TÉCNICO LISBOA

Machine learning assisted identification of bioindicators in metagenomics

A study over transcription factor genes in a nitrate reducing microcosm

Afonso de Oliveira Santos Goulart

Thesis to obtain the Master of Science Degree in

Biological Engineering

Supervisors: Prof. Dr. Ana Luísa Nobre Fred
Prof. Dr. Ulisses Nunes da Rocha

Examination Committee

Chairperson: Prof. Dr. Nuno Gonçalo Pereira Mira
Supervisor: Prof. Dr. Ana Luísa Nobre Fred
Member of the Committee: Prof. Dr. Rodrigo da Silva Costa

July 2021

Preface

The work presented in this thesis was performed at Helmholtz Centre for Environmental Research (Leipzig, Germany), during the period of September 2020 - February 2021, under the supervision of Dr. Ulisses Rocha. The thesis was co-supervised at Instituto Superior Técnico by Prof. Ana Luísa Nobre Fred.

I declare that this document is an original work of my own authorship and that it fullfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Acknowledgments

I would like to thank the entire MDS team at UFZ, namely my supervisor Ulisses for believing in me, giving me a chance to work in such a great group and helping me throughout the way, Rodolfo for all your invaluable technical support, Felipe for all your feedback and help with code and writing, João for all your help with the PredicTF tool and Marcos for helping me understand Linux. I must also thank my IST supervisor Prof. Ana Fred for accompanying me and giving me feedback throughout the writing process.

I also want to thank all the great people that I have met in MEBiol, as I could never have made it without the help of a lot of you. Inês for being the best lab partner all these years and putting up with me. Rafael for all the board game nights and your eternal patience in reading over this thesis multiple times. Henrique, Kiko and Caldas for all the good times we shared.

Outside of MEBiol, I have to thank all my friends, but especially, Picci, Tiggy, Nogueira and Afonso, for always being the best friend group a person could ask for. I have to thank Sofia, not just for being there for me when I needed you throughout this thesis, but also for all the invaluable feedback your provided.

Finally, I want to thank my family for always supporting me in my endeavors, especially my parents, who always supported me and provided me with what I needed to tackle this challenge.

Abstract

Transcription factors are proteins essential in the control of gene expression and a proper mapping and profiling of these proteins could prove invaluable for the understanding and control of gene regulation in a microbiome. Directly analyzing the abundance and presence of different transcription factor genes in metagenomic samples may be a tool that can help provide insight on the external factors that drive transcription factor abundance.

The metagenomic assemblies originating from a benzene-degrading nitrate-reducing succession experiment were analyzed with the newly developed PredicTF tool in order to identify which transcription factors were present in the metagenome. This data was used to conduct a diversity analysis which detected shifts in the abundance of different transcription factor families over time. Machine learning algorithms, such as Random Forest, and statistical tests were then employed to identify potential bioindicators among the different transcription factor families, with two different families being identified as potential bioindicators for benzene degradation/nitrate reduction. This work demonstrates that this technique has potential for determining the impact of external factors on biological samples, motivating further exploration of the proposed approach in broader datasets.

Keywords

Random Forest; Transcription Factor; Microbial Diversity; Machine Learning.

Resumo

Fatores de transcrição são proteínas essenciais no controlo da expressão génica e um mapeamento e caracterização destas proteínas poderá ser essencial na descrição e controlo da regulação génica num microbioma. Analisar diretamente a abundância e presença de diferentes fatores de transcrição poderá ser uma ferramenta que ajude a melhor entender o impacto de fatores externos na expressão de fatores de transcrição.

As montagens metagenómicas obtidas de amostras provenientes de uma experiência de sucessão com um microcosmo com degradação de benzeno e redução de nitrato foram analisadas com a nova ferramenta PredicTF, de forma a identificar os fatores de transcrição presentes nos metagenomas. Estes dados foram usados de forma a realizar uma análise de diversidade, que detetou alterações na abundância de diferentes famílias de fatores de transcrição, com 2 famílias sendo identificadas como possíveis bio-indicadores. Este trabalho demonstra que esta técnica tem o potencial de estudar o impacto de fatores externos em amostras biológicas, motivando estudo dos métodos utilizados num dataset maior.

Palavras Chave

Random Forest; Fator de transcrição; Diversidade Microbiana; Aprendizagem Automática.

Contents

1 Introduction **1**

1.1 Context and Motivation . . . 3

1.2 Objectives and proposed framework . . . 3

1.3 Thesis outline . . . 4

2 Basic Concepts and State-of-the-art **5**

2.1 Biology Background . . . 7

2.1.1 Gene regulation . . . 7

2.1.2 Transcription Factors . . . 8

2.1.2.A Transcriptional regulatory networks . . . 8

2.1.2.B Transcription factors for community modeling . . . 9

2.2 Metagenomics . . . 10

2.2.1 Sequencing and assembly . . . 11

2.3 Diversity analysis . . . 13

2.3.1 Alpha Diversity . . . 13

2.3.2 Beta Diversity . . . 13

2.3.3 Library Rarefaction . . . 14

2.4 Machine Learning Background . . . 14

2.4.1 Machine learning in genomic prediction . . . 14

2.4.2 Supervised learning . . . 15

2.4.3 Feature selection/representation . . . 16

2.4.4 Decision trees . . . 16

2.4.5 Random Forest . . . 18

2.4.5.A Performance evaluation . . . 19

2.4.6 Neural Networks . . . 20

2.4.7 Deep learning . . . 20

2.4.8 PredicTF . . . 21

3 Materials and Methods **23**
3.1 Microcosm setup and sampling . . . 25
3.2 Rarefaction and assembly . . . 26
3.3 PredicTF . . . 26
3.4 Diversity Analysis . . . 27
3.4.1 Alpha and Beta diversity . . . 27
3.4.2 Bray-Curtis dissimilarity index . . . 29
3.5 Identifying bioindicators using machine learning . . . 29
4 Results and Discussion **31**
4.1 PredicTF output . . . 33
4.2 Diversity analysis . . . 35
4.2.1 Alpha diversity . . . 35
4.2.2 Beta Diversity . . . 37
4.2.3 Variance analysis . . . 39
4.3 Bioindicators . . . 42
4.3.1 Random Forest . . . 42
4.3.2 Statistical Test . . . 44
5 Conclusions and future work **51**
A Annex A **67**
A.1 NMDS and PCoA plots . . . 68
A.2 Subfamily Variance analysis . . . 72
A.3 RF Attempts . . . 73

List of Figures

2.1 RNA-poymerase . . . 7
2.2 Transcription Factors in gene expression . . . 9
2.3 Schematic of a computational model of a global transcriptional regulatory network for *E. faecalis* . . . 10
2.4 Reference and de-novo assembly . . . 12
2.5 Schematic of the Machine Learning process . . . 16
2.6 Example of a decision tree. . . . 17
2.7 Overview of a Random Forest model . . . 18
2.8 Example of Bagging . . . 19
2.9 Model of a feed-forward neural network . . . 20
2.10 Comparison between a classical neural network and deep-learning neural network . . . 21
2.11 Summary of training, testing and validation of the PredicTF tool . . . 22

3.1 Schematic description of the proposed methodology. . . . 25

4.1 Absolute abundance of TF families . . . 34
4.2 Relative abundance of TF families . . . 34
4.3 Alpha Diversity of the TF sub-families . . . 35
4.4 Alpha Diversity of the TF families . . . 35
4.5 PCoA graph using relative abundance of TF subfamilies . . . 37
4.6 PCoA graph using absolute abundance of TF subfamilies . . . 38
4.7 PCoA graph using relative abundance of TF families . . . 38
4.8 PCoA graph using absolute abundance of TF families . . . 39
4.9 Bray-Curtis dissimilarity index among the absolute abundance of samples taken in the same day . . . 41
4.10 Bray-Curtis dissimilarity index among the relative abundance of samples taken in the same day . . . 41

4.11 GINI score for Transcription Factor families with 5 variables per tree 43
4.12 GINI score for Transcription Factor families with 20 variables per tree 43
4.13 Relative abundance of NtrC.DctD and LysR 45
4.14 Absolute abundance of the PILR transcription factor subfamily over 5 different sampling days. 47

A.1 PCoA of relative abundance of TF families 68
A.2 PCoA of absolute abundance of TF families 69
A.3 NMDS of absolute abundance of TF families 69
A.4 NMDS of absolute abundance of TF families with time groups 70
A.5 NMDS of relative abundance of TF families 70
A.6 NMDS of relative abundance of TF families with time groups 71
A.7 PCoA of relative abundance of TF families 71
A.8 PCoA of absolute abundance of TF families 72
A.9 Bray-Curtis dissimilarity index among the relative abundance of TF subfamilies in samples taken in the same day. 72
A.10 Bray-Curtis dissimilarity index among the absolute abundance of TF subfamilies in samples taken in the same day. 73

List of Tables

4.1 Pairwise t-test on the Alpha diversity results for TF sub-families 36
4.2 Pairwise t-test on the Alpha diversity results for TF families 36
4.3 Pairwise t-test on the Bray Curtis dissimilarity results for the absolute abundance data . . 40
4.4 Random Forest Confusion Matrix 42
4.5 Least Square Mean analysis of relevant TF families 44
4.6 Least Square Mean analysis of NtrC.DctD subfamilies 46

A.1 Sample groupings used for the building of RFs with 5 variables tested per tree and their OOB error rate 73

Acronyms

bp Base Pairs

DBD DNA-binding domains

DNA Deoxyribonucleic Acid

HMN 2,2,4,4,6,8,8-Heptamethylnonane

LS Means Least Square Means

ML Machine Learning

NMDS Non-Metric Dimensional Scaling

NN Neural Networks

OOB Out-of-bag

PCoA Principal Coordinate analysis

PERMANOVA Permutational multivariate analysis of variance

RF Random Forest

RNA RiboNucleic Acid

TFs Transcription Factors

1

Introduction

Contents

1.1 Context and Motivation 3

1.2 Objectives and proposed framework 3

1.3 Thesis outline 4

1.1 Context and Motivation

The inspiration for this work is the combination of three main factors: The progress and availability of new tools to determine the abundance of transcription factors (TFs); the interest in better modeling of microbial communities' transcription networks; and the question of whether the regulatory and expression patterns of a microbial community are related to the abundance of transcription factors within that community.

Given that one of the major objectives in microbiome manipulation for the ecology and biotechnology fields is to control microbiome function and expression [1] [2], and that TFs are essential in the control of gene expression [3], a proper mapping and profiling of TFs could allow for better understanding and control of gene regulation in a microbiome. Improving our understanding of these regulatory networks and their many components permits a superior level of manipulation, control, and monitoring of these microbiomes. This could prove valuable in many facets, providing benefits in the biotechnology fields, but also allowing better prevention of potential environmental hazards in communities, given that understanding the specific methods of regulation of a microbiome would allow us to better predict the environmental impact of external factors introduced to this environment [1].

When studying TFs, there is a lot of work published on their effect on gene expression [3]. However, research on which external factors directly impact TF abundance and how these factors integrate transcription regulatory networks is often overlooked. With the advent on new tools geared specifically towards the detection of TFs such as PredicTF [2], and with the expansion of TF databases like UniProt [4] and CollectTF [5], analysis of TF abundance and the environmental stresses impacting it can be more properly undertaken.

1.2 Objectives and proposed framework

The main objective of this work is to determine the impact of environmental stresses on the abundance of bacterial TFs and whether these differences in abundance levels could be used to distinguish between samples at different time points of a succession experiment, thus providing direct insight into the regulatory networks of the community. In order to achieve this objective, 4 main steps are taken, according to a proposed methodology graphically summarized in fig 3.1: The elaboration of a succession experiment on a bacterial, benzene degrading, nitrate reducing microcosm performed by a collaborator [6], the sequencing of metagenomic data of these samples followed by metagenomic assembly, the use of a tool to predict TFs in the assemblies (PredicTF [2]) and an extensive diversity analysis and search for bioindicators in the TF abundance results, using multiple diversity metrics, statistical tests and machine learning approaches.

1.3 Thesis outline

This document is divided in 5 chapters. Chapter 2 introduces the background and state of the art of underlying concepts and techniques explored in this work , such as transcription factors, metagenomics, diversity analysis and machine learning. In chapter 3 I describe the methodology employed in this work , including the microcosm setup, assembly, the use of PredicTF tool, the diversity analysis and the search for bioindicators. In the following chapter 4, the results of the previous described work are shown and discussed. Finally, in chapter 5 I summarize the main findings of this work and some potential future questions to be addressed are briefly discussed.

2

Basic Concepts and State-of-the-art

Contents

2.1 Biology Background 7
2.2 Metagenomics 10
2.3 Diversity analysis 13
2.4 Machine Learning Background 14

2.1 Biology Background

2.1.1 Gene regulation

For different cells and organisms, the same genomic sequence can present widely different phenotypes. These variations are caused by specific genomic instructions, which stem from transcriptional regulatory pathways. The metabolism of cells and organisms are therefore affected and regulated by transcriptional regulation [3]. This regulation can be positive regulation (activating gene expression), negative regulation (repressing gene expression), and co-regulation (activating or repressing multiple genes at the same time) [3].

In order to commence transcription, RiboNucleic Acid (RNA) polymerase must bind to a promoter, a site in Deoxyribonucleic Acid (DNA) often adjacent to genes, with an affinity for this enzyme. Promoters also contain, or are near , binding sites for transcription factors, which are DNA-binding proteins that can either help recruit, or repel, RNA polymerase [3].

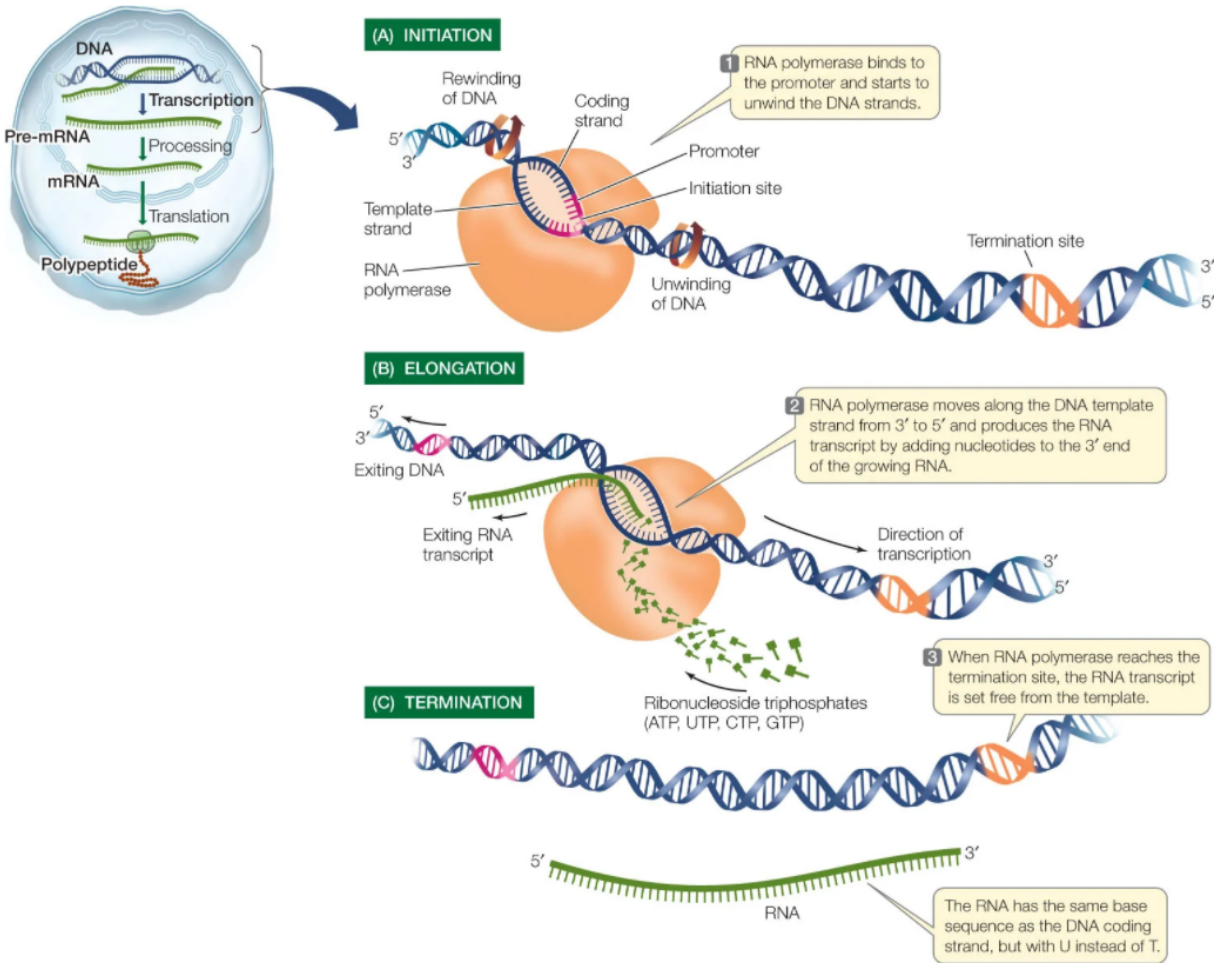

Figure 2.1: RNA-polymerase in DNA transcription [7]

2.1.2 Transcription Factors

Transcription Factors (TFs) are involved in the process of converting/transcribing DNA into RNA and include a wide number of proteins that initiate and regulate the transcription of genes [8]. TFs directly interpret genomes by recognizing specific DNA sequences and controlling chromatin and transcription [9], often functioning as regulators, controlling developmental patterning [10], immune response pathways [11],and drive cell differentiation [12], de-differentiation and trans-differentiation [13]. These TFs can function alone or in combination with other proteins (often other TFs) in order to promote/upregulate (activators) or repress/downregulate (repressors) the recruitment of RNA polymerase in specific genes [14]. RNA polymerase is the enzyme responsible for transcribing DNA into RNA and, therefore, by controlling the recruitment of this enzyme, transcription factors can regulate the expression levels of the genes they target [15] [14]. The process by which this regulation of transcription occurs can be by: directly stabilizing (upregulation) or blocking (downregulation) the binding of RNA polymerase; catalyzing the acetylation (upregulation) or deacetylation (downregulation) [16];recruitment of coactivators and corepressors into the TF DNA complex [17]. An example of how gene expression may be promoted by TFs can be found in fig 2.2.

Although many different proteins also take part in gene regulation such as kinases, coactivators, histone acetyltrasnferases and deacetlyases, TFs are defined by the presence of one or more DNA-binding domains (DBD) [19], a protein domain which recognizes and binds to DNA [20]. In TFs this DBD binds to a specific sequence of DNA adjacent to the gene being regulated, with this sequence being either an enhancer or a promoter [21]. For a large number of promoters and enhancers, a combination of TFs working together allows for the integration and modulation of different signals [22] [23]. Bacterial TFs have also been shown to work together with other DNA-binding proteins with the primary role of sculpting the bacterial folded chromosome [2] [22] [24].

In addition, TFs are known to be associated with or even bioindicators of human diseases such as alzheimer's [25], congenital heart disease [26], Parkinson's [27], among many others.

2.1.2.A Transcriptional regulatory networks

Transcriptional regulatory networks encompass all molecular species and regulatory interactions necessary to accurately describe observed patterns of gene expression [28]. Meaning that any component that can influence if or how a certain gene is expressed is part of this network, which would include genes, transcription factors, promoters, repressors, and many other enzymes and proteins previously mentioned which also take part in gene regulation, as well as all interactions between these different components. These networks are essential in all biological organisms, and understanding them is essential to deciphering the development, functioning and pathological pathways of these organisms [29].

Given the importance of these networks in understanding the biological pathways of organisms and

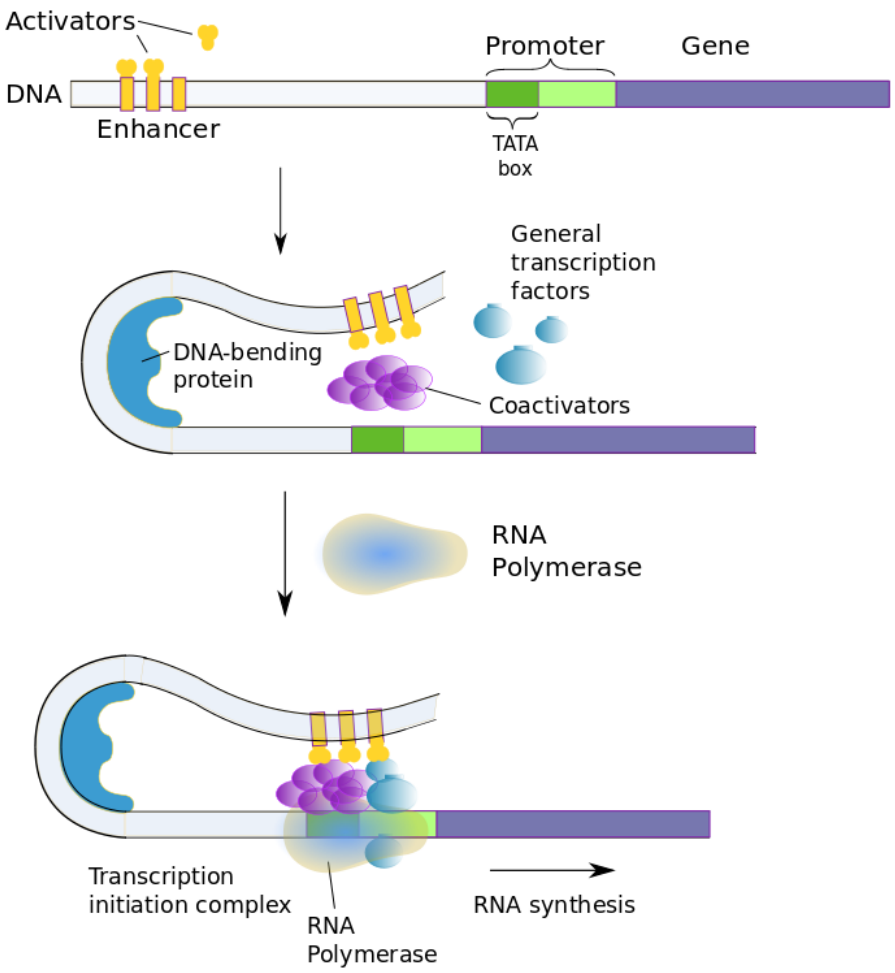

Figure 2.2: Schematic of TF assisted gene expression adapted from [18].

the importance of transcription factors in these networks, a clearer understanding of the external factors affecting TF abundance could provide valuable insight into the metabolism of organisms in communities.

2.1.2.B Transcription factors for community modeling

Recently, research has shown how certain external factors affect expression, namely, that bacterial transcription factors in specific are heavily modulated by their environment, directly responding to factors such as pH levels,oxygen and nutrient concentration, presence of toxins, among others [31]. In addition, work has been done on modeling transcription networks, including those responsible for biofilm formation [32], controlling stromal cell differentiation [33] or handling copper stress [30]. The study of these networks is often associated with the study of the transcription factors present in them, given their integral role.

If the hypothesis that the abundance levels of TFs may be used for community modeling, this would

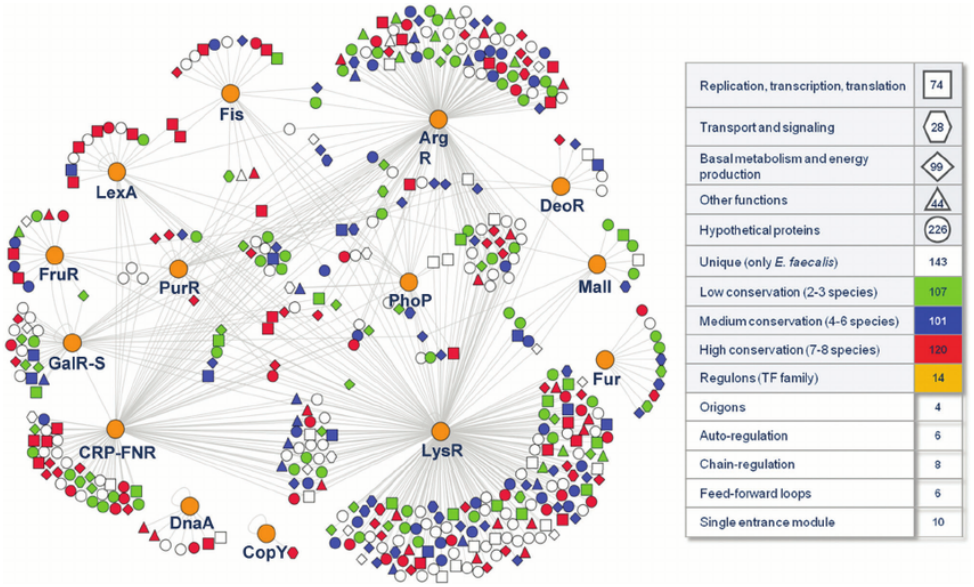

Figure 2.3: Schematic of a computational model of a global transcriptional regulatory network for *Enterococcus faecalis*. Orange circles indicate transcription factor families. Adapted from [30].

prove a very advantageous tool in this field, given that, even if TF abundance alone may not be enough to model a community, it may be used as an extra tool to get a more accurate and complex model of communities and their regulation patterns.

Given the importance of transcription factors in the metabolism of organisms [3], work has been done in an attempt to model the expression of these factors in order to provide more accurate modeling of gene regulatory networks [10] [34] [35] [36]. These works include the study of the importance of specific TF families on plant growth and development [36], determining *Aspergillus fumigatus* transcription factor expression and function during invasion of lungs [35], studies on the transcription factor regulatory networks in order to better comprehend and defend against heat stress [37] and many different environmental conditions [38] in plants.

2.2 Metagenomics

Several methods may be employed to determine the microbial composition of biological samples, with two main methods being considered of great importance, cultivation-dependent and independent approaches, namely, metagenomic sequencing. Although microbial culture is still widely used and the classical method for microbial studies, the majority of species cannot be cultured in a lab, either due to the lack of biological knowledge to culture them or a lack of more advanced laboratory techniques required to simulate the proper culture environment [39] [40]. Contrary to microbial cultures, techniques such as nucleotide sequencing are not affected by the microbial characteristics or environment mimick-

ing technology. Rather, these techniques are limited by the current sequencing technologies available, which are becoming increasingly faster, cheaper and more accurate [41].

Genomics is the study of the complete genetic complement of a given organism which is obtained using high-throughput sequencing of the base pairs of its DNA [42]. Metagenomics, on the other hand, constitutes the sampling of all the genomic sequences of a community of organisms in the same environment [42].

2.2.1 Sequencing and assembly

Currently, the main sequencing strategies for metagenomics involve high-throughput, short read technologies, like Illumina NextSeq and HiSeq, which are able to generate billions of short reads (small fragments of DNA base pairs) in a matter of days [43]. New sequencing techniques are in constant development with some of the most promising technologies including single-molecule sequencing [44], synthetic long reads [45], and Hi-C [46]. Some of these technologies, such as single-molecule sequencing, have been around for a few years but have not seen widespread use, mostly due to complex sample processing requirements and elevated costs [43]. Following DNA sequencing, the obtained reads are assembled in order to reconstruct the genomes of observed microorganisms. Current methods are able to reconstruct DNA segments, including operons, tandem gene arrays, and syntenic blocks [43]. Metagenome assembly is a process used to order the multiple reads previously sequenced, in order to obtain a genome in the correct order. This process is required due to the fact that a lot of current sequencing techniques produce reads that are far shorter than whole genomes, meaning some degree of assembly is required in order to obtain the full picture. These reads are typically 35-1000 Base Pairs (bp) for short reads, while genomes can vary from as low as 2000 bp for bacteriophages, to 100 billion bp for certain eukaryotes [47].

If no reference genomes can be used a *de novo* assembly is required; this type of assembly assumes no *a priori* knowledge of the correct order of sequenced reads and relies solely on read and contig overlap to construct the reference. Using de-novo assembly is also required for a metagenomics analysis, given that the diversity found in complex microbial communities would lead to a lot of chimeric contigs (contigs that combine sequences from more than one genome) if reference genomes were used [48].

A contig is the name given to a set of overlapping reads. When reads share a significantly large enough subset of nucleotide base pairs, they are considered to have been sequenced from the same ”location” within the genome, creating a contig. These contigs contain at least two reads, but may be comprised of more, with a greater number of reads overlapping in a certain region translating to a greater level of confidence in the nucleotide sequence in this given region [49] [50]. At times reads may have to be reversed in order for a matching orientation to be obtained [49].

When multiple contigs are joined, a scaffold is created. Once again, these contigs may have to be

reversed in or to match orientation and, in a scaffold, not all contigs have to overlap all other contigs. Finally, multiple scaffolds are matched and gaps between them filled in order to create chromosomes.

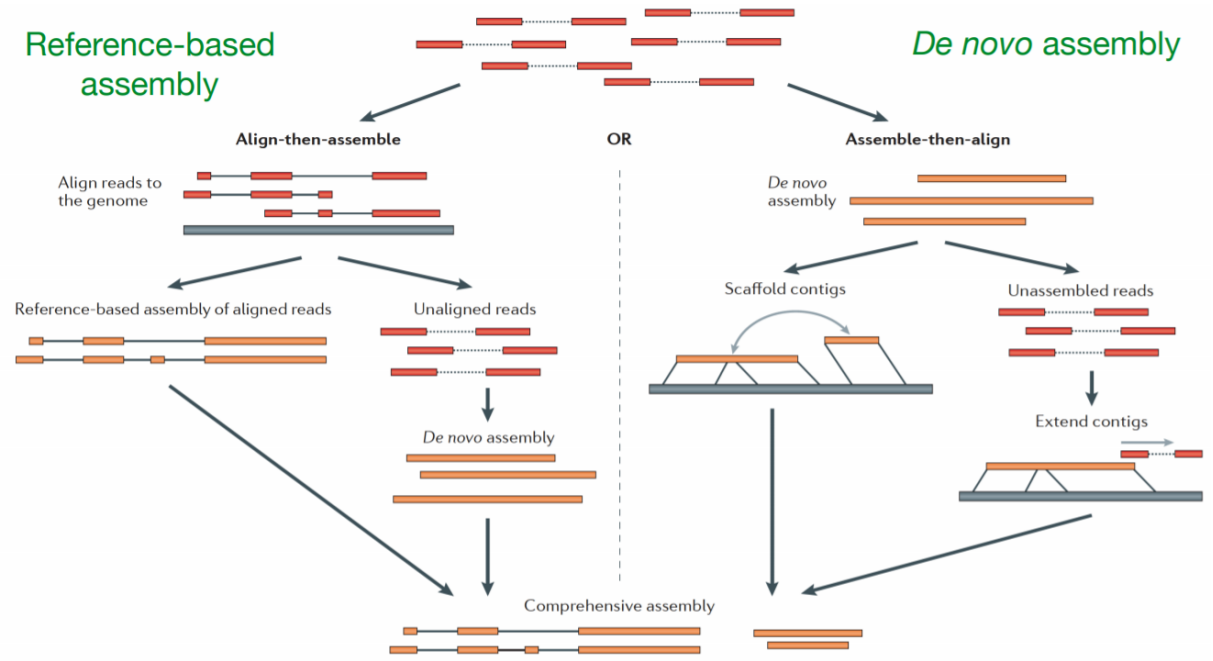

Figure 2.4: Reference based and *de-novo* assembly. Adapted from [51].

Given the complex nature of metagenomic assembly, it is currently considered impossible to have an error-free assembly, regardless of the sequencing method or assembly algorithm used [43]. One of the greatest challenges for genomic assemblies are intragenomic repeats, repeated strands of non-coding DNA, sometimes forming arrays that are megabases long [52]. This problem is only further exacerbated in metagenomic assemblies due to the presence of both intragenomic and intergenomic repeats. In fact, Nagarajan N [53] has shown that the complexity of an assembly is directly related to the ratio between the read length and the length of repeats. Due to these potential errors, in order to verify the quality of assembled metagenomes, validation is required. This validation can be done in one of two ways: de-novo validation; or reference based. De-novo validation, much like de-novo assemblies, relies solely on the original data's features, attempting to ascertain internal inconsistencies within the assembled data [54]. Reference based validation use databases of previously assembled genes and genomes, in order to identify possible differences between the assembly and the reference data [55]. As a general rule, reference based validation is very effective, but can only be used provided a reference genome has already been properly cataloged. Otherwise, a de-novo validation will have to be used. It is also difficult to determine whether certain minor differences between a reference genome and the newly assembled genome are due to legitimate assembly mistakes or simply regular genetic differences between the reference and studied organisms [55].

2.3 Diversity analysis

In ecology, diversity measures the richness and evenness of species in a given community [56]. Given that a comparison between every single possible living form within a given area would not only be unfeasible, but also meaningless, as, for instance, a comparison between the richness of proteins and of plants in a given community would have no meaningful information. Therefore, a precise definition of community must exist. In general, a community is considered to represent all the organisms belonging to a specific taxonomic group within a defined area [56] [57]. To perfectly determine the richness of every single species in a given community, a complete and thorough survey of the entire community would be required, which is impossible in most scenarios [56]. Therefore, richness is usually estimated based on environmental samples, using sampling designs and methods that are specific for the given context/situation in order to avoid biasing of the results, however leading to a decrease in the accuracy of the analysis [58] [59] [60].

Diversity analysis is generally classified as alpha or beta diversity analysis.

2.3.1 Alpha Diversity

Alpha diversity metrics evaluate either the species richness (number of species, taxonomic groups, transcription factor families, etc), evenness (abundance of species in relation to each other) or a combination of both in a single sample or community. Since disturbances to a community often affect its alpha diversity, analyzing this metric in amplicon sequencing data tends to a be a first step when studying differences between environments [61]. The simplest calculation of alpha diversity richness is to count the number of different species in a community (observed species). Evenness, on the other hand, is calculated with methods such as Shannon Index [56]; This index measures species biodiversity in a given community according to the formula in 2.1, with H representing the Shannon Diversity Index (entropy) and p_{i} being the proportion of the entire community made up of species i. The lower the entropy the more evenly distributed the species are [56].

$$H = -\Sigma(p_{\mathrm{i}}{}^{*} \ln(p_{\mathrm{i}})) \tag{2.1}$$

2.3.2 Beta Diversity

Beta diversity metrics are used to quantify differences between different samples within the same community , leading to several techniques being called ”beta diversity” analyses [62] [63], such as (Non-Metric Dimensional Scaling (NMDS) [64], Principal Coordinate analysis (PCoA) [65], etc). As a general rule, an indice relating to compositional heterogeneity between different samples can be classi-

fied as beta diversity [66] [67] [68], although such a broad classification is challenged by certain authors [62] [69].

2.3.3 Library Rarefaction

A common bias introduced in diversity analysis when using sequence data is library size. So much so, that if not accounted for, this bias is often the most determining factor in the result of diversity analyses [70]. To help minimize this bias, rarefaction can be employed as a method for normalizing library sizes across different samples. This method involves randomly discarding reads from samples until a certain threshold is hit. This threshold should be equal to or less than the number of reads of the sample with the least amount of reads (before rarefaction) [71]. Using these subsets of samples, diversity metrics may be more accurate [72]. However, it should be noted that although rarefaction is necessary for comparing absolute abundances in samples, its use with relative abundance analyses of diversity is highly contested, with many authors discouraging the use of rarefaction [61] [73] [74], while pipelines such as QIIME2 and other authors advocate for its use [75] [60].

2.4 Machine Learning Background

2.4.1 Machine learning in genomic prediction

DNA sequencing produces sequences of unknown function. Genomic prediction and annotation, the prediction of genes, proteins and transcription elements in an organism, is the process of identifying functional elements in these sequences and describing their function. This is an essential technique in many scientific fields, such as biomedical engineering, ecology, biotechnology, among others.

Genomic prediction was initially performed using computational annotation of long coding genes on a single genome and the experimental annotation of short regulatory elements and their impacts on a small number of genes [76] [77]. However, with the rise of high-throughput technologies, the data for gene, protein, and transcription elements' interactions has grown significantly and these methods are not as widely used due to very low turn-over rate rendering these techniques incredibly time-consuming [76]. Consequently, network-based techniques using Machine Learning (ML) have gained considerable prominence [77] [78] [79] [80] [81] [82]. These new techniques allow for the population annotation of elements as small as single nucleotides on thousands of different genomes [76]. Machine learning approaches, such as deep learning are being used for identification of previously unknown genes in metagenomic data [79] [82], microRNA genes [78], prediction of gene-disease interactions [81] [80] or identification of transcription factors (TF) [2] and their binding sites [83] [84]. Further examples of applications of ML include the use of techniques like support vector machines to improve

genome annotation of commercially relevant species [85], exploration of supervised learning methods to differentiate between cured and at-risk cancer patients by analyzing gene expression [86], and classification of *Drosophila melanogaster* essential genes using Random Forests, support vector machines and artificial neural networks [87].

The following sections detail the techniques to be explored along the thesis by building on the state-of-the-art and the most promising ML techniques used in this context.

2.4.2 Supervised learning

Overall, machine learning comprises the study of algorithms which use data and experience to improve automatically [88]. More specifically, machine learning algorithms construct models based on training data with the intent of making predictions which they have not been explicitly programmed to [89].

Supervised learning is characterized by having fully labeled datasets, meaning its training data consists of data samples and associated class labels (corresponding to the desired output of the classifier), and the algorithm builds a model that best matches input data to its desired output [90]. This model is often achieved by iteratively optimizing an objective function in order to obtain the best possible score in the training data [91]. Some examples of supervised learning include classification, regression and active learning [92].

When training a machine learning model, the dataset used to build this model is often divided into two subsets, training and testing. First, a training set, which will be used to train the model and tune its parameters to best predict the outputs of the training set. And second, a testing set, which will be used to evaluate the performance of the algorithm and can help ensure the model isn't overfit to training data. In order to evaluate the performance of the model error metrics are used, with some of the most common metrics being: error probability, the probability of the model making a wrong decision; confusion matrices, a table layout that allows visualization of the performance of an algorithm, with each row of the matrix representing the instances in an actual class and each column representing the instances in a predicted class; and out-of-bag error estimates (explained in detail below as a performance metric for Random Forests). Overfitting is the process which may occur when the available training data is too small to express all patterns of variability and, therefore, the trained model is unable to generalize for data beyond the training set. As a result, the efficiency and accuracy of the model decrease when tested with data it has never seen before (testing set) [93].

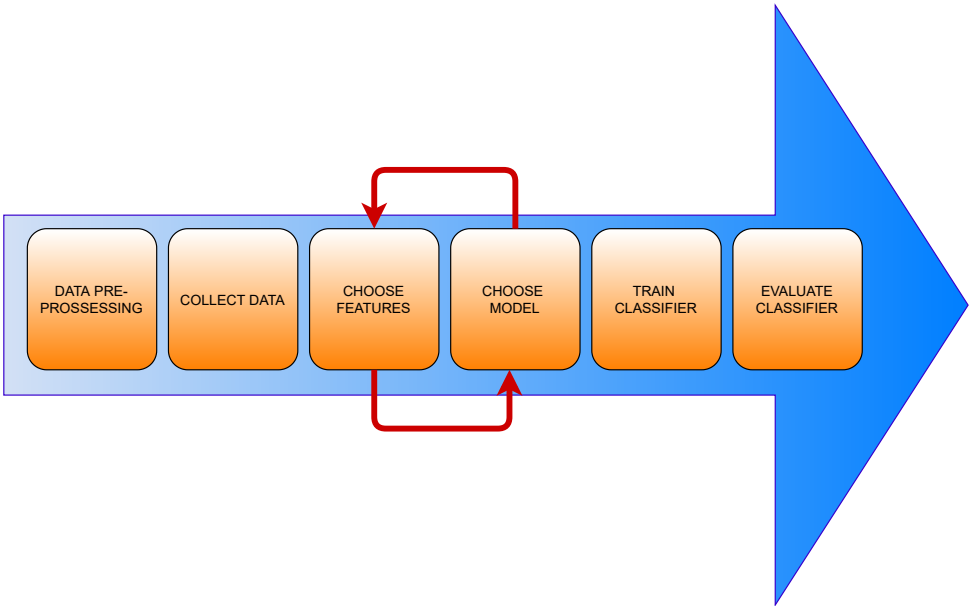

Figure 2.5: Schematic of the Machine Learning process

These algorithms are employed in many different applications, and are especially useful and prevalent in tasks considered too hard to program in a rule-based algorithm. Some of these tasks include image processing [94], email filters [95] and fraud detection [96].

2.4.3 Feature selection/representation

Feature selection, the process which allows a system to automatically discover the most relevant features for classification from raw data, is often required in machine learning models. This is due to the fact that these algorithms commonly require mathematically or computationally precise input, but raw data often has many features which are completely irrelevant, therefore, a process of selection of relevant features is often required before the classification and prediction steps [97]. Similarly, several learning algorithms have the ability to discover better feature representations of the inputs provided in the training set, using a process called feature extraction [97]. Some examples include principal components/coordinate analysis and clustering methods. The algorithms attempt to preserve the input information while transforming it in an attempt to make it more useful/relevant.

2.4.4 Decision trees

A decision tree is a predictive model designed to associate a prediction or decision over an observation, the decision being organized in a hierarchy that forms a tree. The two main types of decision trees are classification (the output is a discrete set of variables) and regression (the outputs are continuous values). Given the intelligibility of these algorithms and the ability to visualize decision making, decision

trees constitute some of the most popular machine learning algorithms [98] [99]. Each tree is composed of decision nodes, branches and leaf nodes. At decision nodes tests are made on a specific feature, the outcome of which results in a branch connecting to other nodes in the tree; the first decision node is called a root node and the terminal nodes with a class label are called leaf nodes [100].

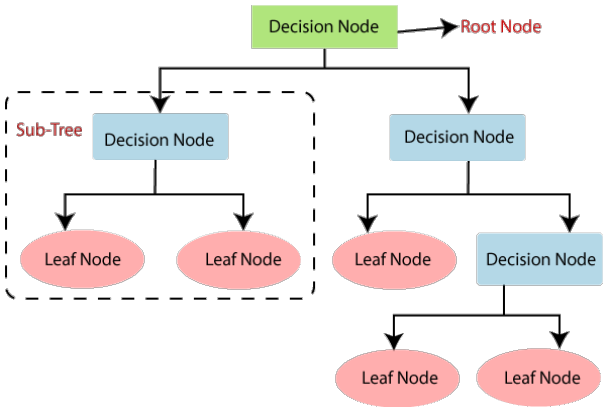

Figure 2.6: Example of a decision tree. Adapted from [101].

The construction of these trees tends to start at the root node and builds its way down to the leaf nodes, choosing the best test to perform at each decision node in order to split the set of items to be classified in the optimal manner [102]. Depending on the algorithm used, different metrics are used for determining the best split, including information gain, variance reduction and Gini impurity.

The Gini impurity metric is the probability of an incorrect classification of a new input of a random variable, if that classification were done according to the current distribution of the data set's class labels [103]. It is calculated using equation 2.2, with G representing the Gini impurity, c being the total number of classes and p(i) the probability of picking a data point with class i.

$$G = \sum_{i=1}^{C} p(i) * (1 - p(i)) \tag{2.2}$$

At each decision node the Gini gain is calculated by taking the initial Gini impurity of the branch upstream from the decision node and subtracting the Gini impurity of the resulting subsets for each possible split of each class [103].

Given that these metrics provide a way to measure how relevant each feature is, decision trees or techniques utilizing decision trees (i.e. Random Forests which will be further explained below) can therefore be used for feature selection. A process of feature selection can be performed by observing the impact each feature has on the Gini value of a node and establishing a threshold value of Gini impact above which the features are selected [104].

2.4.5 Random Forest

One technique often employed in machine learning is the use of Ensemble methods. These are techniques that aim at improving the accuracy of results in models by combining the prediction of multiple models instead of a single one. One way to do this is to train multiple weaker models and then incorporating all their predictions into a single result. This approach has been shown to improve result accuracy significantly by reducing bias and variance, consequently, boosting the accuracy of regression and classification models [105].

A Random Forest (RF) is an ensemble learning method, constructed with multiple decision trees, which seeks to correct the tendency of decision trees to overfit to the training data [106]. It has a singular output which will be either the mode (classification) or the mean (regression) of all outputs [100] [106]. Each tree is constructed by selecting a random subset of all features, meaning that no tree takes into account all features. The amount of features selected for each tree (tree depth) is a heuristic based on the total number of features, which can be altered in order to obtain better accuracy. Typically, with **N** as the total number of features, $\sqrt{\mathbf{N}}$ or log(**N**) features tend to be used for classification problems and **N/3** features for regression [107]. Since no tree has all features, the trees are de-correlated which helps prevent over-fitting to the training data [107] [100]. In addition, the trees are constructed utilizing a process called bagging or bootstrap aggregating, meaning that each tree is also given a random independent subset of all samples within the training set, so for each tree, there is a subset of samples with which it was not trained (the "out-of-bag" samples), and can therefore be used for model validation [108].

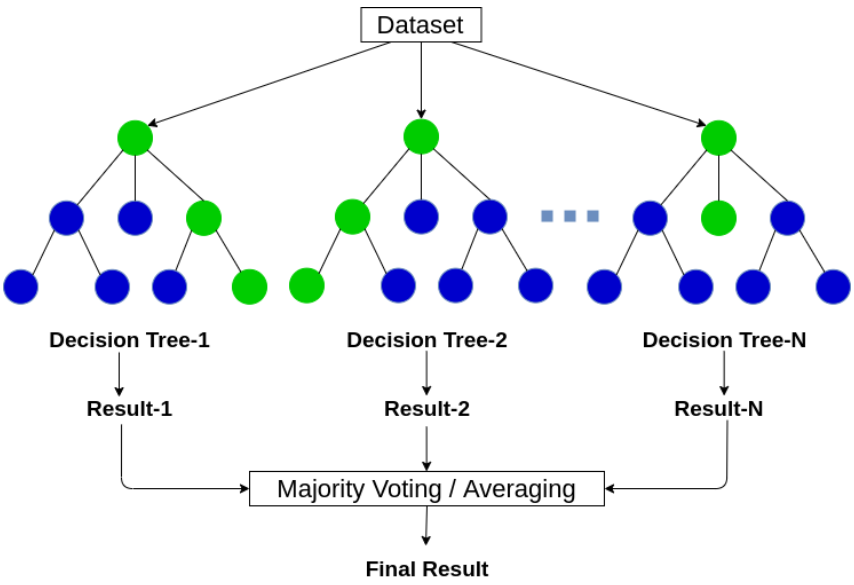

Figure 2.7: Schematic of a Random Forest model.Taken from [109].

2.4.5.A Performance evaluation

Many different approaches can be used to evaluate the performance of machine learning models, such as cross-validation and Out-of-bag (OOB) error.

Cross-validation methods involve the partitioning of samples into subsets, with a particular instance of this being a simple train-test set split of the available training data. Using the latter method, one subset is called the training set and is used to train the model, while the other subset (testing or validation set) is used to validate the generated model. These methods often resort to multiple different partitions and rounds of cross-validation, with the performance results being averaged over all rounds in an attempt to reduce variability [110].

OOB error is used in error prediction in machine learning models utilizing bagging, such as random forests. Since in a RF, each tree utilizes a subset of all samples smaller than the total number of samples, then for each tree there is a subset of samples with which it was not trained. The OOB error is measured by validating each tree with the subset of samples which had not been part of that tree's training set, using this set to obtain the decision for each tree and then determining the mean prediction error using the fusion of all these individual decisions into a singular class label [108]. A visual representation of the allocation of samples for training and validation can be found in 2.8.

The OOB error is one of the most common means of analyzing prediction error in RFs [108], mainly due to requiring far less computational power than methods like cross-validation and allowing the testing of the model during the training process. This is mainly due to the subsets used in this method already being created during the bagging phase. It should be noted that despite its frequent use, this metric has been shown to overestimate the prediction error rate [111].

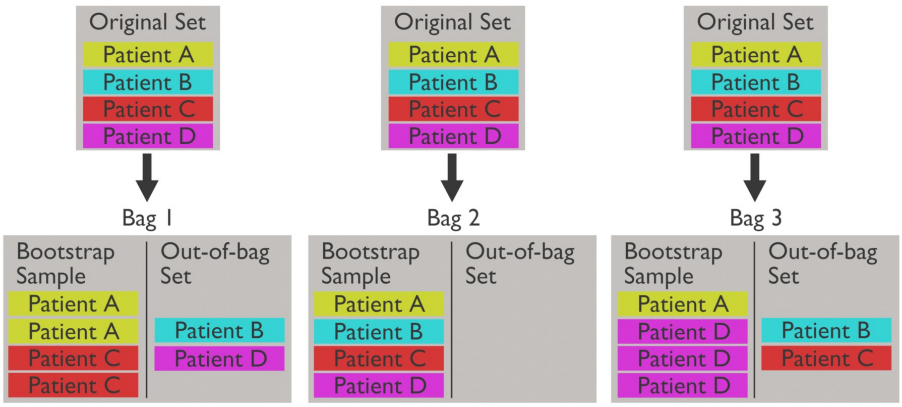

Figure 2.8: Visualization of the in-bag and out-of-bag sets. Adapted from [108].

2.4.6 Neural Networks

Neural Networks (NN) are a series of interconnected layers of nodes, which loosely model a brain. For this reason, the nodes are often referred to as (artificial) neurons, while the connections between these neurons are often called edges [112].

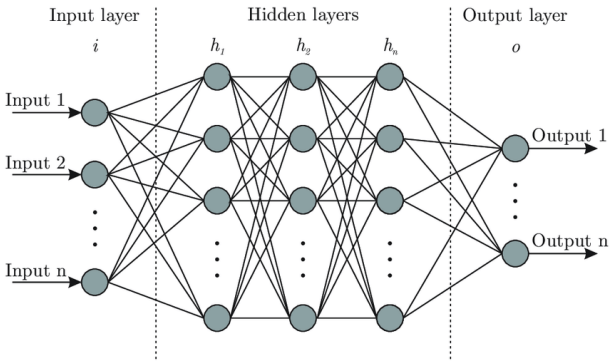

Figure 2.9: Model of a feed-forward neural network. Adapted from [113].

In the simplest NN models, feed-forward NNs, signals travel from the input layer to the output layer, with the signal being transformed in each layer. Each neuron can receive a signal, a real number, process it and output a new signal to all connected neurons in the following layer. The output at a given neuron is calculated by a function of the sum of it's inputs. Both the neurons and the edges will generally have a weight, which either increases or decreases the "strength" of a signal or even stops it from being sent if the aggregate signal is above/below a certain threshold. For most cases, these weights are the variables adjusted by the algorithm as the training process takes place, increasing or decreasing these weights in an attempt to better model the training data and receive the expected outputs [112].

The training process generally consists of using examples with a known input and result. The input of these examples is used as the "initial signal" and produces an output. This output is compared with the expected result and the difference between the two is considered the error. Following this, the model changes it's weights using this error and a learning rule (backpropagation). Iterating this process leads to the neural network's output to become increasingly similar to the result [112]. The simplest form of a neural network has 3 layers, the input, hidden and output layers. A neural network with multiple hidden layers is called a deep neural network [112].

2.4.7 Deep learning

As of recent years, deep learning is one the most used approaches in machine learning applications [92]. This is a technique that uses neural networks with multiple layers to progressively extract higher-level features from the raw data [112]. An example comparing a traditional neural network with a deep learning model can be found in figure 2.10. As can be seen, in a Deep Learning model, there are multiples

layers which transform their input into a more abstract and composite representation in order to output a prediction.

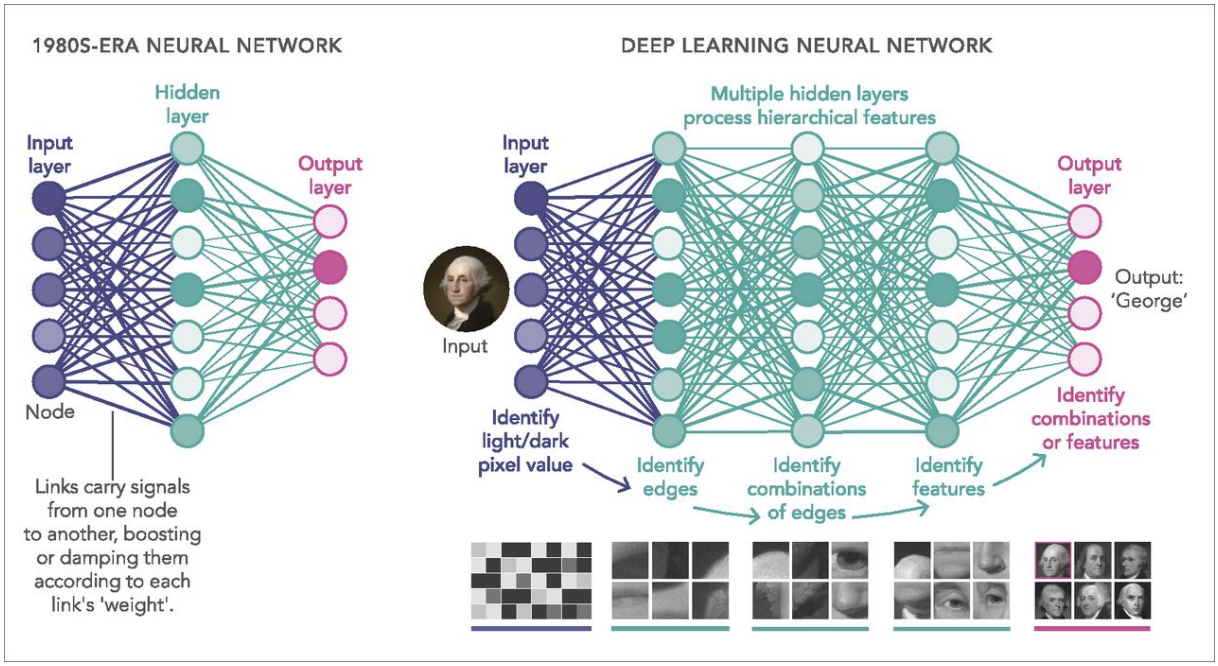

Figure 2.10: Comparison between a classical neural network and deep-learning neural network. Adapted from [114].

2.4.8 PredicTF

Deep Learning techniques have been employed in prediction of DNA sequence affinities [115], to identify TF binding sites in humans [116] and even gene regulation [117], however, until the creation of the PredicTF tool, it had never been used as an approach to predict bacterial TFs.

PredicTF is a software tool used for the prediction of novel TFs and their respective families using genomic and metagenomic data. The deep learning approach used is similar to the one described in the DeepARG tool [118]. The reason why deep learning was used as opposed to other machine learning techniques is due to its ability to extract relevant features without the need for human intervention [119] [120] [121] and its ability to resolve multiclass classification problems [122] [119]. Initially, characterization had to be performed, in order to represent DNA sequences as numerical values (features). With this purpose, dissimilarity based classification [123] was used, with any given sequence being represented by its identity distance to known TFs. The PredicTF model then consisted of 4 dense hidden layers with 2000, 1000, 500 and 100 nodes, propagating the input score distribution into dense and abstract features.

For the purpose of training and validation, a transcription factor database was created (BacTFDB) by merging and curating TFs in the UniProt [4] and CollectTF [5] databases. UniProt is ”a comprehensive,

high-quality and freely accessible resource of protein sequence and functional information" [4] and CollecTF is a database with well described and characterized TFs, which have been validated *in vivo* [2]. BacTFDB was then used as the training set for development of a deep learning model capable of TF prediction, with the accuracy and performance being tested with five model organisms (*Escherichia coli, Bacillus subtillis, Pseudomonas fluorescens, Azotobacter vinelandii and Caulobacter crescentus*), a clinical isolate (*P. aeruginosa* PAO1) and a metagenomic sample from an anaerobic ammonium oxidation community [2]. In addition, this tool was also tested in transcriptomes and metatranscriptomes. During the training, to prevent overfitting, random hidden units were removed from the model at different rates using the dropout technique [124], an algorithm for training neural networks by randomly dropping units during training to prevent their co-adaptation. Finally, the output layer of the deep neural network outputs the most likely subfamily of TF, as well as its confidence in this result. This output layer utilizes a softMax [125] activation function (converting the output values into probabilities) that computes the probability of the input sequence against each TF category and using this value assigns a TF category to the sequence [2].

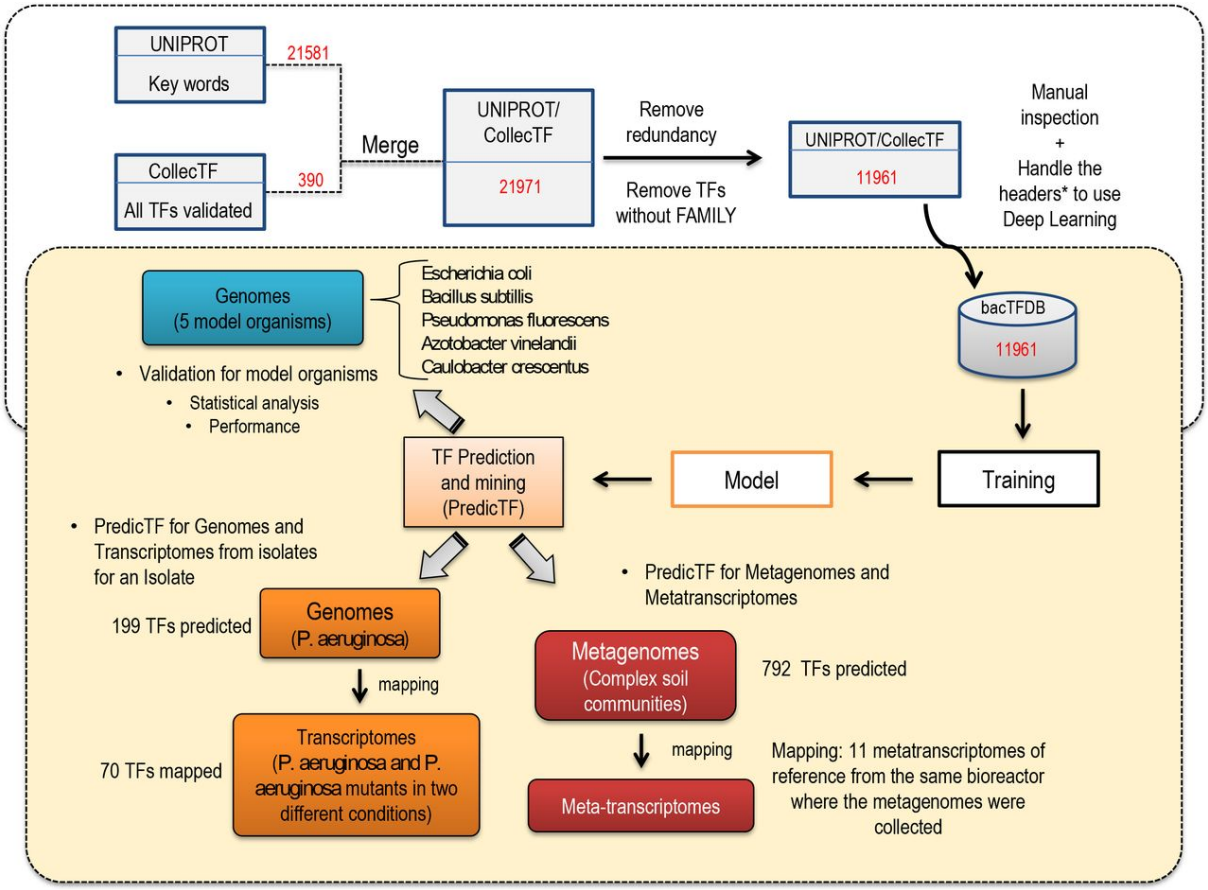

Figure 2.11: Summary of training, testing and validation of the PredicTF tool [2].

3

Materials and Methods

Contents

3.1 Microcosm setup and sampling 25
3.2 Rarefaction and assembly 26
3.3 PredicTF 26
3.4 Diversity Analysis 27
3.5 Identifying bioindicators using machine learning 29

In this chapter the experimental design and data analysis methods are detailed. The proposed methodology and associated main steps, summarized in figure 3.1, are detailed next.

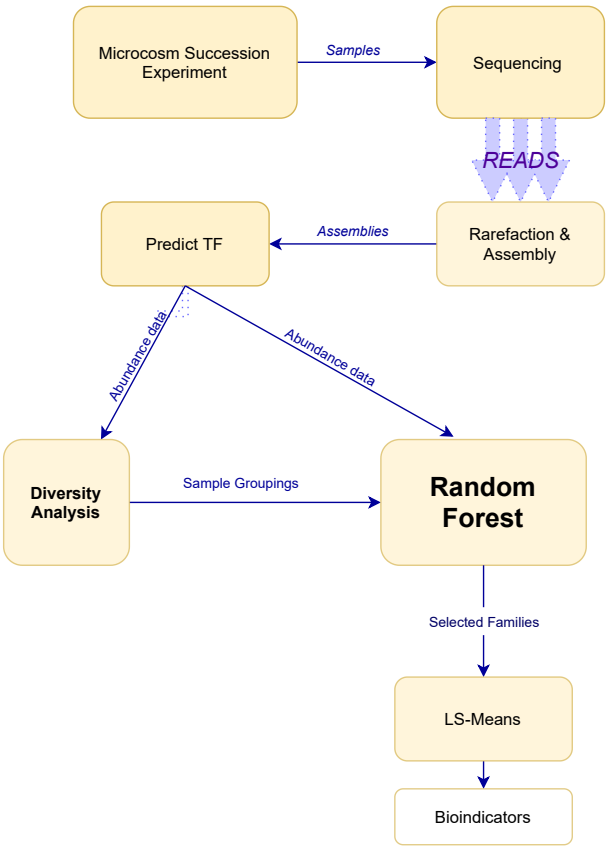

Figure 3.1: Schematic description of the proposed methodology.

3.1 Microcosm setup and sampling

The initial microcosm setup was performed by Dennis Metze and colleagues [6].A controlled mineralization experiment was set up with a benzene-mineralizing nitrate reducing culture. This culture was enriched using an on-site column system with coarse sand which had been percolated with groundwater at a benzene-contaminated aquifer and had been maintained under nitrate-reducing conditions in the laboratory for several years [126]. A concentration of 1% (v/v) benzene was obtained, using 2,2,4,4,6,8,8-Heptamethylnonane (HMN) as a carrier phase. Samples for DNA extraction were taken as 5 mL of liquid. Technical triplicates were obtained per each sampling time (0, 70, 76, 96 and 124 days) and were immediately stored at -80 ℃ until DNA extraction. Likewise, samples for metaproteome analysis were taken at the same time. Microbial community diversity was analyzed by paired-end sequencing of reads obtained thorugh shotgun, full metagenome sequencing [126].

3.2 Rarefaction and assembly

With the sequencing of metagenomic data, $1.3 \cdot 10^8$ reads were sequenced, among 15 different samples in 5 different time points, for an average of $8.6 \cdot 10^6$ reads per sample.

Following sequencing and quality control, initially, the reads were assembled using MetaSpades, ran on the PredicTF tool and diversity analysis was started. However, following visualization of the initial results, a decision was made to rarefy the original reads and repeat this process. The initial reads were normalized by rarefaction to the lowest sample size of $6 \cdot 10^6$. This was performed in order to, as previously explained, remove library size bias from our analysis.

After the rarefaction step, assembly was performed using the Metaspades [127] algorithm, with 30GB of allocated memory over 10 hours.

3.3 PredicTF

In this section we explain how PredicTF was used to determine the abundance of TF families and subfamilies in the metagenomic assemblies.

This tool is used as a platform for prediction and classification of bacterial TFs in microbial communities [2]. Following the assembly process, the assemblies are used as input for PredicTF. PredicTF outputs the predicted TF subfamily and family, their query position, closest hit in the database, probability of a real match, alignment length, e-value and in which bin/assembly it was predicted. Any output with a probability lower than 97% is disregarded as well as any e-value (probability due to chance, that there is another alignment with a similarity greater than the given score) higher than 10^{-10}.

For the diversity analysis, 4 different indices are used: absolute abundance of TF subfamilies, relative abundance of TF subfamilies, absolute abundance of TF families and relative abundance of TF families. Given the fact that, to the best of our knowledge, no research has yet been published using this tool or measuring TF abundance for a diversity analysis, these multiple metrics were used in an attempt to ascertain which could provide the most meaningful results. Note that the absolute abundance of the TFs and their families could only be used given that the reads were previously rarefied [128].

For the TF families, any family with absolute abundance lower than 3 is not taken into account in downstream analyses. Due to the limited amount of data in this dataset and the large amount of TF subfamilies, when analyzing the PredicTF output, the absolute abundance of most subfamilies was below 3 and the vast majority of subfamilies had at most 2 instances in each sample. Unfortunately, these low abundance levels make it difficult to differentiate between the different samples. Consequently, further along in the diversity analysis, the results concerning TF families are generally considered to be more trustworthy and relevant. In any future work with this tool this is a concern that has to be taken into account, which may potentially be dealt with by either having a larger dataset or simply using only the

abundance of TF families, as is eventually done in this work. It should be noted that due to the broad range of functions that members of certain TF families (such as LysR) can have, being able to ascertain differences in abundance levels of individual subfamilies should be more relevant than entire families, therefore being able to analyze these subfamilies in more detail with more data should be an objective of any future work and not something to be disregarded initially.

3.4 Diversity Analysis

In this section, an explanation of how the diversity analysis was conducted, using the PredicTF output (abundance data of specific TF subfamilies and families in each sample). First, the alpha diversity of the samples is analyzed by both absolute number of TFs and Shannon Index. Following this analysis, an analysis on the beta diversity of the samples is undertaken, using Non-metric multidimensional scaling (NMDS) and Principal coordinate analysis (PCoA) methods. Finally, the variance between samples belonging to the same time point is observed.

3.4.1 Alpha and Beta diversity

Data analyses were completed using the phyloseq [129], vegan [130] and RandomForest [131] packages in R software [132]. In an attempt to assess richness and evenness of the samples, the alpha diversity indices (observed number of TFs and Shannon Index) were estimated using the rarefied data. The statistical significance of the differences in alpha diversity measures between samples was then estimated using a pairwise t-test ("rstatix" R package) [133]. The beta diversity of the samples was studied using both NMDS and PCoA ordination, using the Bray-Curtis dissimilarity method to calculate distance [133].

PCoA is an ordination method which preserves dissimilarity measures between objects [65]. It is used to represent all points in an Euclidean space and can produce 2-dimensional (2D) reduced ordinations of multi-dimensional objects, allowing for the data to be far more easily visualized, with the most relevant coordinates (those accounting for the most variance) being represented in the 2D space [65]. NMDS is also an ordination technique, however, unlike PCoA where many axes are calculated but only a few are viewed, in NMDS, a small number of axes are explicitly chosen prior to the analysis and the data is fitted to those dimensions; there are no hidden axes of variation [64].

Given that there are 27 TF families and well over 100 subfamilies, in order to represent them graphically in a manner that could be understandable there was need for the dimensionality reduction of the PCoA and NMDS methods, allowing us to visualize the most relevant features of our data in a 2-dimensional plot. This visualization is used in order to determine whether or not clusters or decision surfaces can be observed, separating the different time points at which samples were collected. They

may also be used to cluster multiple time points. For example, the visualizations may show a clear division in samples taken before a certain date vs those taken at later time points.

Following the visualization of the results, the samples were divided between Early (taken at 0,70 and 76 days) and Late (taken at 96 and 124 days).

The statistical significance of the principal coordinates in the different time points and groupings created was then assessed using a Permutational multivariate analysis of variance (PERMANOVA) [134], as this statistical test has been shown to be very powerful as a tool to detect changes in community structures [133]. PERMANOVA is a non-parametric multivariate statistical test, used to compare groups of objects and test the null hypothesis that the centroids and dispersion of the groups as defined by measure space are equivalent for all groups [133]. Meaning that if this null hypothesis is rejected then the centroid and/or spread of species is different among the different groups. This statistic is calculated as follows:

With **p** groups and **n** objects in each group out **N** total objects, with d_{ij}^2 the squared distance between objects **i** and **j** and ϵ_{ij} being equal to 1 if **i** and **j** are in the same group and 0 otherwise, the total sum-of-squares(SS_T) and the within groups sum-of-squares (SS_W) are determined as:

$$SS_T = \frac{1}{N} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_{ij}^2 \tag{3.1}$$

$$SS_W = \frac{1}{n} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_{ij}^2 \epsilon_{ij} \tag{3.2}$$

Then, the between groups sum-of-squares SS_A can be calculated as the difference between the overall and the within groups sum-of-squares: $SS_A = SS_T - SS_W$ This value is used to calculate the Pseudo_F value:

$$F = \frac{\left(\frac{SS_A}{p-1}\right)}{\left(\frac{SS_W}{N-p}\right)} \tag{3.3}$$

Finally, this Pseudo-F value is used to obtain a p-value, by performing multiple permutations of the data with each permutation shuffling the items between groups and calculating the **F** value. The p-value is then calculated by:

$$P = \frac{(\text{ count } F^p \geq F) + 1}{(\text{ total } \text{ count } F^p) + 1} \tag{3.4}$$

Where **F** is the F statistic obtained from the original data and F^p is a permutation F statistic.

Overall this diversity analysis will allow us to determine whether there really is a statistically significant difference in the abundance levels of certain TFs between the different time points and groupings, allowing for a further study forward on which specific TF (sub)families have differences in their abundance

level.

3.4.2 Bray-Curtis dissimilarity index

Bray–Curtis dissimilarity is used to determine the compositional dissimilarity between two different samples, based on each sample's counts. As defined by the authors [135], the index of dissimilarity is:

$$BC_{ij} = 1 - \frac{2C_{ij}}{S_i + S_j} \tag{3.5}$$

With C_{ij} representing the sum of the lesser value between all species present in both samples. So, for instance, if we have **Sample 1** with 5 **A** species and 3 **B** species and **Sample 2** with 2 **A** species, 7 **B** species and 3 **C** species, C_{ij} would be equal to the sum of 2 (smallest number of species **A** in either sample) + 3 (smallest number of species **B** in either sample). **C** species are not used for the calculation of C_{ij}, as they are not present in both samples. S_i and S_j represent the total number of specimens in each sample.

In order to analyze variance, the Bray-Curtis dissimilarity of each sample within the same time point was graphed. This is done as it reveals which time points had the greatest variance of abundance level between samples indicating potential sampling or analysis errors.

3.5 Identifying bioindicators using machine learning

Since the division of the samples into the Early and Late time groups leads to the overall abundance levels of TF families to change over time, an analysis on which families have the most significant change is conducted. With this objective, a Random Forest is constructed to distinguish the samples between Early and Late, using their relative abundance, with the families that contribute the most GINI information gain to the creation of this forest being selected. Following this selection, those with statistically significant differences in their abundance levels are considered potential bioindicators and the difference in abundance between their Early and Late samples is tested using Least Square Means (LS Means) tests.

A Random forest was generated with the intent of identifying TF families which may be potential bioindicators. This analysis was performed using the "RandomForest" package [136]. As previously mentioned in chapter 2 , Random Forests can be used for feature selection. This method was also chosen as it is very useful when dealing with limited datasets where the number of features is larger than the number of samples [100], which is the case in this work, as there are only 15 samples but 27 families (and 165 subfamilies). The number of variables tried at each split (also known as tree depth) is generally recommended to be close to the square root of total, so in this case $\sqrt{27} \approx 5$. In order to

analyze the impact of tree depth in the final performance of the model and associated study, two models were implemented, corresponding to a depth of 5 and 20.

When determining the number of trees in each forest, an initial value of 2000 was used. However, this value was increased to take into account all the possible combinations of 5 or 20 variables out of the initial 27 families. This value corresponds to 80730 combinations of 5 variables and 888030 combinations of 20 variables. With this in mind, forests were created with values ranging from the initial 2000 to 800000.

Following this, a plot of the mean decrease GINI value for each family was obtained from the Random Forests with the lowest out-of-bag error rate and all families that caused an increase of +5% in the mean Gini value were selected. This was done since distinctive features identified by random forests are visualized by their Gini score [131]. Prediction success was estimated with the out-of-bag error (how often a subsample was misclassified) [111].

Following the Random Forest generation and selection of the relevant TF families by using their Mean GINI value, a Least Squares Means (LS means) test is performed on all relevant families, comparing their early and late abundance values. Least-square means or marginal means are the group means after having controlled for a covariate [137]. In the instance of this work, this entails calculating the mean abundance value for each family in all samples and then calculating the average value for each time group (Early and Late). Least-squares means are predictions on a linear model and can be used in unbalanced datasets [138]. Following a pairwise t-test on the LS means results, with false discovery rate adjustment, the families with statistically significant ($p<0.01$) differences between these two time groups were considered to be bioindicators.

After the selection of these bioindicator families, one family (NtrC.DctD) was chosen as it had the most significant results and few subfamilies. These subfamilies were also tested with a pairwise LS-means statistical test, in order to determine whether their abundance varied significantly between the Early and Late time groups.

4

Results and Discussion

Contents

4.1 PredicTF output 33

4.2 Diversity analysis 35

4.3 Bioindicators 42

The present chapter is divided into 3 main sections, in accordance with the framework layed out to achieve the main goal of this thesis – the exploration of potential applications of newly developed transcription factor predicting tools in order to determine the impact of environmental stresses on the abundance of bacterial TFs and whether these differences in abundance levels could be used to distinguish between samples at different time points of a succession experiment. In section 4.1, the PredicTF output results are initially analyzed and visualized. In section 4.2, the diversity analysis of the samples is conducted. In section 4.3, Random Forests and statistical tests are used in order to determine which TF families could be potential bioindicators in this specific succession experiment, in order to understand how the abundance of TF families and subfamilies is altered when a community is exposed to a new stress in its environment.

4.1 PredicTF output

Upon using the PredicTF tool to predict transcription factors on the 15 metagenomic assemblies, a total of 1569 TFs were found, spread out over 159 subfamilies and 27 families. These results can be found in Supplementary Material A, as the tables are too large to be conveniently shown in the main document. The number of TFs of each family and their relative abundance in every sample can be found in figures 4.1 and 4.2, respectively.

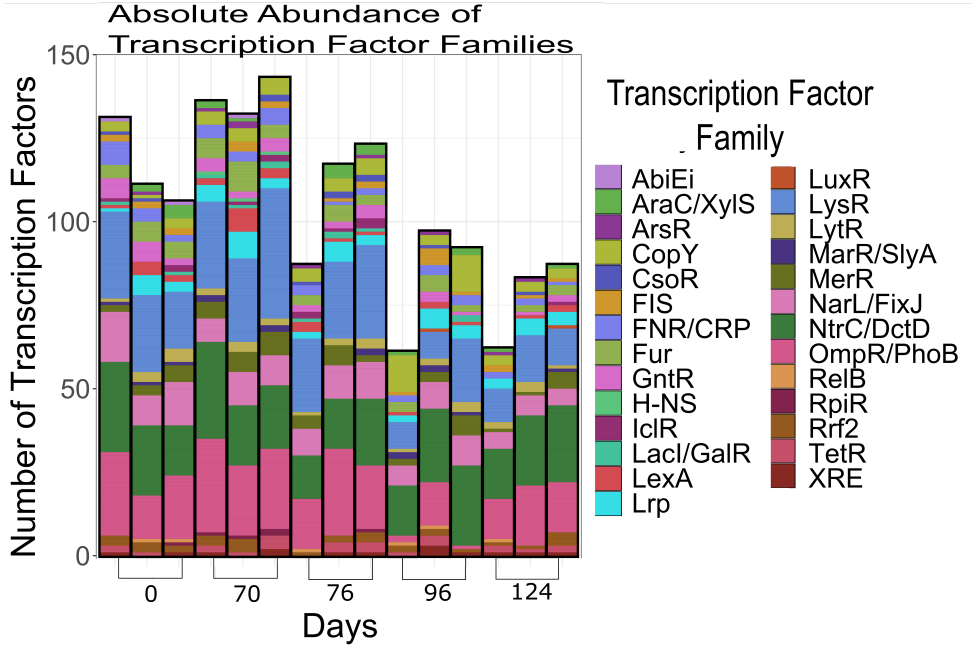

Figure 4.1: Absolute abundance of Transcription Factor families for samples collected at the different days of the experiment. Each column represents one sample. The divisions (line) inside each family represent subfamilies.

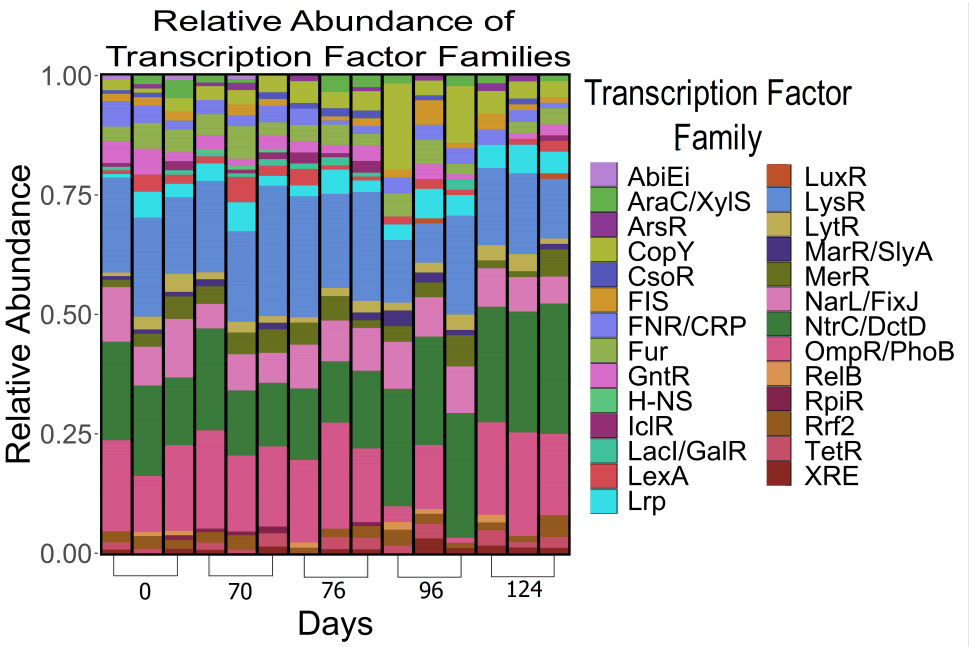

Figure 4.2: Relative abundance (%) of Transcription Factor families for samples collected at the different days of the experiment. Each column represents one sample. The divisions (line) inside each family represent subfamilies.

4.2 Diversity analysis

4.2.1 Alpha diversity

The Alpha diversity measures for the TF subfamilies can be found in figure 4.3 and for the TF families can be found in figure 4.4.

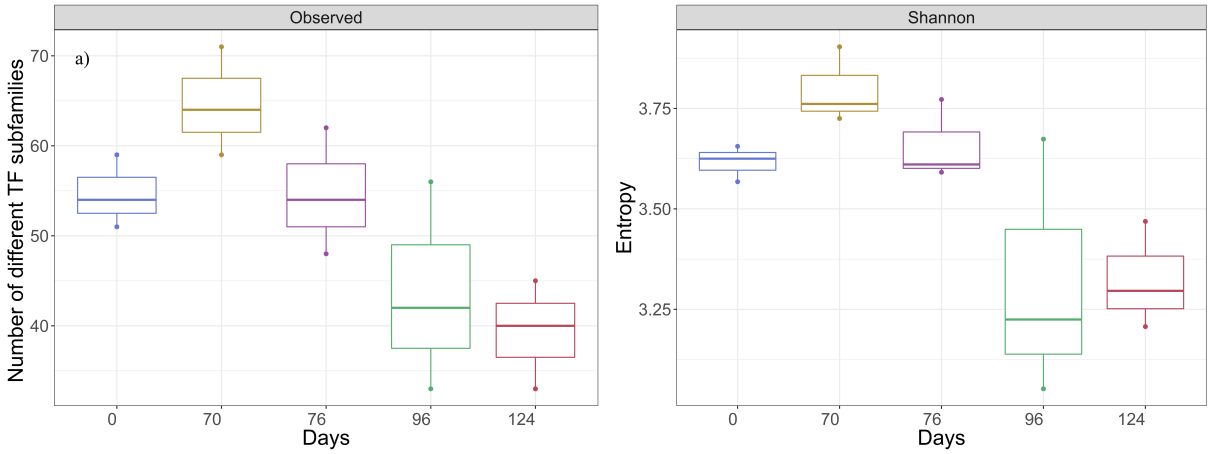

Figure 4.3: Alpha Diversity of the TF sub-families, measured in number of Observed Transcription Factors (a) and Shannon Index (b). The line inside the box represents the median value, and the box contains the 25th to 75th percentiles of the dataset.

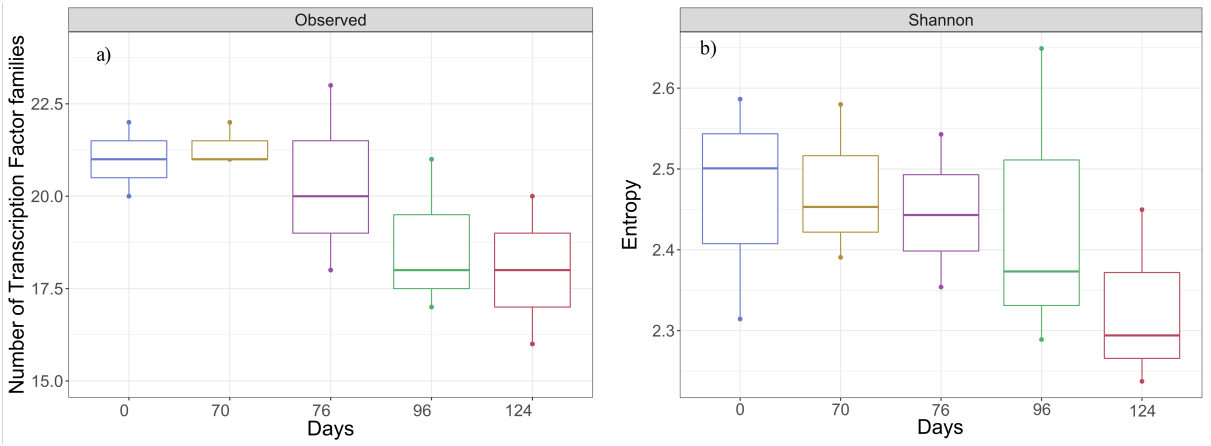

Figure 4.4: Alpha Diversity of the TF families, measured in number of Observed Transcription Factors (a) and Shannon Index. The line inside the box represents the median value, and the box contains the 25th to 75th percentiles of the dataset.

The pairwise t-test results can be found in tables 4.1 and 4.2.

Days	Days	p-value (Observed)	Adjusted p value (Observed)	p-value (Shannon)	Adjusted p value (Shannon)
0	70	0.184	1	0.146	1
0	76	1	1	0.658	1
0	96	0.309	1	0.224	1
0	124	0.119	1	0.097	0.969
70	76	0.09	0.897	0.002	**0.019**
70	96	0.145	1	0.134	1
70	124	0.018	0.184	0.003	**0.026**
76	96	0.24	1	0.233	1
76	124	0.003	**0.033**	0.006	0.057
96	124	0.539	1	0.974	1

Table 4.1: Pairwise t-test on the Alpha diversity results for TF sub-families (Observed Number of TF families and Shannon Index), using Bonferroni correction for adjusted p-values. Statistically significant (adjusted p-value$<$0.05) values in bold.

As can be seen in table 4.1, there are statistically significant differences in the alpha diversity of sub-families among certain samples, for both the observed and Shannon metrics. This indicates a significant difference in the number of different sub-families between samples taken at 76 and 124 days. as well a significant difference in the diversity of the samples taken at 70 days vs the samples taken at 76 and 124.

Days	Days	p-value (Observed)	Adjusted p value (Observed)	p-value (Shannon)	Adjusted p value (Shannon)
0	70	0.742	1	0.927	1
0	76	0.635	1	0.567	1
0	96	0.296	1	0.846	1
0	124	0.096	0.848	0.184	1
70	76	0.58	1	0.712	1
70	96	0.094	0.848	0.638	1
70	124	0.109	0.848	0.282	1
76	96	0.444	1	0.95	1
76	124	0.02	0.198	0.113	1
96	124	0.691	1	0.575	1

Table 4.2: Pairwise t-test on the Alpha diversity results for TF families (Observed Number of TF families and Shannon Index), using Bonferroni correction for adjusted p-values.

Despite the results in figure 4.4 seemingly indicating a decrease in the alpha diversity over time, the pairwise t-test shows no significant statistical difference in the Alpha Diversity values of the different time points for families, in either Observed or Shannon measures. This indicates that the overall species diversity seems to be the same among all samples. Since these are alpha diversity measures, the results indicate that over time in this succession experiment, the number of different transcription factor families does not seem to change significantly. This does not however mean that the abundance of certain families has not changed. In order to observe whether different families are being expressed more or less, we'll have to resort to beta diversity metrics, comparing the diversity between samples.

4.2.2 Beta Diversity

In order to perform an analysis on the beta diversity of the samples, both PCoA and NMDS plots were created, using Bray-Curtis dissimilarity as a measure of distance. The PCoA results can be seen in figures 4.5, 4.6, 4.7 and 4.8. Following analysis of the graphs, there was a decision to split the samples into two distinct groupings. These groupings are the Early (0, 70 and 76 days) samples and the Late (96 and 124 days) samples. The reason for these groupings came from realizing that these 2 groupings of samples could always be divided by a decision surface in all 4 PCoA graphs, but most importantly, this split is very apparent in the family plots (4.7 and 4.8) and given that in these graphs the 2 first principal coordinates account for the greatest amount of variance(60% for the family analysis vs 30% for the subfamily analysis) and they also have the most statistically significant results in the PERMANOVA analysis, they were considered the most relevant. Once again, the lower variance of the 2 main principal coordinates and the less statistically significant values of the subfamily analysis are believed to be due to the somewhat limited nature of this dataset and the large number of different subfamilies with very low abundance present, which makes it difficult to meaningfully discern between different samples. For this reason, the graphs are colored according to this split, with the different shapes in each datapoint representing the day at which it was collected. In addition, the PERMANOVA results for both the individual time samples and the time groups are shown as inset boxes. The graphs with time sample coloring and the NMDS plots can be found in annex A.1, as their results are generally consistent with those observed in following PCoA plots.

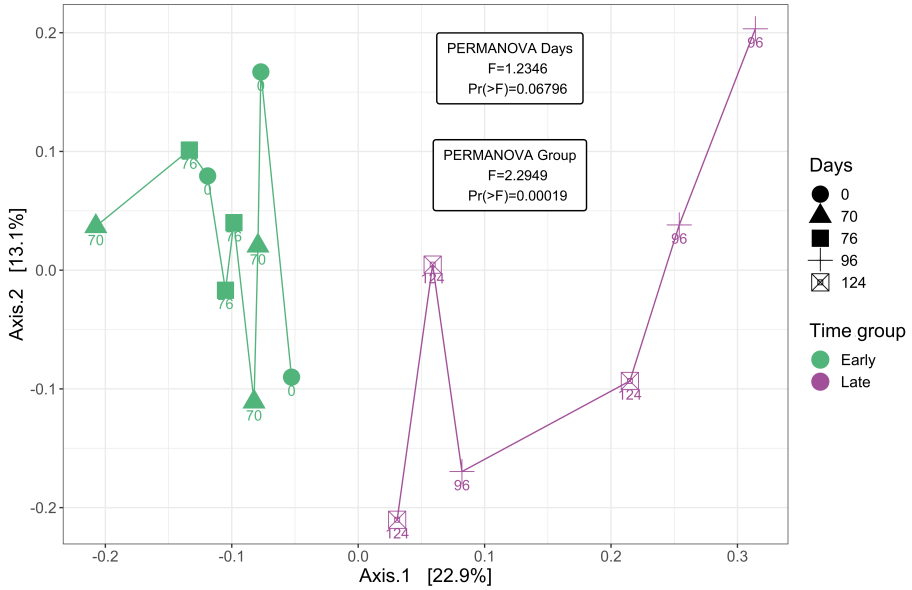

Figure 4.5: Principal Coordinate analysis of Bray-Curtis dissimilarity comparing the relative abundance of TF subfamilies in different samples. We defined two time groups as Early (Days $\leq$ 76) and Late (Days>76). Permutational multivariate analysis of variance (PERMANOVA) of both the individual Days and the time groups are shown in the figure.

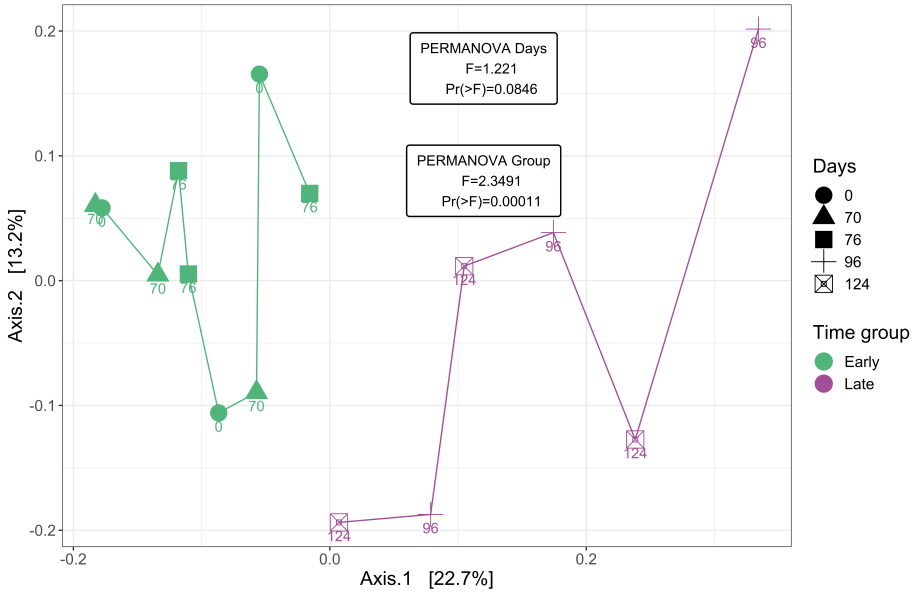

Figure 4.6: Principal Coordinate analysis of Bray-Curtis dissimilarity comparing the absolute abundance of TF subfamilies in different samples. We defined two time groups as Early (Days $\leq$ 76) and Late (Days>76). Permutational multivariate analysis of variance (PERMANOVA) of both the individual Days and the time groups are shown in the figure.

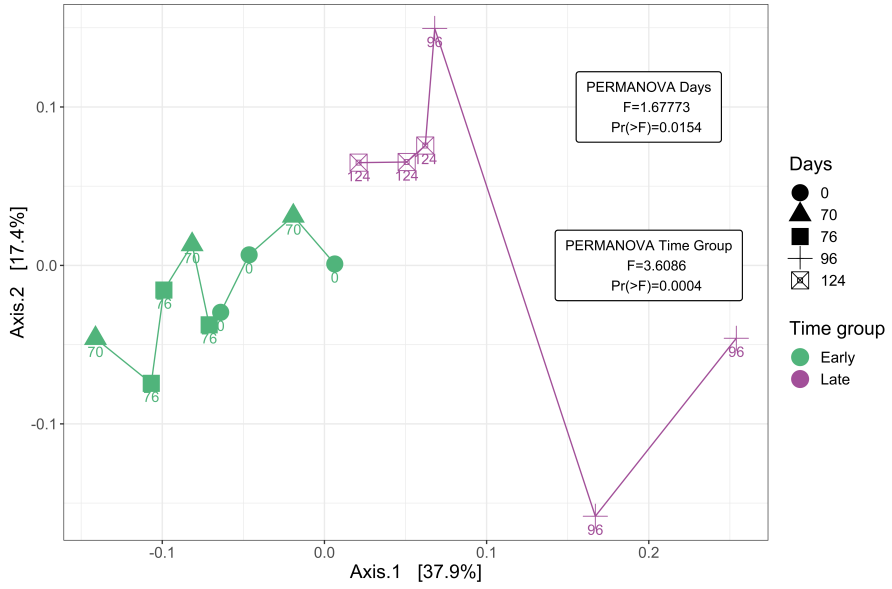

Figure 4.7: Principal Coordinate analysis of Bray-Curtis dissimiliarity comparing the relative abundance of TF families in different samples. We defined two time groups as Early (Days $\leq$ 76) and Late (Days>76). Permutational multivariate analysis of variance (PERMANOVA) of both the individual Days and the time groups are shown in the figure.

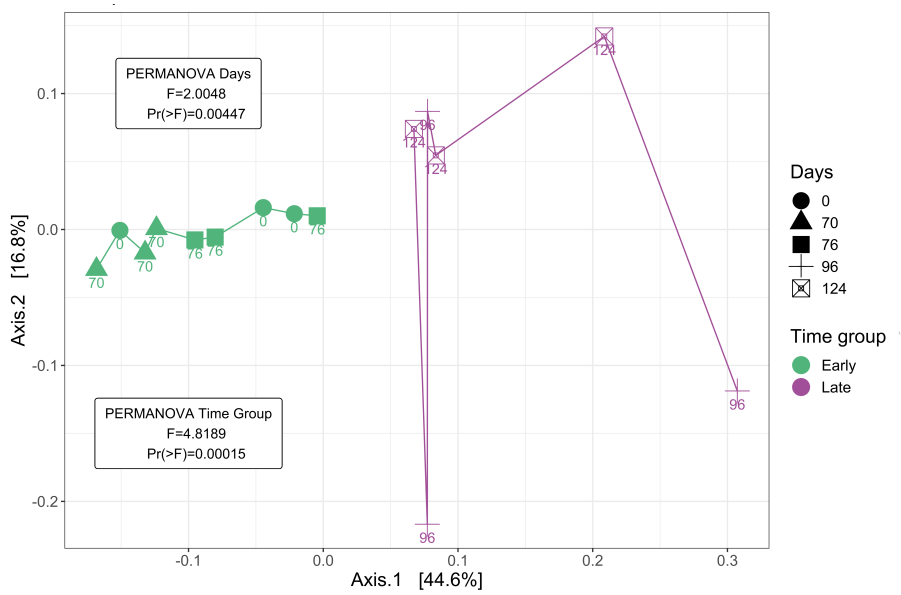

Figure 4.8: Principal Coordinate analysis of Bray-Curtis dissimilarity comparing the absolute abundance of TF families different samples. We defined two time groups as Early (Days $\leq$ 76) and Late (Days$>$76). Permutational multivariate analysis of variance (PERMANOVA) of both the individual Days and the time groups are shown in the figure.

Principal coordinates analysis using Bray-Curtis dissimilarity revealed contrasts associated with the sampling time, both for the individual time samples and time groups. Samples collected before and at 76 days differed significantly from those collected afterwards. These differences among the different time points and time groups were confirmed in a PERMANOVA analysis on Bray distances, for the different time samples (Pseudo-F=1.88;P=0.01149) and for the time groups (Pseudo-F=3.61;P=0.0005). This indicates there is a clear difference in the abundance levels of certain families in the samples among the different time points/group, potentially as a result of the stress of the environment altering abundance levels over time. Following these results, a decision was made to, for the most part, disregard the abundance of subfamily values in the proceeding analyses, as these are not considered as relevant, for the reasons previously stated.

4.2.3 Variance analysis

The variance at each time point is analyzed using the previously determined Bray-Curtis dissimilarities between samples taken on the same day.

As can be seen in figures 4.9 and 4.10, all time points seem to have low variance between their different samples, with the exception of the time point corresponding to 96 days, which has far higher variance among its different samples. This pattern was also observed in the PCoA plots 4.7 and 4.8, as the samples taken at 96 days have a high level of dispersion in the 2 axes represented. However, it should always be noted that in these PCoA graphs, only about 60% variance is shown in the 2 axes and,

therefore, these samples could have lower levels of dispersion in the remaining principal coordinates, hence why an analysis of the Bray-Curtis dissimilarity was performed. The differences in variance values was then tested using a pairwise T-test, verifying the statistical significant difference between the samples taken at 96 days and those taken at 70, 76 and 124 days for the absolute values, but no statistical significance between time points was found for relative abundance data.

Days	Days	p value	Adjusted p-value	Signficance
0	70	0.064	0.128	ns
0	76	0.204	0.337	ns
0	96	0.040	0.100	ns
0	124	0.236	0.337	ns
70	76	0.825	0.917	ns
70	96	0.014	**0.047**	*
70	124	0.810	0.917	ns
76	96	0.013	**0.047**	*
76	124	0.979	0.979	ns
96	124	0.014	**0.047**	*

Table 4.3: Pairwise t-test on the Bray Curtis dissimilarity results for the absolute abundance data, with the statistically significant results (Adjusted p-value$<$0,05) in bold.

This increase in variance of the absolute abundance values at 96 days could be due to multiple reasons, such as sampling errors or the community undergoing a shift in the abundance of its transcription factors to adapt to the stress in the environment. In order to further analyze this effect, besides the Early and Late split, other groupings of samples will be used when generating random forests later on, including grouping all samples except those taken at 96 days vs those taken at 96 days, in an attempt to determine which grouping may have the most relevant differences in their abundance.

Despite what was previously mentioned, this analysis was also performed on the subfamily abundance values, annex A.2, with the overall average distance being larger when compared to the family abundance values. This is to be expected due to the larger number of classes and more disperse nature of these values, with a lot of TF subfamilies not being detected in all samples.

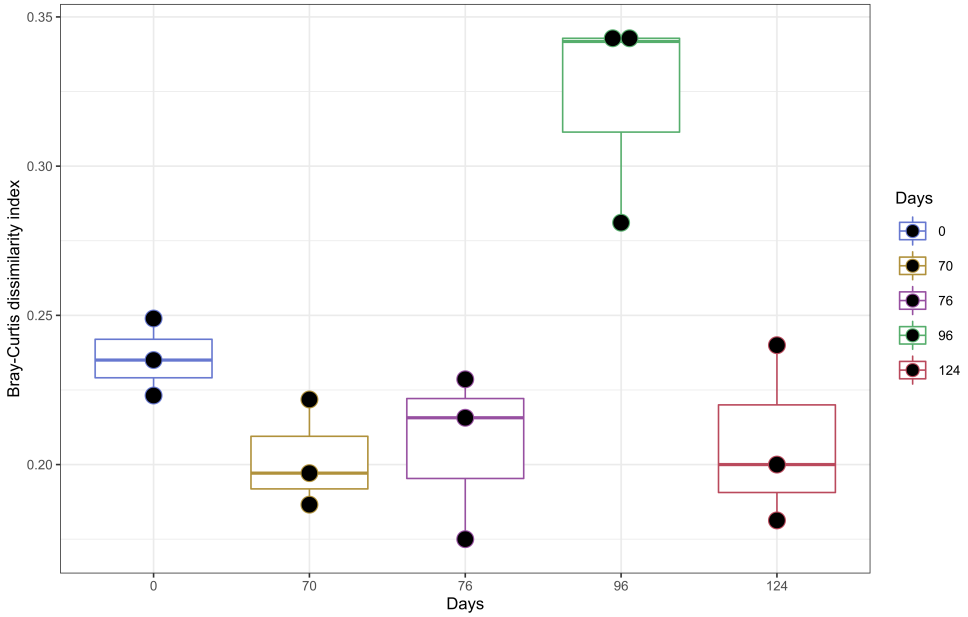

Figure 4.9: Bray-Curtis dissimilarity index among the absolute abundance of TF families in samples taken in the same day. The line inside the box represents the median value, and the box contains the 25th to 75th percentiles of the dataset.

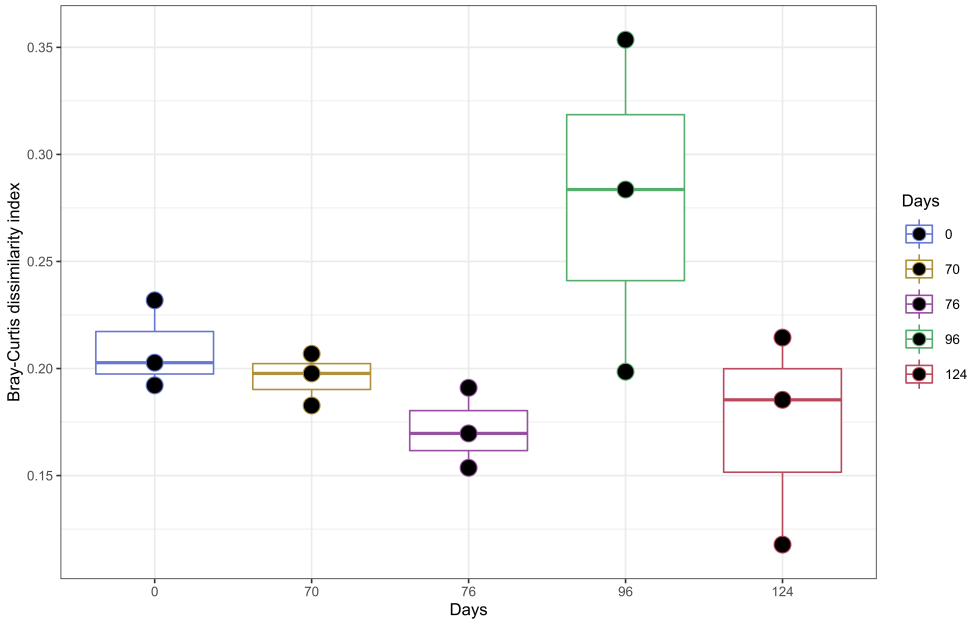

Figure 4.10: Bray-Curtis dissimilarity index among the relative abundance of TF families in samples taken in the same day. The line inside the box represents the median value, and the box contains the 25th to 75th percentiles of the dataset.

4.3 Bioindicators

4.3.1 Random Forest

By classifying the samples as late or early, as previously established, a random forest with the following confusion matrix was obtained, as shown in table 4.4, using the relative abundance of TF families. This confusion matrix corresponds to an out-of-bag error of 6,7%, due to a missclassification of 1 late sample as early. Interestingly, this exact confusion matrix (1 missclassification of a late sample) was always obtained with either 5 or 20 variables per tree, with the number of trees ranging from 2000 to 800000. This shows that, at least for this dataset, corresponding the number of trees in a forest to the number of possible combinations of features was excessive and a lower value should be picked as it provides the same results while being less computationally intensive. It should however be noted that this effect may be due to the very high GINI index value of the NtrC.DctD family, as it is possible to determine the class of a sample by simply taking into account the relative abundance of this family, something which will be further discussed below. In addition, several different divisions of the samples were created in an attempt to create a RF classifier with a confusion matrix with a smaller OOB error rate than the Early vs Late division RF in an attempt to determine whether a more optimal division of the samples could be found. One such example was the grouping of all samples except the ones taken at 96 days and comparing them against the samples taken at 96 days, given that, as demonstrated in 4.10 and the pairwise t-tests in table 4.3, the samples taken at this time point seem to have higher variance than at any other period. All different subdivisions, however, proved unsuccessful as all forests generated OOB errors above 10%. These results can be found in annex A.3.

The mean decrease GINI values for the 2 RFs with Early and Late time groups are plotted in figures 4.11 and 4.12, and the relevant families are determined by evaluating which families contribute more than 5% to the sum of Mean GINI values so far. This results in 7 families being selected as relevant(NtrC.DtcD, LysR, XRE, GntR, LacI.GalR, CopY and Fur). In addition, the confusion matrix obtained from both these RFs can be seen in table 4.4.

Table 4.4: Confusion Matrix following the Random forest analysis of the time groups. Out-of-bag error= 6,7%. $_a$ Early time group, consisting of the samples taken at 0, 70 and 76 days, $_b$ Late time group, consisting of the samples taken at 96 and 124 days, $_c$ The proportion of instances misclassified over the whole set of instances.

Confusion Matrix	Early$_a$	Late$_b$	Classification error$_c$
Early	**9**	**0**	**0**
Late	**1**	**5**	**0,17**

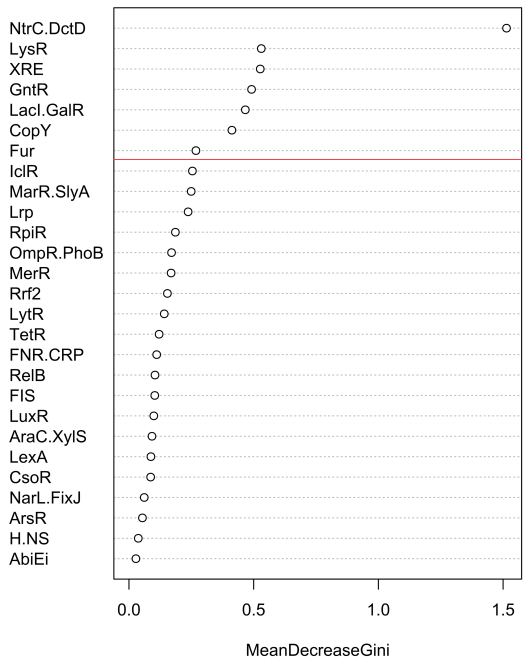

Figure 4.11: Mean Gini Decrease values for all 27 Transcription Factor families, representing their impact on the Random Forest algorithm, using 5 variables per tree. The Transcription Factor families below the red line add less than 5% for the sum of the Mean Gini Decrease values.

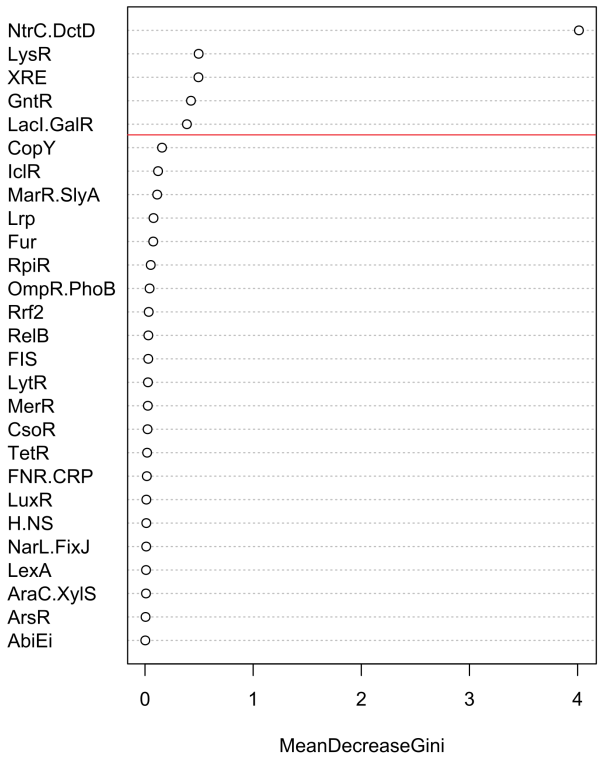

Figure 4.12: Mean Gini Decrease values for all 27 Transcription Factor families, representing their impact on the Random Forest algorithm, using 20 variables per tree. The Transcription Factor families below the red line add less than 5% for the sum of the Mean Gini Decrease values.

As can be seen, the graph created with the RandomForest with 20 variables (4.12) has a far higher value of GINI impact for the NtrC.DctD family than the RF with 5 variables per tree(4.11); the reasons for this are discussed further below. Due to higher GINI value of this family, not as many families are selected in 4.12 compared to 4.11(5 vs 7), since each subsequent family is less relevant for the overall GINI score in 4.12. It was decided to use the families selected in 4.11 as it includes all the families selected by 4.12 and was performed with a heuristic that is generally considered to be more accurate, as it de-correlates the individual trees more effectively.

4.3.2 Statistical Test

Following the selection of the families in the previous section, the LS means test was applied, with results being shown in table 4.5, where **SE** is the standard error of the sampling distribution [139] and **t-ratio** is the ratio of departure of the estimated value of a given parameter from its hypothesized value to its standard error. The greater the absolute value, the greater the confidence that coefficient observed is different than zero [139].

Table 4.5: Least square mean analysis of selected Transcription Factor families between the Early (Days$\leq$76) and Late (Days$>$76) time groups. Values in bold indicate TF families with relative abundance values statistically different between the Early and Late time groups.

TF family	p-value	SE	t ratio
NtrC.DctD	**0.00004**	0.01417	-6.08016
LysR	**0.00644**	0.02002	3.24074
XRE	0.07771	0.00388	-1.91529
GntR	0.02975	0.00678	2.44025
LacI.GalR	0.13747	0.00352	1.58285
CopY	0.05081	0.02058	-2.15160
Fur	0.20420	0.00853	1.33683

The two families selected as bioindicators are the NtrC.DctD and LysR families. In order to visualize the difference in abundance levels of these two families between early and late, a box plot was constructed, as can be seen in figure 4.13. Interestingly, the change in the number of variables tried at each tree led to no difference in the final results as to which families are considered bioindicators, since the families added by lowering this value did not have statistically significant ($p < 0.01$) differences in their early vs late abundance levels.

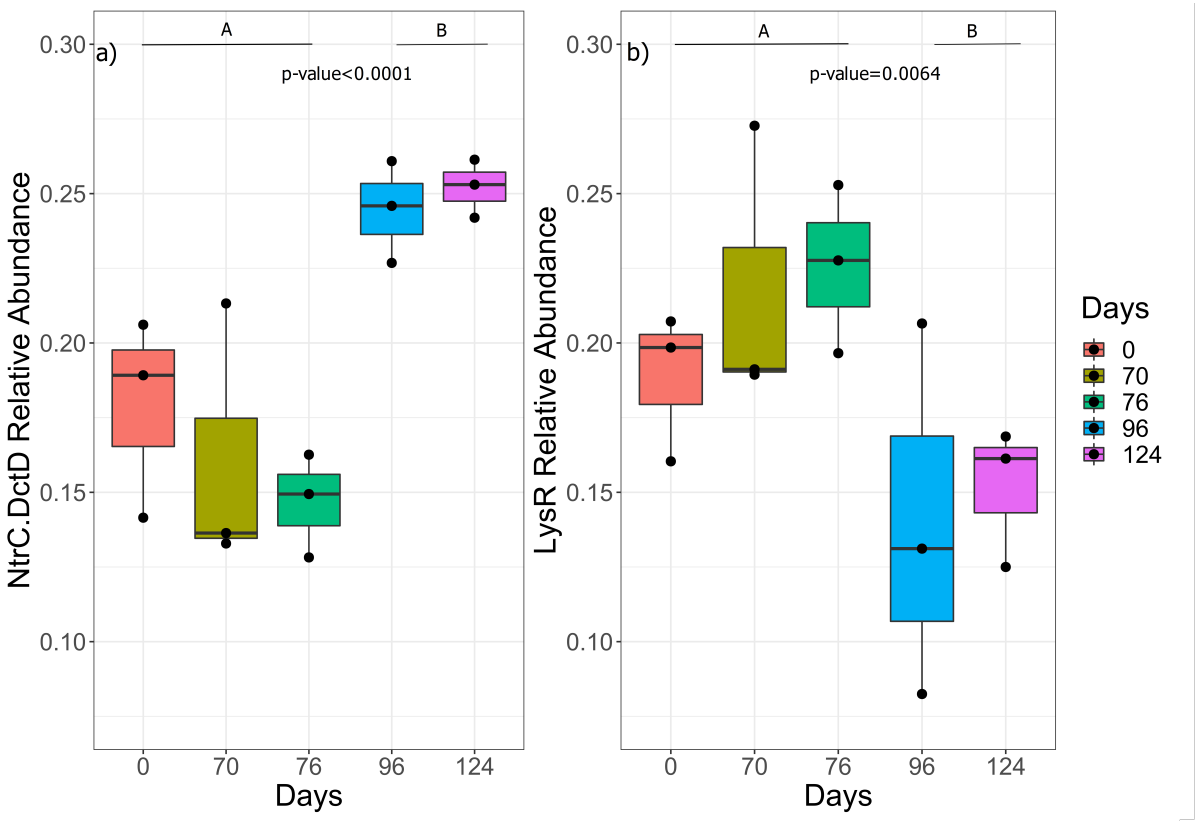

Figure 4.13: Relative abundance of the (a) NtrC.DctD and (b) LysR transcription factor families over 5 different sampling days. Different capital letters (A or B) represent a statistically significant difference between the two time groups (Early and Late), with the respective p-value represented in the graph.

As can be seen in figure 4.13, the relative abundance of the NtrC.DctD family rises in the late time group while decreasing for the LysR family. Of note, there is a clear cutoff for the abundance of the NtrC.DctD family since all early samples have a relative abundance below 22%, while all late samples have a relative abundance above 22%. This clear cutoff may account for why this family had such a high GINI value following the construction of the Random Forest, as a clear cutoff would have made any node which took into account the relative abundance of this family be able to immediately classify a sample as early or late. This effect is only exacerbated in the forest which had 20 variables per tree, as most of the trees would have this family as a node, thus increasing its impact in the decision making even further, showing why a smaller number of variables should be used in each tree, as it allows them to be further de-correlated, as previously mentioned.

This effect is further confirmed by the building of a decision tree using all families selected by the GINI plot (before the LS-means test), in order to be able to visualize the decision making process and potentially visualize some patterns regarding whether a family was up or downregulated over time. However, due to the clear cutoff previously mentioned, a decision tree which includes the NtrC.DctD family is just a single node, using that family's abundance value to distinguish between samples, with 100%

accuracy.

The NtrC.DctD family is not only relevant in the RFs, but also had relatively few subfamilies detected (7) when compared to families like LysR which had around 40 subfamilies. Therefore, an LS-means analysis on the absolute abundance values of NtrC.DctD subfamilies with more than 5 instances comparing the Late and Early time groups was conducted. This analysis could prove relevant in revealing which specific subfamilies were up- or downregulated over time. The results can be found in table 4.6.

Table 4.6: Least square mean analysis of selected Transcription Factor families between the Early (Days$\leq$76) and Late (Days$>$76) time groups. Values in bold indicate TF families with absolute abundance values statistically different between the Early and Late time groups.

NtrC.DctD Subfamily	p-value	SE	t-ratio
PILR	**0.01201**	0.68563	-2.91703
GNFM	0.16114	0.48603	1.48595
XYLR	0.71193	0.58875	0.37745
CBRB	0.05081	0.02058	-2.15160

Only one subfamily had a statistically significant($p<0,05$) difference in it's abundance between Early and Late time groups, the PILR subfamily. This subfamily is responsible for activation of the pilin gene. This gene has been implicated in playing a key role during the initial stages of colonization of a host by the pathogen *Pseudomonas aeruginosa*, by mediating attachment of the bacterium to host cell receptors [140], meaning the study of this gene is considered to be important in understanding chronic respiratory disease in patients with cystic fibrosis [140]. In order to visualize it's abundance, the following graph was produced 4.14.

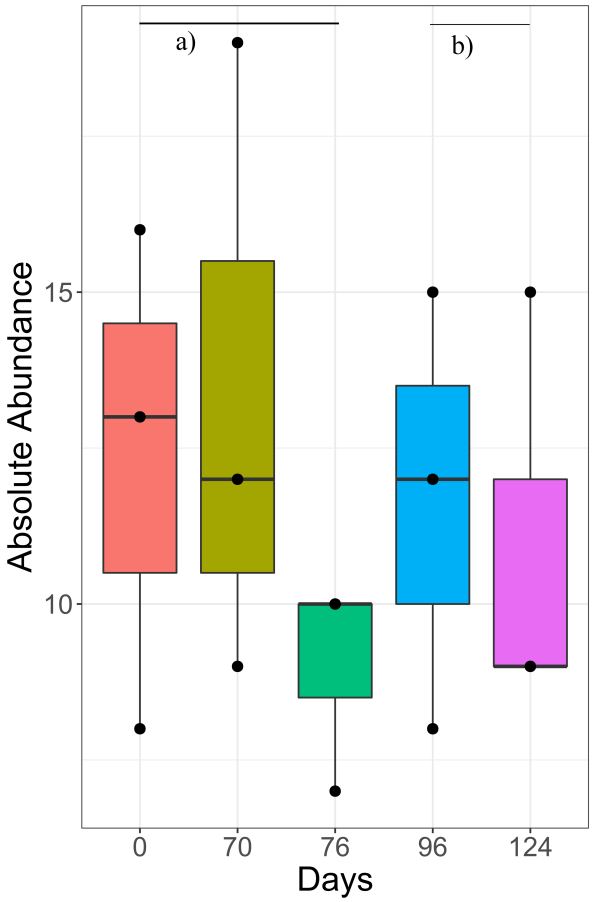

Figure 4.14: Absolute abundance of the PILR transcription factor subfamily over 5 different sampling days. Different letters (a or b) represent a statistically significant difference between the two time groups (Early and Late).

Although not as statistically significant as the results from the 2 bioindicator families, the abundance of the PILR subfamily seems to decrease over time. This could be due to a multitude of reasons: the increased stress of this environment led to the microorganisms having to ”focus” their expression to other TFs which would help them deal with the pollutant in the ecosystem; it's also possible the motility conferred by the *piln* gene activated by the PILR subfamily was not as advantageous in the new benzene-degrading conditions of the community, leading to either it not being expressed as much or even to a decrease in the population of microorganisms which express this TF, given that it has only been identified in certain species [140] [141].

The NtrC.DctD family represents transcription factors which are involved in nitrogen regulation [2]. As previously mentioned, the family NtrC.DctD has a higher relative abundance in the late time group. Considering these samples originate from a microcosm with nitrate-reducing conditions, this result may be partly attributed to environmental conditions. Due to the use of nitrate as a substrate, the pathways that regulate nitrogen consumption/nitrate reduction would be expected to have higher importance

among the different organisms present in this ecosystem. It is important to note that this family contains TFs that both upregulate and downregulate multiple different steps in the nitrogen regulation pathways and therefore an increase in the abundance levels of NtrC.DctD TFs can have multiple causes [142]. Although, it should be noted that most TFs have been shown to upregulate genes [143]. One potential cause for this increase could be that, due to the increased importance of nitrogen regulation in a nitrate-reducing environment, the organisms present within the microcosm had their NtrC.DctD transcription factors' abundance increased in order to more accurately and better regulate the nitrate-reducing pathways. This increase in abundance can be observed gradually over time, indicating why the abundance levels of the TFs are higher in samples taken at later times.

Another potential bioindicator is the LysR family, which has a statistically significant difference in its relative abundance when comparing the early and late time groups. Unlike the NtrC.DctD family, the relative abundance of the LysR family decreases for the later time group, possibly indicating a decline in the transcription or regulation of genes regulated by this family in a nitrate reducing/benzene present environment. The LysR family of transcriptional regulators is the most abundant family in the prokaryotic kingdom and regulates a very diverse set of genes, involved in functions such as virulence, metabolism, quorum sensing and motility [144]. Due to the varied amount of gene functions that the LysR family regulates, it is hard to determine whether the regulation of certain specific functions within this family is being suppressed or activated due to the environmental stresses and it is therefore harder to draw conclusions as to what specific functions could be affecting the abundance levels of TFs of this family. One way to more precisely map out which functions' TFs are being affected is to analyze the abundance levels of the LysR subfamilies, as each will have more precise functions whose importance in this system may be more easily studied. With this purpose in mind, the previously mentioned analysis on the subfamilies of all TF families was performed; however, due to the size of the dataset and in some families' (namely LysR) case, the very large number of subfamilies, there isn't enough data to draw any significant conclusions. This could potentially be remedied with a larger dataset to work with, less specific subfamily groupings, or better-quality reads. Nonetheless, a list of all LysR and NtrC.DctD TF subfamilies detected by PredicTF was made, and the functions of the genes they regulate (if known) were annotated in Supplementary Material B. Of note, we can find LysR TFs such as:

- **PCAQ-RHIRD**, a TF which activates the pcaDCHGB operon, which is involved in the catabolism of aromatic compounds. Given the high benzene concentration of this microcosm, it is possible the activation of this operon was useful or even necessary for certain microorganisms to process benzene as an energy source [145].

- **NAC-KLEAE**, a TF activator which is downregulated in the presence of nitrogen sources [146].

- **CYNR-ECOLI**, a TF activator for genes which permit *E.coli* to use cyanate as a sole nitrogen

source and is downregulated in the presence of nitrogen [147].

It should always be noted that an increase or decrease in the abundance of a certain TF in the genome does not necessarily equate to that TF being more or less translated (though the 2 are often correlated) [42]. Therefore an analysis of the transcriptome and proteome (as was originally planned to be performed in this work) can accompany this sort of work in order to more accurately determine whether or not these TFs are indeed being transcribed into RNA and proteins.

The technique performed in this work for TF analysis has proved useful, as it allows a better understanding of regulatory networks and the mechanisms which govern TF abundance, helping in the modeling and description of these complex networks, by providing insight that techniques such as phylogenetic or gene expression analyses do not, namely TF abundance values and how they are affected.

5

Conclusions and future work

The main objectives of this work were to determine the impact of environmental stresses on the abundance of bacterial TFs and whether these differences in abundance levels could be used to distinguish between samples at different time points of a succession experiment, thus providing direct insight into the regulatory networks of the community. With this objective in mind, data from a microcosm experiment was used as input for the newly developed PredicTF tool in order to detect TF abundance. These abundance values are then used to perform a diversity analysis and to determine bioindicators using Random Forests and statistical tests.

The diversity analysis results, mainly the beta diversity analysis using PCoA, show that the TF family abundance levels of the community are indeed changing over time, indicating that external factors do indeed affect bacterial TF abundance, as had been shown in literature before [31]. In addition, this analysis also showed that there appears to be a clear contrast between samples taken earlier than 96 days and those taken later.

When searching for bioindicators, 2 TF families were found whose abundance altered significantly over time: NtrC.DctD and LysR. While the NtrC.DctD change in abundance may be caused by the nitrate reducing environment of the microcosm experiment the differences in abundance of LysR TFs may be harder to explain, due to the very wide variety of functions different TFs in this family regulate. Regardless, these results show that not only does TF abundance change as a result of external factors, but this fact may potentially be used in the realms of ecology, given that if the abundance of specific families changes with specific external factors (e.g. overabundance of NtrC.DctD TFs in the presence of a nitrate reducing environment), then the abundance of these families may be used as a tool for better monitoring of the regulatory networks of microbial communities.

In this work, the initial reads were rarefied, in part due to the fact that at least part of the analysis would use absolute abundance data. However, considering that a lot of the work ended up being completed using relative abundance data, which may ,as previously mentioned, be affected by this initial normalization step, any future work should take into account whether or not normalizing the initial reads is a necessary step.

When determining the statistical significance of the Beta Diversity values a PERMANOVA analysis was performed. However, after the grouping into the Early and Late time groups, an unbalanced dataset was created (one of the groups had 9 samples while the other had 6). Due to this, an analysis using ANOSIM may have proven more accurate, as it has been shown to better handle unbalanced datasets when compared to PERMANOVA [133].

As previously stated, just because a TF is present in the metagenome, this does not mean it is necessarily transcribed into a protein. In future work one could complement the metagenomic samples with transcriptomic (determine whether the TF genes are being transcribed) and proteomic (determine whether the TFs are being translated into proteins) samples in order to have a more accurate under-

standing of which TFs are in fact being expressed. In fact, a proteomics and transcriptomics analysis was part of the original objective of this work, however, due to poor quality proteomics data and primarily the extended lockdowns in Germany due to the COVID-19 pandemic, which barred access to the necessary equipment, it was not possible to perform these extra steps. In addition, in this work due to the limited nature of the dataset, the abundance values of the TF subfamilies were not extensively used, namely in the search for bioindicators. Despite this, the abundance values of the subfamilies have the potential to be more relevant than the more generalized family groupings and should not be disregarded immediately in any future work, especially if a larger dataset with greater abundance of subfamilies can be produced. This is due to the subfamilies affecting the expression of a smaller amount of genes, often with similar functions, so a variation on the abundance of a specific subfamily can be more easily linked to a specific function or characteristic of the community.

Another analysis that may be performed alongside this work is a phylogenetic analysis of the species present in the microbiome. Such an analysis provides insight into the microbial composition of the community over time. When coupled with the PredicTF tool, it may allow for determination of, for instance, whether a decrease of the abundance of a specific TF is due to that TF not being upregulated in the conditions of microcosm, or, due to the population of microorganisms in which that TF is present decreasing or disappearing over time [148]. A phylogenetic analysis could have provided such insight into the causes of the decreased abundance of the PILR subfamily over time.

Bibliography

[1] S. Widder, R. J. Allen, T. Pfeiffer, T. P. Curtis, C. Wiuf, W. T. Sloan, O. X. Cordero, S. P. Brown, B. Momeni, W. Shou *et al.*, "Challenges in microbial ecology: building predictive understanding of community function and dynamics," *The ISME journal*, vol. 10, no. 11, pp. 2557–2568, 2016.

[2] L. M. O. Monteiro, J. Saraiva, R. B. Toscan, P. F. Stadler, R. Silva-Rocha, and U. N. da Rocha, "Predictf: a tool to predict bacterial transcription factors in complex microbial communities," *bioRxiv*, 2021.

[3] L. J. Hawkins, R. Al-Attar, and K. B. Storey, "Transcriptional regulation of metabolism in disease: From transcription factors to epigenetics," *PeerJ*, vol. 6, p. e5062, 2018.

[4] U. Consortium *et al.*, "Uniprot: the universal protein knowledgebase," *Nucleic acids research*, vol. 46, no. 5, p. 2699, 2018.

[5] S. Kılıç, E. R. White, D. M. Sagitova, J. P. Cornish, and I. Erill, "Collectf: a database of experimentally validated transcription factor-binding sites in bacteria," *Nucleic acids research*, vol. 42, no. D1, pp. D156–D160, 2014.

[6] D. Metze, D. Popp, L. Schwab, N.-S. Keller, U. N. da Rocha, H.-H. Richnow, and C. Vogt, "Temperature management potentially affects carbon mineralization capacity and microbial community composition of a shallow aquifer," *FEMS Microbiology Ecology*, 2020.

[7] "Transcription and translation, visited on 14/03/2021," https://microbenotes.com/prokaryotic-transcription-enzymes-steps-significance/, 2018.

[8] G. A. Maston, S. K. Evans, and M. R. Green, "Transcriptional regulatory elements in the human genome," *Annu. Rev. Genomics Hum. Genet.*, vol. 7, pp. 29–59, 2006.

[9] S. A. Lambert, A. Jolma, L. F. Campitelli, P. K. Das, Y. Yin, M. Albu, X. Chen, J. Taipale, T. R. Hughes, and M. T. Weirauch, "The human transcription factors," *Cell*, vol. 172, no. 4, pp. 650–665, 2018.

[10] T. I. Lee and R. A. Young, “Transcriptional regulation and its misregulation in disease,” *Cell*, vol. 152, no. 6, pp. 1237–1251, 2013.

[11] H. Singh, A. A. Khan, and A. R. Dinner, “Gene regulatory networks in the immune system,” *Trends in immunology*, vol. 35, no. 5, pp. 211–218, 2014.

[12] A. P. Fong and S. J. Tapscott, “Skeletal muscle programming and re-programming,” *Current opinion in genetics & development*, vol. 23, no. 5, pp. 568–573, 2013.

[13] K. Takahashi and S. Yamanaka, “A decade of transcription factor-mediated reprogramming to pluripotency,” *Nature reviews Molecular cell biology*, vol. 17, no. 3, p. 183, 2016.

[14] R. G. Roeder, “The role of general initiation factors in transcription by rna polymerase ii,” *Trends in biochemical sciences*, vol. 21, no. 9, pp. 327–335, 1996.

[15] D. Nikolov and S. Burley, “Rna polymerase ii transcription initiation: a structural view,” *Proceedings of the National Academy of Sciences*, vol. 94, no. 1, pp. 15–22, 1997.

[16] G. J. Narlikar, H.-Y. Fan, and R. E. Kingston, “Cooperation between complexes that regulate chromatin structure and transcription,” *Cell*, vol. 108, no. 4, pp. 475–487, 2002.

[17] L. Xu, C. K. Glass, and M. G. Rosenfeld, “Coactivator and corepressor complexes in nuclear receptor function,” *Current opinion in genetics & development*, vol. 9, no. 2, pp. 140–147, 1999.

[18] “Dna expression and regulation– protein synthesis, visited on 12/03/2021,” https://natural-universe.net/the-scientific-view-of-the-universe/the-geological-present/what-biochemistry-and-cellular-biology-tell-us/dna-expression-protein-synthesis/, 2016.

[19] A. H. Brivanlou and J. E. Darnell, “Signal transduction and the control of gene expression,” *Science*, vol. 295, no. 5556, pp. 813–818, 2002.

[20] A. A. Travers and M. Buckle, *DNA-protein interactions: a practical approach*. Practical Approach (Paperback), 2000.

[21] G. Gill, “Regulation of the initiation of eukaryotic transcription,” *Essays in biochemistry*, vol. 37, pp. 33–44, 2001.

[22] D. F. Browning and S. J. Busby, “The regulation of bacterial transcription initiation,” *Nature Reviews Microbiology*, vol. 2, no. 1, pp. 57–65, 2004.

[23] S. J. Browning, “Local and global regulation of transcription initiation in bacteria,” *Nature Reviews Microbiology*, vol. 14, no. 10, p. 638, 2016.

[24] D. F. Browning, D. C. Grainger, and S. J. Busby, "Effects of nucleoid-associated proteins on bacterial chromosome structure and gene expression," *Current opinion in microbiology*, vol. 13, no. 6, pp. 773–780, 2010.

[25] K. Maiese, "Forkhead transcription factors: new considerations for alzheimer's disease and dementia," *Journal of translational science*, vol. 2, no. 4, p. 241, 2016.

[26] D. J. McCulley and B. L. Black, "Transcription factor pathways and congenital heart disease," *Current topics in developmental biology*, vol. 100, pp. 253–277, 2012.

[27] C. R. Scherzer, J. A. Grass, Z. Liao, I. Pepivani, B. Zheng, A. C. Eklund, P. A. Ney, J. Ng, M. McGoldrick, B. Mollenhauer *et al.*, "Gata transcription factors directly regulate the parkinson's disease-linked gene α-synuclein," *Proceedings of the National Academy of Sciences*, vol. 105, no. 31, pp. 10 907–10 912, 2008.

[28] H. Bolouri, *Computational modeling of gene regulatory networks-a primer*. World Scientific Publishing Company, 2008.

[29] A. Irrthum, L. Wehenkel, P. Geurts *et al.*, "Inferring regulatory networks from expression data using tree-based methods," *PloS one*, vol. 5, no. 9, p. e12776, 2010.

[30] M. Latorre, J. Galloway-Peña, J. Roh, M. Budinich, A. Reyes-Jara, B. Murray, A. Maass, and M. González, "Enterococcus faecalis reconfigures its transcriptional regulatory network activation at different copper levels," *Metallomics : integrated biometal science*, vol. 6, 01 2014.

[31] D. F. Browning, M. Butala, and S. J. Busby, "Bacterial transcription factors: regulation by pick "n" mix," *Journal of molecular biology*, vol. 431, no. 20, pp. 4067–4077, 2019.

[32] E. Mancera, I. Nocedal, S. Hammel, M. Gulati, K. F. Mitchell, D. R. Andes, C. J. Nobile, G. Butler, and A. D. Johnson, "Evolution of the complex transcription network controlling biofilm formation in candida species," *Elife*, vol. 10, p. e64682, 2021.

[33] A. Rauch and S. Mandrup, "Transcriptional networks controlling stromal cell differentiation," *Nature Reviews Molecular Cell Biology*, pp. 1–18, 2021.

[34] J.-Y. Lee, J. Colinas, J. Y. Wang, D. Mace, U. Ohler, and P. N. Benfey, "Transcriptional and post-transcriptional regulation of transcription factor expression in arabidopsis roots," *Proceedings of the National Academy of Sciences*, vol. 103, no. 15, pp. 6055–6060, 2006.

[35] H. Liu, W. Xu, V. M. Bruno, Q. T. Phan, N. V. Solis, C. A. Woolford, R. L. Ehrlich, A. C. Shetty, C. McCraken, J. Lin *et al.*, "Determining aspergillus fumigatus transcription factor expression and

function during invasion of the mammalian lung," *PLoS pathogens*, vol. 17, no. 3, p. e1009235, 2021.

[36] A. F. Samad, M. Sajad, N. Nazaruddin, I. A. Fauzi, A. Murad, Z. Zainal, and I. Ismail, "Microrna and transcription factor: key players in plant regulatory network," *Frontiers in plant science*, vol. 8, p. 565, 2017.

[37] N. Ohama, H. Sato, K. Shinozaki, and K. Yamaguchi-Shinozaki, "Transcriptional regulatory network of plant heat stress response," *Trends in plant science*, vol. 22, no. 1, pp. 53–65, 2017.

[38] Z. Xie, T. M. Nolan, H. Jiang, and Y. Yin, "Ap2/erf transcription factor regulatory networks in hormone and abiotic stress responses in arabidopsis," *Frontiers in plant science*, vol. 10, p. 228, 2019.

[39] E. J. Stewart, "Growing unculturable bacteria," *Journal of bacteriology*, vol. 194, no. 16, pp. 4151–4160, 2012.

[40] A. Mohsen, J. Park, Y.-A. Chen, H. Kawashima, and K. Mizuguchi, "Impact of quality trimming on the efficiency of reads joining and diversity analysis of illumina paired-end reads in the context of qiime1 and qiime2 microbiome analysis frameworks," *BMC bioinformatics*, vol. 20, no. 1, pp. 1–10, 2019.

[41] E. R. Mardis, "A decade's perspective on dna sequencing technology," *Nature*, vol. 470, no. 7333, pp. 198–203, 2011.

[42] P. Hugenholtz and G. W. Tyson, "Metagenomics," *Nature*, vol. 455, no. 7212, pp. 481–483, 2008.

[43] N. D. Olson, T. J. Treangen, C. M. Hill, V. Cepeda-Espinoza, J. Ghurye, S. Koren, and M. Pop, "Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes," *Briefings in bioinformatics*, vol. 20, no. 4, pp. 1140–1150, 2019.

[44] A. Ameur, W. P. Kloosterman, and M. S. Hestand, "Single-molecule sequencing: towards clinical applications," *Trends in biotechnology*, vol. 37, no. 1, pp. 72–85, 2019.

[45] S. L. Amarasinghe, S. Su, X. Dong, L. Zappia, M. E. Ritchie, and Q. Gouil, "Opportunities and challenges in long-read sequencing data analysis," *Genome biology*, vol. 21, no. 1, pp. 1–16, 2020.

[46] O. Oluwadare, M. Highsmith, and J. Cheng, "An overview of methods for reconstructing 3-d chromosome and genome structures from hi-c data," *Biological procedures online*, vol. 21, no. 1, pp. 1–20, 2019.

[47] T. Cavaller-Smith, “The evolution of genome size,” 1985.

[48] T. Thomas, J. Gilbert, and F. Meyer, “Metagenomics-a guide from sampling to data analysis,” *Microbial informatics and experimentation*, vol. 2, no. 1, pp. 1–12, 2012.

[49] M. Pop, A. Phillippy, A. L. Delcher, and S. L. Salzberg, “Comparative genome assembly,” *Briefings in bioinformatics*, vol. 5, no. 3, pp. 237–248, 2004.

[50] S. A. Stanhope, “Occupancy modeling, maximum contig size probabilities and designing metagenomics experiments,” *PLoS One*, vol. 5, no. 7, p. e11652, 2010.

[51] M. Piper, “In-depth-ngs-data-analysis-course,” https://github.com/hbctraining/In-depth-NGS-Data-Analysis-Course, 2014.

[52] C. Kingsford, M. C. Schatz, and M. Pop, “Assembly complexity of prokaryotic genomes using short reads,” *BMC bioinformatics*, vol. 11, no. 1, pp. 1–11, 2010.

[53] N. Nagarajan and M. Pop, “Parametric complexity of sequence assembly: theory and applications to next generation sequencing,” *Journal of computational biology*, vol. 16, no. 7, pp. 897–908, 2009.

[54] A. Mikheenko, V. Saveliev, and A. Gurevich, “Metaquast: evaluation of metagenome assemblies,” *Bioinformatics*, vol. 32, no. 7, pp. 1088–1090, 2016.

[55] D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, and G. W. Tyson, “Checkm: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes,” *Genome research*, vol. 25, no. 7, pp. 1043–1055, 2015.

[56] L. Fattorini, “Statistical analysis of ecological diversity,” *Environmetrics; El-Shaarawi, AH, Jureckova, J., Eds*, pp. 18–29, 2003.

[57] E. Pielou, “The latitudinal spans of seaweed species and their patterns of overlap,” *Journal of Biogeography*, pp. 299–311, 1977.

[58] N. L. Bachmann, R. J. Rockett, V. J. Timms, and V. Sintchenko, “Advances in clinical sample preparation for identification and characterization of bacterial pathogens using metagenomics,” *Frontiers in public health*, vol. 6, p. 363, 2018.

[59] N. Conceição-Neto, M. Zeller, H. Lefrère, P. De Bruyn, L. Beller, W. Deboutte, C. K. Yinda, R. Lavigne, P. Maes, M. Van Ranst *et al.*, “Modular approach to customise sample preparation procedures for viral metagenomics: a reproducible protocol for virome analysis,” *Scientific reports*, vol. 5, no. 1, pp. 1–14, 2015.

[60] H. Shimadzu and R. Darnell, “Attenuation of species abundance distributions by sampling,” *Royal Society open science*, vol. 2, no. 4, p. 140219, 2015.

[61] A. D. Willis, “Rarefaction, alpha diversity, and statistics,” *Frontiers in microbiology*, vol. 10, p. 2407, 2019.

[62] H. Tuomisto, “A diversity of beta diversities: straightening up a concept gone awry. part 1. defining beta diversity as a function of alpha and gamma diversity,” *Ecography*, vol. 33, no. 1, pp. 2–22, 2010.

[63] H. Toumisto, “A diversity of beta diversities: straightening up a concept gone awry. part 2. quantifying beta diversity and related phenomena,” *Ecography*, vol. 33, no. 1, pp. 23–45, 2010.

[64] S. M. Holland, “Non-metric multidimensional scaling (mds),” *Department of Geology, University of Georgia, Athens, Tech. Rep. GA*, pp. 30 602–2501, 2008.

[65] S. Dray, P. Legendre, and P. R. Peres-Neto, “Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (pcnm),” *Ecological modelling*, vol. 196, no. 3-4, pp. 483–493, 2006.

[66] R. Gorelick, “Commentary: Do we have a consistent terminology for species diversity? the fallacy of true diversity,” *Oecologia*, vol. 167, no. 4, pp. 885–888, 2011.

[67] M. J. Anderson, T. O. Crist, J. M. Chase, M. Vellend, B. D. Inouye, A. L. Freestone, N. J. Sanders, H. V. Cornell, L. S. Comita, K. F. Davies *et al.*, “Navigating the multiple meanings of β diversity: a roadmap for the practicing ecologist,” *Ecology letters*, vol. 14, no. 1, pp. 19–28, 2011.

[68] P. Koleff, K. J. Gaston, and J. J. Lennon, “Measuring beta diversity for presence–absence data,” *Journal of Animal Ecology*, vol. 72, no. 3, pp. 367–382, 2003.

[69] G. Jurasinski and M. Koch, “Commentary: do we have a consistent terminology for species diversity? we are on the way,” *Oecologia*, vol. 167, no. 4, pp. 893–902, 2011.

[70] R. Lande, “Statistics and partitioning of species diversity, and similarity among multiple communities,” *Oikos*, pp. 5–13, 1996.

[71] S. H. Hurlbert, “The nonconcept of species diversity: a critique and alternative parameters,” *Ecology*, vol. 52, no. 4, pp. 577–586, 1971.

[72] S. Weiss, Z. Xu, S. Peddada, A. Amir, K. Bittinger, A. Gonzalez, C. Lozupone, J. Zaneveld, Y. Vázquez-Baeza, A. Birmingham *et al.*, “Normalization and microbial differential abundance strategies depend upon data characteristics. microbiome 5: 27,” 2017.

[73] P. J. McMurdie and S. Holmes, “Waste not, want not: why rarefying microbiome data is inadmissible,” *PLoS Comput Biol*, vol. 10, no. 4, p. e1003531, 2014.

[74] L. Beule and P. Karlovsky, “Improved normalization of species count data in ecology by scaling with ranked subsampling (srs): application to microbial communities,” *PeerJ*, vol. 8, p. e9593, 2020.

[75] E. S. Cameron, P. J. Schmidt, B. J.-M. Tremblay, M. B. Emelko, and K. M. Müller, “To rarefy or not to rarefy: Enhancing microbial community analysis through next-generation sequencing,” *bioRxiv*, 2020.

[76] J. F. Abril and S. Castellano Hereza, “Genome annotation.” Elsevier, 2019.

[77] D.-H. Le, “Machine learning-based approaches for disease gene prediction,” *Briefings in functional genomics*, vol. 19, no. 5-6, pp. 350–363, 2020.

[78] M. D. Saçar and J. Allmer, “Machine learning methods for microrna gene prediction,” in *miRNomics: MicroRNA Biology and Computational Analysis*. Springer, 2014, pp. 177–187.

[79] K. J. Hoff, M. Tech, T. Lingner, R. Daniel, B. Morgenstern, and P. Meinicke, “Gene prediction in metagenomic fragments: a large scale machine learning approach,” *BMC bioinformatics*, vol. 9, no. 1, pp. 1–14, 2008.

[80] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, “Machine learning applications in cancer prognosis and prediction,” *Computational and structural biotechnology journal*, vol. 13, pp. 8–17, 2015.

[81] P. Luo, Y. Li, L.-P. Tian, and F.-X. Wu, “Enhancing the prediction of disease–gene associations with multimodal deep learning,” *Bioinformatics*, vol. 35, no. 19, pp. 3735–3742, 2019.

[82] A. Al-Ajlan and A. El Allali, “Cnn-mgp: Convolutional neural networks for metagenomics gene prediction,” *Interdisciplinary Sciences: Computational Life Sciences*, vol. 11, no. 4, pp. 628–635, 2019.

[83] F. Schmidt, F. Kern, P. Ebert, N. Baumgarten, and M. H. Schulz, “Tepic 2—an extended framework for transcription factor binding prediction and integrative epigenomic analysis,” *Bioinformatics*, vol. 35, no. 9, pp. 1608–1609, 2019.

[84] F. Jing, S. Zhang, Z. Cao, and S. Zhang, “An integrative framework for combining sequence and epigenomic data to predict transcription factor binding sites using deep learning,” *IEEE/ACM transactions on computational biology and bioinformatics*, 2019.

[85] G. Rätsch, S. Sonnenburg, J. Srinivasan, H. Witte, K.-R. Müller, R.-J. Sommer, and B. Schölkopf, "Improving the caenorhabditis elegans genome annotation using machine learning," *PLoS Comput Biol*, vol. 3, no. 2, p. e20, 2007.

[86] M. A. Shipp, K. N. Ross, P. Tamayo, A. P. Weng, J. L. Kutok, R. C. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G. S. Pinkus *et al.*, "Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning," *Nature medicine*, vol. 8, no. 1, pp. 68–74, 2002.

[87] O. Aromolaran, T. Beder, M. Oswald, J. Oyelade, E. Adebiyi, and R. Koenig, "Essential gene prediction in drosophila melanogaster using machine learning approaches based on sequence and functional features," *Computational and structural biotechnology journal*, vol. 18, pp. 612–621, 2020.

[88] T. M. Mitchell *et al.*, "Machine learning," 1997.

[89] J. R. Koza, F. H. Bennett, D. Andre, and M. A. Keane, "Automated design of both the topology and sizing of analog electrical circuits using genetic programming," in *Artificial Intelligence in Design'96*. Springer, 1996, pp. 151–170.

[90] S. Russell and P. Norvig, "Artificial intelligence: a modern approach," 2002.

[91] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning*. MIT press, 2018.

[92] E. Alpaydin, *Introduction to machine learning*. MIT press, 2020.

[93] D. M. Hawkins, "The problem of overfitting," *Journal of chemical information and computer sciences*, vol. 44, no. 1, pp. 1–12, 2004.

[94] J. Mairal, "Sparse coding for machine learning, image processing and computer vision," Ph.D. dissertation, Cachan, Ecole normale supérieure, 2010.

[95] E. G. Dada, J. S. Bassi, H. Chiroma, A. O. Adetunmbi, O. E. Ajibuwa *et al.*, "Machine learning for email spam filtering: review, approaches and open research problems," *Heliyon*, vol. 5, no. 6, p. e01802, 2019.

[96] J. Stilgoe, "Machine learning, social learning and the governance of self-driving cars," *Social studies of science*, vol. 48, no. 1, pp. 25–56, 2018.

[97] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[98] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, S. Y. Philip *et al.*, "Top 10 algorithms in data mining," *Knowledge and information systems*, vol. 14, no. 1, pp. 1–37, 2008.

[99] H. Wagner, H. Köke, S. Dähne, S. Niemann, C. Hühne, and R. Khakimova, "Decision tree-based machine learning to optimize the laminate stacking of composite cylinders for maximum buckling load and minimum imperfection sensitivity," *Composite Structures*, vol. 220, pp. 45–63, 2019.

[100] Y. Qi, "Random forest for bioinformatics," in *Ensemble machine learning*. Springer, 2012, pp. 307–323.

[101] "Machine learning decision tree classification algorithm - javatpoint." [Online]. Available: https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

[102] L. Rokach and O. Maimon, "Top-down induction of decision trees classifiers-a survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 35, no. 4, pp. 476–487, 2005.

[103] R. I. Lerman and S. Yitzhaki, "A note on the calculation and interpretation of the gini index," *Economics Letters*, vol. 15, no. 3-4, pp. 363–368, 1984.

[104] P. M. Rosado, D. C. Leite, G. A. Duarte, R. M. Chaloub, G. Jospin, U. N. da Rocha, J. P. Saraiva, F. Dini-Andreote, J. A. Eisen, D. G. Bourne *et al.*, "Marine probiotics: increasing coral resistance to bleaching through microbiome manipulation," *The ISME journal*, vol. 13, no. 4, pp. 921–936, 2019.

[105] T. G. Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*. Springer, 2000, pp. 1–15.

[106] J. Friedman, T. Hastie, R. Tibshirani *et al.*, *The elements of statistical learning*. Springer series in statistics New York, 2001, vol. 1, no. 10.

[107] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[108] L. Briman, "Out-of-bag estimation," 1996.

[109] S. Gupta, "Random forest (easily explained), visited on 24/05/2021," https://medium.com/@gupta020295/random-forest-easily-explained-4b8094feb90, 2020.

[110] M. W. Browne, "Cross-validation methods," *Journal of mathematical psychology*, vol. 44, no. 1, pp. 108–132, 2000.

[111] S. Janitza and R. Hornung, “On the overestimation of random forest’s out-of-bag error,” *PloS one*, vol. 13, no. 8, p. e0201904, 2018.

[112] L. Deng and D. Yu, “Deep learning: methods and applications,” *Foundations and trends in signal processing*, vol. 7, no. 3–4, pp. 197–387, 2014.

[113] F. Bre, J. M. Gimenez, and V. D. Fachinotti, “Prediction of wind pressure coefficients on building surfaces using artificial neural networks,” *Energy and Buildings*, vol. 158, pp. 1429–1441, 2018.

[114] M. M. Waldrop, “News feature: What are the limits of deep learning?” *Proceedings of the National Academy of Sciences*, vol. 116, no. 4, pp. 1074–1077, 2019.

[115] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, “Predicting the sequence specificities of dna-and rna-binding proteins by deep learning,” *Nature biotechnology*, vol. 33, no. 8, pp. 831–838, 2015.

[116] X. Pan and H.-B. Shen, “Rna-protein binding motifs mining with a new hybrid deep learning based cross-domain knowledge integration approach,” *BMC bioinformatics*, vol. 18, no. 1, pp. 1–14, 2017.

[117] J. Zrimec, C. S. Börlin, F. Buric, A. S. Muhammad, R. Chen, V. Siewers, V. Verendel, J. Nielsen, M. Töpel, and A. Zelezniak, “Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure,” *Nature communications*, vol. 11, no. 1, pp. 1–16, 2020.

[118] G. Arango-Argoty, E. Garner, A. Pruden, L. S. Heath, P. Vikesland, and L. Zhang, “Deeparg: a deep learning approach for predicting antibiotic resistance genes from metagenomic data,” *Microbiome*, vol. 6, no. 1, pp. 1–15, 2018.

[119] S. Min, B. Lee, and S. Yoon, “Deep learning in bioinformatics,” *Briefings in bioinformatics*, vol. 18, no. 5, pp. 851–869, 2017.

[120] A. Coates, A. Ng, and H. Lee, “An analysis of single-layer networks in unsupervised feature learning,” in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2011, pp. 215–223.

[121] Y. Sun, X. Wang, and X. Tang, “Deep learning face representation from predicting 10,000 classes,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1891–1898.

[122] F. Buggenthin, F. Buettner, P. S. Hoppe, M. Endele, M. Kroiss, M. Strasser, M. Schwarzfischer, D. Loeffler, K. D. Kokkaliaris, O. Hilsenbeck *et al.*, “Prospective identification of hematopoietic lineage choice by deep learning,” *Nature methods*, vol. 14, no. 4, pp. 403–406, 2017.

[123] L. Sørensen, M. Loog, P. Lo, H. Ashraf, A. Dirksen, R. P. Duin, and M. De Bruijne, "Image dissimilarity-based quantification of lung disease from ct," in *international conference on medical image computing and computer-assisted intervention*. Springer, 2010, pp. 37–44.

[124] P. Baldi and P. Sadowski, "The dropout learning algorithm," *Artificial intelligence*, vol. 210, pp. 78–122, 2014.

[125] B. Gao and L. Pavel, "On the properties of the softmax function with application in game theory and reinforcement learning," *arXiv preprint arXiv:1704.00805*, 2017.

[126] A. H. Keller, S. Kleinsteuber, and C. Vogt, "Anaerobic benzene mineralization by nitrate-reducing and sulfate-reducing microbial consortia enriched from the same site: comparison of community composition and degradation characteristics," *Microbial ecology*, vol. 75, no. 4, pp. 941–953, 2018.

[127] S. Nurk, D. Meleshko, A. Korobeynikov, and P. A. Pevzner, "metaspades: a new versatile metagenomic assembler," *Genome research*, vol. 27, no. 5, pp. 824–834, 2017.

[128] A. Chao, N. J. Gotelli, T. Hsieh, E. L. Sander, K. Ma, R. K. Colwell, and A. M. Ellison, "Rarefaction and extrapolation with hill numbers: a framework for sampling and estimation in species diversity studies," *Ecological monographs*, vol. 84, no. 1, pp. 45–67, 2014.

[129] P. J. McMurdie and S. Holmes, "phyloseq: an r package for reproducible interactive analysis and graphics of microbiome census data," *PloS one*, vol. 8, no. 4, p. e61217, 2013.

[130] J. Oksanen, F. G. Blanchet, R. Kindt, P. Legendre, B. O'Hara, G. Simpson, P. Solymos, H. Stevens, and H. Wagner, *Vegan: Community Ecology Package*, 01 2010, vol. 1.

[131] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002. [Online]. Available: https://CRAN.R-project.org/doc/Rnews/

[132] R. C. Team *et al.*, "R: A language and environment for statistical computing," 2013.

[133] M. J. Anderson and D. C. Walsh, "Permanova, anosim, and the mantel test in the face of heterogeneous dispersions: what null hypothesis are you testing?" *Ecological monographs*, vol. 83, no. 4, pp. 557–574, 2013.

[134] M. J. Anderson, "Permutational multivariate analysis of variance (permanova)," *Wiley statsref: statistics reference online*, pp. 1–15, 2014.

[135] P. Legendre and L. Legendre, *Numerical ecology*. Elsevier, 2012.

[136] A. Liaw, M. Wiener *et al.*, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[137] R. V. Lenth, “Using lsmeans,” *J stat Softw*, vol. 69, pp. 1–33, 2017.

[138] R. Lenth, “Least-squares means. r package ‘lsmeans’. 2016,” 2016.

[139] D. G. Altman and J. M. Bland, “Standard deviations and standard errors,” *Bmj*, vol. 331, no. 7521, p. 903, 2005.

[140] K. S. Ishimoto and S. Lory, “Identification of pilr, which encodes a transcriptional activator of the pseudomonas aeruginosa pilin gene.” *Journal of Bacteriology*, vol. 174, no. 11, pp. 3514–3521, 1992.

[141] D. Sakai and T. Komano, “The pill and piln genes of inci1 plasmids r64 and colib-p9 encode outer membrane lipoproteins responsible for thin pilus biogenesis,” *Plasmid*, vol. 43, no. 2, pp. 149–152, 2000.

[142] M. R. Atkinson, E. S. Kamberov, R. L. Weiss, and A. J. Ninfa, “Reversible uridylylation of the escherichia coli pii signal transduction protein regulates its ability to stimulate the dephosphorylation of the transcription factor nitrogen regulator i (nri or ntrc).” *Journal of Biological Chemistry*, vol. 269, no. 45, pp. 28 288–28 293, 1994.

[143] M. J. Mann, “Transcription factor decoys: a new model for disease intervention,” *Annals of the New York Academy of Sciences*, vol. 1058, no. 1, pp. 128–139, 2005.

[144] S. E. Maddocks and P. C. Oyston, “Structure and function of the lysr-type transcriptional regulator (lttr) family proteins,” *Microbiology*, vol. 154, no. 12, pp. 3609–3623, 2008.

[145] A. M. MacLean, G. MacPherson, P. Aneja, and T. M. Finan, “Characterization of the β-ketoadipate pathway in sinorhizobium meliloti,” *Applied and environmental microbiology*, vol. 72, no. 8, pp. 5403–5413, 2006.

[146] A. Schwacha and R. A. Bender, “The nac (nitrogen assimilation control) gene from klebsiella aerogenes,” *Journal of bacteriology*, vol. 175, no. 7, pp. 2107–2115, 1993.

[147] Y.-C. Sung and J. A. Fuchs, “The escherichia coli k-12 cyn operon is positively regulated by a member of the lysr family,” *Journal of bacteriology*, vol. 174, no. 11, pp. 3645–3650, 1992.

[148] J. P. Huelsenbeck, J. Bull, and C. W. Cunningham, “Combining data in phylogenetic analysis,” *Trends in Ecology & Evolution*, vol. 11, no. 4, pp. 152–158, 1996.

A

Annex A

A.1 NMDS and PCoA plots

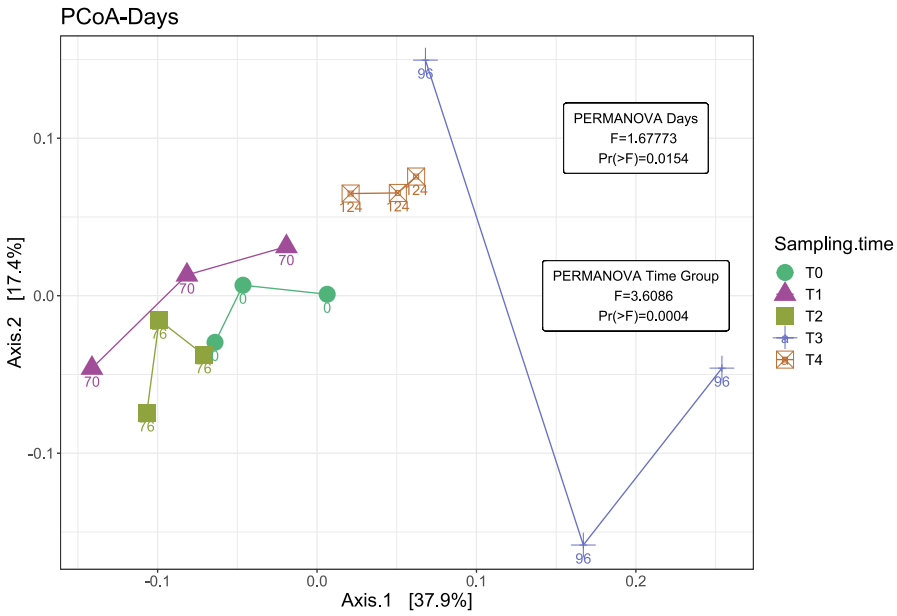

Figure A.1: Principal Coordinate analysis of Bray distances comparing the relative abundance of TF families in different samples.

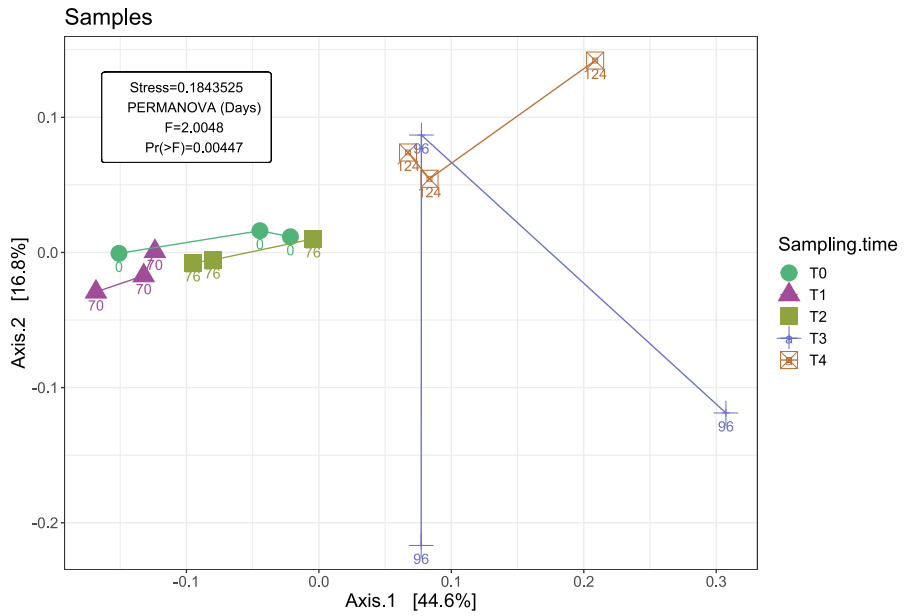

Figure A.2: Principal Coordinate analysis of Bray distances comparing the absolute abundance of TF families in different samples.

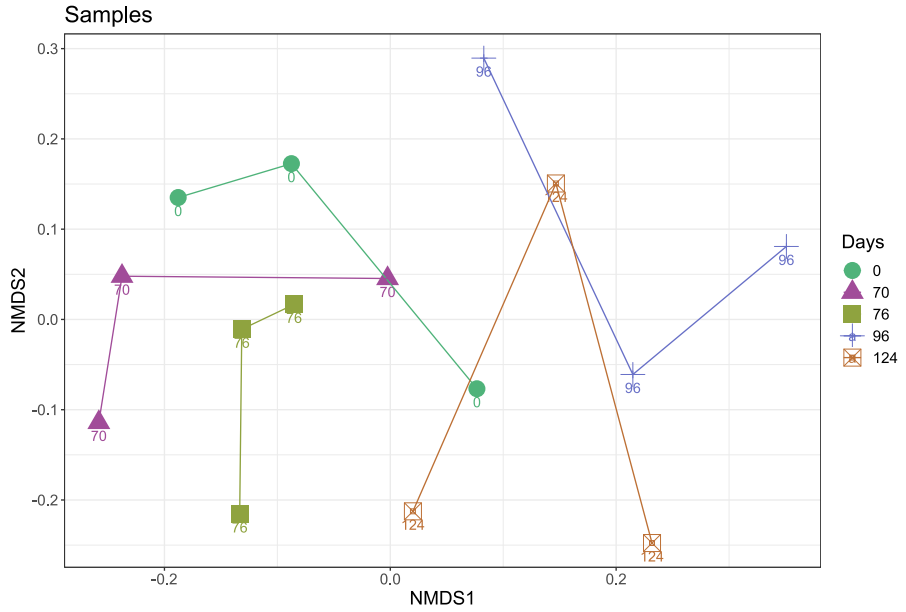

Figure A.3: NMDS analysis of Bray distances comparing the absolute abundance of TF families in different samples.

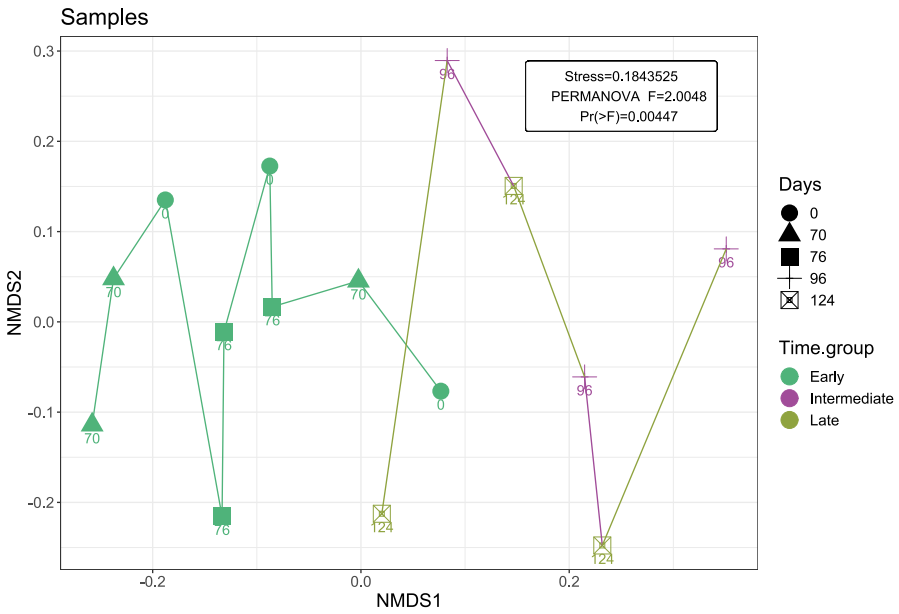

Figure A.4: NMDS analysis of Bray distances comparing the absolute abundance of TF families in different samples. We defined two time groups as Early (Days $\leq$ 76) and Late (Days$>$76).

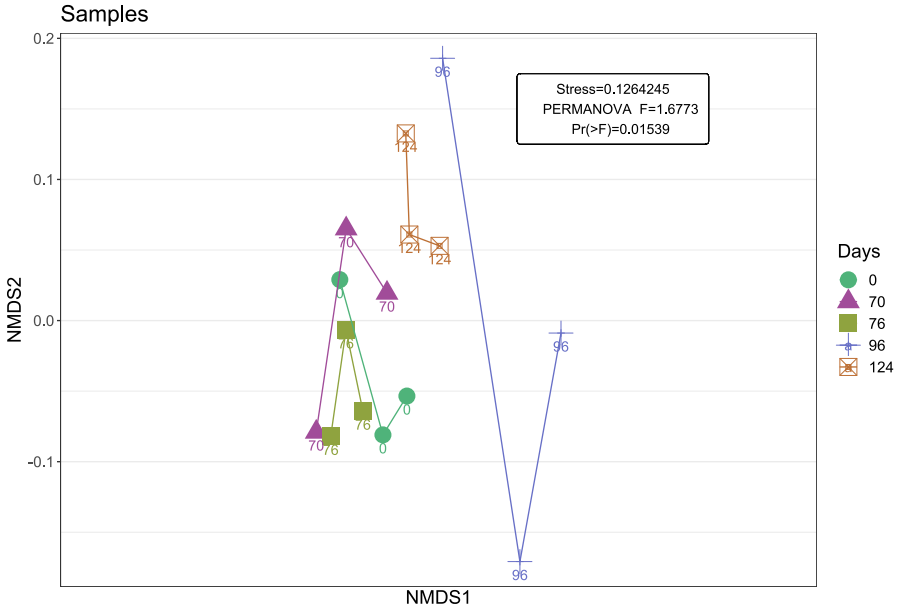

Figure A.5: NMDS analysis of Bray distances comparing the relative abundance of TF families in different samples.

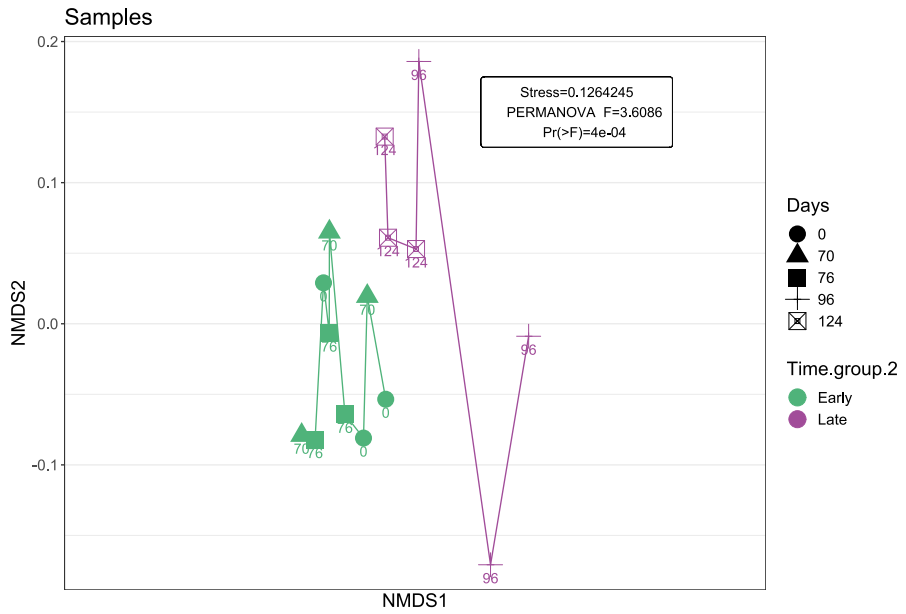

Figure A.6: NMDS analysis of Bray distances comparing the relative abundance of TF families in different samples. We defined two time groups as Early (Days $\leq$ 76) and Late (Days$>$76).

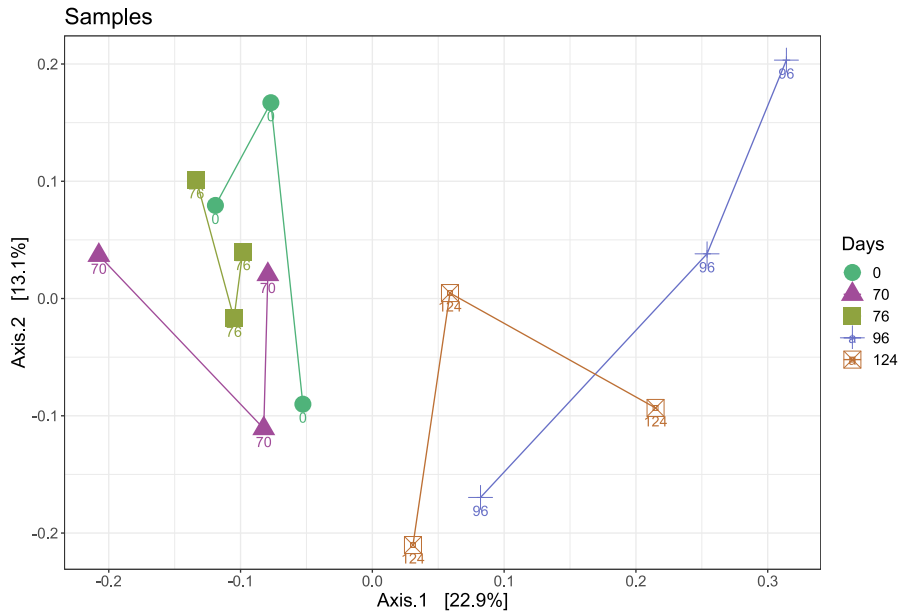

Figure A.7: Principal Coordinate analysis of Bray distances comparing the relative abundance of TF subfamilies in different samples.

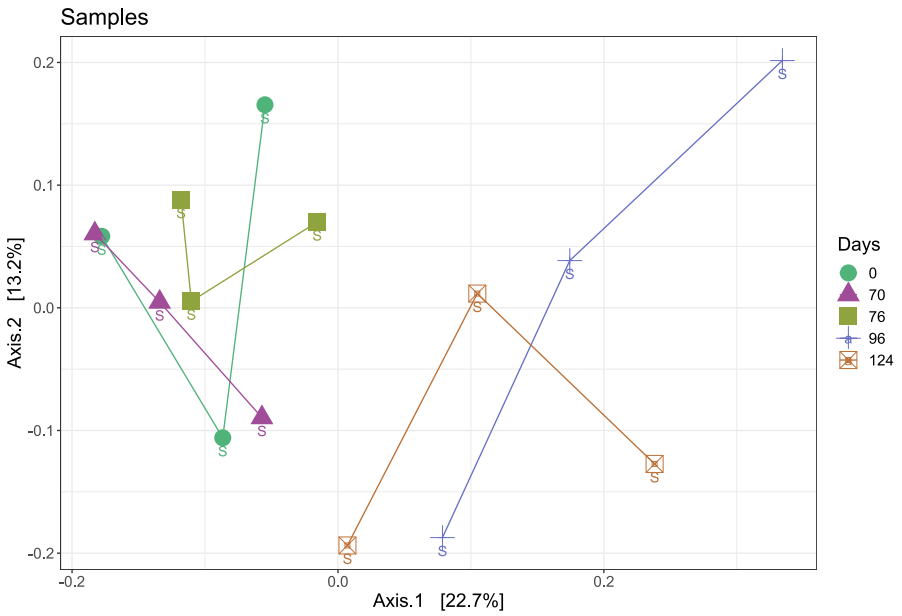

Figure A.8: Principal Coordinate analysis of Bray distances comparing the absolute abundance of TF subfamilies in different samples.

A.2 Subfamily Variance analysis

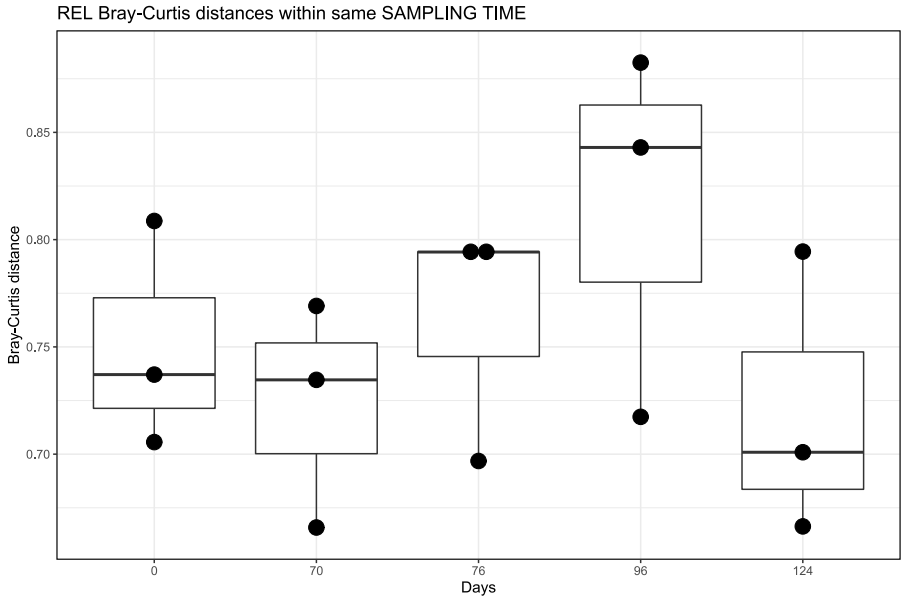

Figure A.9: Bray-Curtis dissimilarity index among the relative abundance of TF subfamilies in samples taken in the same day.

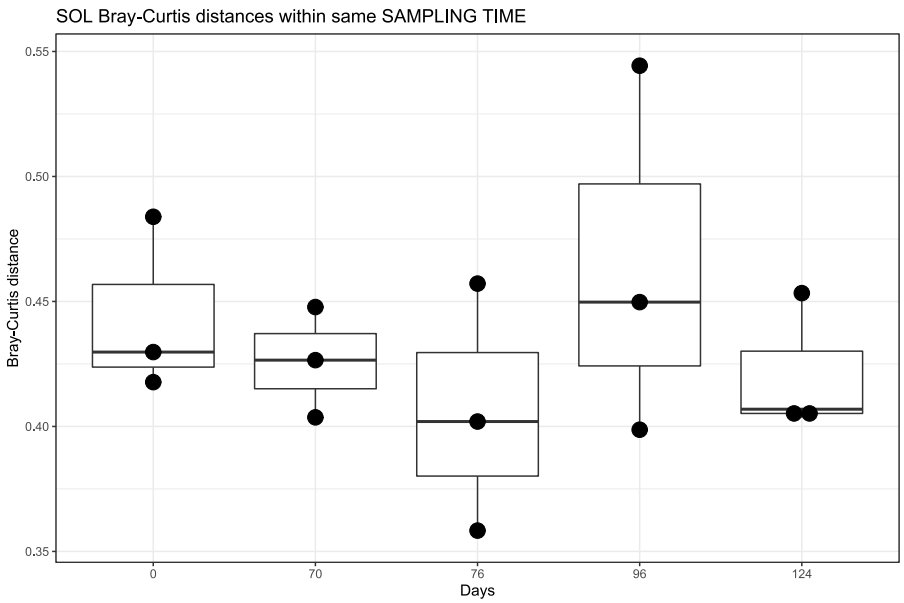

Figure A.10: Bray-Curtis dissimilarity index among the absolute abundance of TF subfamilies in samples taken in the same day.

A.3 RF Attempts

Table A.1: Sample groupings used for the building of RFs with 5 variables tested per tree and their OOB error rate

Sample grouping (Days)	Sample grouping 2 (Days)	Sample grouping 3 (Days)	OOB error(%)
Early (0,70,76)	Late(96,124)	-	6
Early + 124	96	-	13
Early	96	124	27
70 + 76	Late	-	25
70 + 76	124	-	11
Early	124	-	16