pages 25 and 97)

Tsai, K.-N.; Lin, S.-H.; Liu, W.-C.; and Wang, D., 2015. Inferring microbial interaction network from microbiome data using RMN algorithm. *BMC Systems Biology*, 9, 1 (2015), 54. doi:10.1186/s12918-015-0199-2. http://www.biomedcentral.com/1752-0509/9/54. (cited on pages 12, 31, and 37)

Tsimring, L. S., 2014. Noise in biology. doi:10.1088/0034-4885/77/2/026601. /pmc/articles/PMC4033672//pmc/articles/PMC4033672/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4033672/. (cited on page 5)

Van Der Maaten, L., 2009. Learning a parametric embedding by preserving local structure. In *Journal of Machine Learning Research*, vol. 5, 384–391. (cited on pages 27 and 97)

Van Der Maaten, L. and Hinton, G., 2008. Visualizing Data using t-SNE. Technical report. (cited on pages 58, 91, and 115)

Vidanaarachchi, R.; Shaw, M.; and Halgamuge, S., 2019. Exploring Computational Inference of Microbial Interactions and their Dynamics. In *2019 IEEE 14th International Conference on Industrial and Information Systems: Engineering for Innovations for Industry 4.0, ICIIS 2019 - Proceedings*, 28–33. Institute of Electrical and Electronics Engineers Inc. doi:10.1109/ICIIS47346.2019.9063336. (cited on pages 6, 8, and 47)

Vidanaarachchi, R.; Shaw, M.; Tang, S. L.; and Halgamuge, S., 2020. IMPARO: Inferring Microbial Interactions through Parameter Optimisation. *BMC Molecular and Cell Biology*, (2020). doi:10.1186/s12860-020-00269-y. (cited on pages 6, 8, 13, 29, 35, 42, 45, 51, 52, 53, 56, 57, 58, 60, 62, 65, and 87)

Wang, T.; Goyal, A.; Dubinkina, V.; and Maslov, S., 2019. Evidence for a multilevel trophic organization of the human gut microbiome. *bioRxiv*, (4 2019), 603365. doi:10.1101/603365. https://www.biorxiv.org/content/10.1101/603365v1. (cited on page 48)

Wangersky, P. J., 1978. Lotka-Volterra Population Models. *Annual Review of Ecology and Systematics*, 9 (1978), 189–218. http://www.jstor.org.virtual.anu.edu.au/stable/2096748. (cited on pages 49 and 50)

Wilkinson, C.; Woodruff, S. D.; Brohan, P.; Claesson, S.; Freeman, E.; Koek, F.; Lubker, S. J.; Marzin, C.; and Wheeler, D., 2011. Recovery of logbooks and international marine data: The RECLAIM project. *International Journal of Climatology*, 31, 7 (6 2011), 968–979. doi:10.1002/joc.2102. http://doi.wiley.com/10.1002/joc.2102. (cited on pages 16 and 97)

Wilson, S. A., 2009. THE NATURE OF VIRTUE A Classical Confucian Contribution to Contemporary Ethical Reflection. 23, 2 (2009), 263–289. (cited on pages 23 and 82)

Yang, R.; Gao, R.; Cui, S.; Zhong, H.; Zhang, X.; Chen, Y.; Wang, J.; and Qin, H., 2019. Dynamic signatures of gut microbiota and influences of delivery and feeding modes during the first 6 months of life. *Physiological Genomics*, 51, 8 (2019), 368–378. doi:10.1152/physiolgenomics.00026.2019. https://pubmed.ncbi.nlm.nih.gov/31226006/. (cited on pages 25 and 97)

Yang, S. H.; Tseng, C. H.; Huang, C. R.; Chen, C. P.; Tandon, K.; Lee, S. T.; Chiang, P. W.; Shiu, J. H.; Chen, C. A.; and Tang, S. L., 2017. Long-term survey is necessary to reveal various shifts of microbial composition in corals. *Frontiers in Microbiology*, 8, JUN (2017), 1–11. doi:10.3389/fmicb.2017.01094. (cited on page 48)

Yokobayashi, Y., 2019. Applications of high-throughput sequencing to analyze and engineer ribozymes. *Methods*, (2 2019). doi:10.1016/J.YMETH.2019.02.001. https://www.sciencedirect.com/science/article/pii/S104620231830272X. (cited on page 31)

Zheng, P.; Yang, J.; Li, Y.; Wu, J.; Liang, W.; Yin, B.; Tan, X.; Huang, Y.; Chai, T.; Zhang, H.; Duan, J.; Zhou, J.; Sun, Z.; Chen, X.; Marwari, S.; Lai, J.; Huang, T.; Du, Y.; Zhang, P.; Perry, S. W.; Wong, M. L.; Licinio, J.; Hu, S.; Xie, P.; and Wang, G., 2020. Gut Microbial Signatures Can Discriminate Unipolar from Bipolar Depression. *Advanced Science*, 7, 7 (4 2020), 1902862. doi:10.1002/advs.201902862. https://onlinelibrary.wiley.com/doi/abs/10.1002/advs.201902862. (cited on page 25)

Zhuang, Y.; Chai, J.; Cui, K.; Bi, Y.; Diao, Q.; Huang, W.; Usdrowski, H.; and Zhang, N., 2020. Longitudinal investigation of the gut microbiota in goat kids from birth to postweaning. *Microorganisms*, 8, 8 (7 2020), 1–18. doi:10.3390/microorganisms8081111. https://www.mdpi.com/2076-2607/8/8/1111/htmhttps://www.mdpi.com/2076-2607/8/8/1111. (cited on page 19)

Zou, Y.; Donner, R. V.; Marwan, N.; Donges, J. F.; and Kurths, J., 2019. Complex network approaches to nonlinear time series analysis. doi:10.1016/j.physrep.2018.10.005. (cited on pages 26 and 97)