



UAI
UNIVERSIDAD ADOLFO IBÁÑEZ
FACULTAD DE INGENIERÍA Y CIENCIAS

UNIVERSIDAD ADOLFO IBÁÑEZ

DOCTORAL THESIS

**Contributions to Metagenomics:
Classification, Dynamics and Applications**

Author:
Dante Travisany

Committee:
Eric GOLES
Alejandro MAASS
Gonzalo RUZ
Pedro MONTEALEGRE
Pablo MARQUET

*A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy*

April 15, 2019

Abstract

A microbiome is a community of microorganisms that inhabit a particular environment like soils, oceans or the human body. Those communities are made of a trillion of microorganisms divided into a dozen or even hundreds of different species which interact among them. The community behaves as a complex adaptive system and fluctuates in response to changes in environmental factors such as acidity, pressure, temperature or, when are host related it is sensible to perturbations like antibiotics, changes in the diet and lifestyle factors. The recovery of the genetic material of the microbial communities directly from their environment, the classification and study of these samples is called Metagenomics. The digitalization of the genetic material is done by high throughput sequencing technologies, such as Illumina which shreds the DNA into fragments called reads.

In the first part of this work we propose a alignment-free method capable of assign taxa to each read in a metagenomic sample by analyzing the statistical properties of the reads. Given an environment, we collect genomes from public available databases and generate synthetic genomic fragments libraries. Then, statistics of k-mer frequencies are computed and stored in an environment-associated dataset used to build a robust machine learning procedure based on multiple CART trees.

In the second part we study the dynamics of a gut microbiome under antibiotic perturbation and *Clostridium difficile* infection (CDI). Here, interactions play a key role in the development of the disease. Using a previously Boolean network model for CDI we demonstrate that this model is in fact a threshold Boolean network (TBN). Once the TBN model is set, we further explore the space of possible interactions generating an evolutionary algorithm to identify alternative TBNs. Allowing the construction of a neutral space conformed by a set of models that differ in their interactions, but share the final community states of the gut microbiome under antibiotic perturbation and CDI. We organize the resulting TBNs into clusters that share similar dynamic behaviors and the most relevant interactions are identified. Finally, we discuss how these interactions can either affect or prevent CDI.

In the third part we examined the microbial community across a transect of few kilometers long, where there is a remarkable abiotic variations like pH, temperature and humidity gradient, with acidic soils, to name a few. We studied the compositional structure of the community and constructed two co-occurrence networks representing two sections that divided the transect. Using network analysis, we examined changes in putative ecological interactions among microbial Operational Taxonomic Units (OTUs), as well as their associations to physicochemical and nutritional variables. Network comparisons allowed us to examine the nature of the ecological rearrangements that take place in the microbial community when facing contrasting environments. L-GRAAL, the graph alignment method we used provides a comprehensive way to understand topological shifts among members from two networks. We show here that this method provides a glimpse into the nature of the changes in microbial communities that can foster resistance and resilience to contrasting environmental conditions.

Acknowledgements

I would like to express my sincere gratitude to Eric for his continuous support and constant bullying. Also for his patience, motivation and charisma. His guidance construct outstanding researchers, but more important is the confident environment generated by his presence.

I am profoundly grateful to Alejandro for his example of rightness, methodology and professionalism. He is an exemplary leader.

I would like to thank Pedro and Patricia for their warm welcome in Orleans and their selfless aid and help during my internship there.

I greatly appreciate to Dr. Gonzalo Ruz and Dr. Pablo Marquet for being interested in my work and being part of the committee.

I gratefully acknowledge the funding received towards the Centro de Modelamiento Matemático CONICY PIA 170001 and BASAL PBF003, to the Centro para la Regulación del Genoma FONDAP 15090007, the Adolfo Ibañez excellence scholarship and CONICYT PFCHA/Beca Doctorado Nacional 2015/FOLIO 21150895.

Many thanks to my lab mates, María Paz, Vicente, Ricardo, Natalia for their time, good vibes and recomendations. Many thanks also to the former lab members Alex, Marko and Karina.

Thanks to my family, Carolina for her unconditional support; to Sofía, Florencia and Dante for their love and joy.

Thanks to my parents for being there always.

Contents

Abstract	iii
Acknowledgements	v
1 Introduction	1
1.1 Microbiome and Metagenomics	1
1.1.1 Binning	2
Unsupervised Methods	2
Supervised Methods	2
Semi-supervised Methods	3
Classification Trees	3
1.2 Network Inference	3
1.2.1 Co-occurrence Networks	5
1.3 Dynamical Models	7
1.3.1 Continuous Models	7
1.4 Discrete Models	7
1.4.1 Boolean Networks	7
1.5 The concept of Neutral Space	9
1.6 Contributions to Microbiome Analysis and Modelling	10
2 Predicting the Metagenomics Content Using Multiple CART Trees	15
2.1 Insights	15
2.1.1 Organization of the Chapter	16
2.2 Methods	16
2.2.1 Metagenomics data acquisition and processing	16
2.2.2 Building multiple CART trees	18
2.2.3 Implementation	20
2.2.4 DNA-patterns importance	20
2.2.5 Validation and comparison with other binning methods	20
2.3 Results	20
2.4 Conclusions and Discussion	23
3 Generation and Robustness of Boolean Networks to Model CDI	27
3.1 Insights	27
3.2 Organization of the Chapter	27
3.3 Human Gut Microbiome	28
3.3.1 Boolean Model of the Gut Microbiome	28
3.3.2 <i>Clostridium difficile</i> Infection Model	28
3.4 The CDI Model as a Threshold Boolean Network	31

3.5	Analysis of the <i>Clostridium difficile</i> Infection Model	34
3.6	Neutral Space	36
3.6.1	Algorithm to Explore the Neutral Space	38
3.6.2	Clustering of Threshold networks	39
3.7	Neutral Space Analysis	40
3.7.1	General Insights Across the Sampling of the Neutral Space	40
3.7.2	Interactions	40
3.7.3	Dynamics	40
3.7.4	Classes	42
	Class EC_1	43
	Class EC_2	44
	Class EC_3	44
	Class EC_4	44
	Class EC_5	45
	Class EC_6	45
	Class EC_7	45
	Class EC_8	45
3.7.5	Lessons	46
3.8	Illustration of the Classes	48
3.9	Conclusions	56
3.10	Availability	56
3.11	Additional Information	57
3.11.1	Clustering Analysis Related Figures	57
3.11.2	Classes Features	59
3.11.3	Interactions across the Classes	61
3.11.4	Higher Degree TBNs	67
4	Structure and co-occurrence patterns in microbial communities	73
4.1	Insights	73
4.2	Organization of the Chapter	74
4.3	Methods	74
4.3.1	Background and site description	74
4.3.2	Sample collection	75
4.3.3	Environmental and soil physicochemical measurements	76
4.3.4	Soil DNA extraction and sequencing	76
4.3.5	Processing of Illumina sequence data	77
4.3.6	Sequence analysis and taxonomic prediction using Greengenes DB	77
4.3.7	Microbial diversity and composition and Multivariate analyses	77
4.3.8	Co-occurrence microbial networks	78
4.3.9	Topological graph alignment of co-occurrence networks	79
4.3.10	Functional structure of TLT	80
4.4	Results	80
4.4.1	Talabre-Lejía transect (TLT) exhibits environmental variability	80
4.4.2	TLT microbial diversity is modulated by environmental stressors	89
4.4.3	TLT co-occurrence microbial networks from Section 1 and Section 2 are dissimilar	93
4.4.4	Nitrogen metabolism highlights in the core bacterial community	94

4.4.5	Impact of environmental variables on OTUs taxonomic composition	95
4.5	Discussion	96
5	Perspectives	103
5.1	Atacama Desert	103
5.2	Tara Oceans	103
5.3	Wine Microbiomes	104