Master's thesis

2019

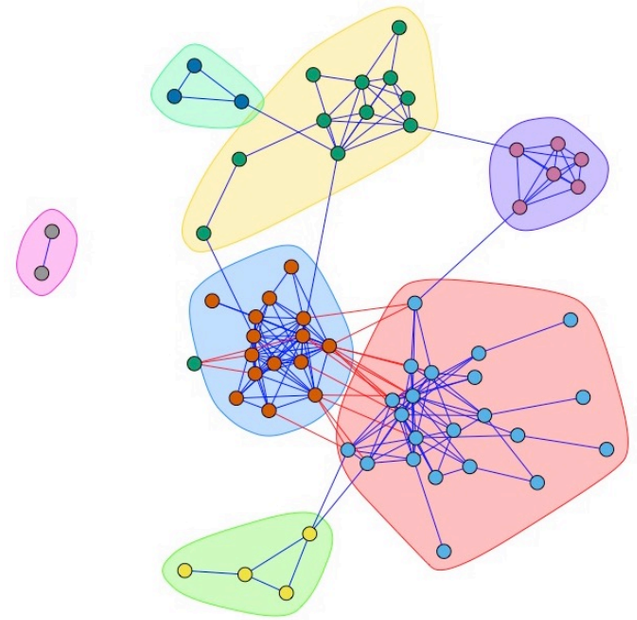
Jakob Peder Pettersen

NTNU
Norwegian University of
Science and Technology
Faculty of Natural Sciences
Department of Biotechnology and Food Science

Jakob Peder Pettersen

A study of the interactions and dynamics of microbial communities

December 2019



NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of
Science and Technology

NTNU

Norwegian University of
Science and Technology

A study of the interactions and dynamics of microbial communities

Jakob Peder Pettersen

Biotechnology - MBIOT5
Submission date: December 2019
Supervisor: Eivind Almaas
Co-supervisor: Olav Vadstein
Ingrid Bakke

Norwegian University of Science and Technology
Department of Biotechnology and Food Science

Acknownledgements

First of all, I would like to thank my main supervisor Eivind Almaas for his friendliness and good advise during this project. In addition, I have appreciated the input and feedback from my co-supervisors Olav Vadstein and Ingrid Bakke (mind you, there is no such thing as a stupid question). Without my special interest in mathematics, writing a thesis such as this one would not have been possible. I therefore thank my best friend Fredrik Østrem for awaking my mathematical talent. Also, I remind the educational politicians that spending half a year more than the normed length of study, was absolutely necessary in order to understand the underlying mathematics of this thesis. Madeleine Stenshorne Gundersen is worth a honorable mention. Not only has she struggled in the lab in order to provide the largest and best dataset used in the thesis She is also a good friend of mine and has provided me valuable insight into the experimental procedure and how to deal with the data.

Even though the main objective in my life as a student has been to obtain knowledge and academic skills, I still have a life beside the studies. In this respect, the practical and emotional support from my mother Reidun Elisabeth Simonsen has been invaluable and essential for the fact that I could complete this thesis. Last, but not least, I would like to thank my fellow students (too many to mention) for their welcoming and respectful attitude towards me, even though I am different from most of the others.

Abstract

The structure, dynamics and interactions of microbial communities have the latest years been a hot topic in the scientific communities. Industries as diverse as aquaculture, water treatment and healthcare would benefit to a major extent if we could utilize a widened understanding of community assembly and interactions to provide a microbial environment fitting our demands. Yet, the field of studying community interactions and dynamics is young and immature. No solid and well-tested theory or approach exist, encouraging further research into this topic.

The data for this project were OTU tables from three different lab-scale reactors experiments, created by 16S rDNA sequencing. ReBoot, a cross-sectional algorithm, combined with an ensemble of similarity measures, was utilized to provide the co-occurrence and co-exclusion patterns of the microbial communities. In addition, a time-series based approach was tired out; the generalized Lotka-Volterra equation was fitted with the experimental data in order to determine whether the dynamics of the communities could be reproduced.

The interactions inferred from the similarity measures through ReBoot were for the most part positive. The similarity measures had the power to distinguish OTUs being dominant under different selection regimes, even though the results from ReBoot were not any better in this respect than the naive application of similarity measures. The predicted interaction networks generally consisted of clusters of closely related OTUs having positive interactions among themselves and negative interactions to the other clusters. The observation was most likely an artifact of the fact that closely related bacteria often have almost the same niche and therefore appear in the same samples. For this reason, the co-occurrence patterns found, most likely did not correspond to causal interactions.

It was shown that the datasets possessed temporal dynamics indicating deterministic behavior and convergence toward a steady state. In addition, the algorithm managed to predict the Lotka-Volterra coefficients accurately on a small artificial test community. However, the Lotka-Volterra modeling gave no reasonable results for the experimental data, as the cross-validation to obtain the tuning parameters provided no optimum and the predicted time series did not resemble the real ones. Too long time spans between sampling points, possibly combined with a too large number of OTUs, is the most plausible reason for the failure.

Sammendrag

De seneste årene har strukturen, sammensetningen og interaksjonene i mikrobielle samfunn vært et hett tema i de vitenskapelig miljøene. En rekke næringer som havbruk, vannbehandling og helse har mye å tjene dersom en utvidet forståelse av interaksjoner mellom mikrober kan gjøre oss i stand til å forme de mikrobielle samfunnene til å bli gunstige for våre formål. Studiet av mikrobielle interaksjoner i hele samfunn er et nytt og umodent fagfelt. Derfor finnes i dag ingen vel underbygd og testet teori eller framgangsmåte, noe som fordrer videre forskning på feltet.

For prosjektet ble det brukt OTU-tabeller fra tre forskjellige lab-skala reaktorforsøk, generert ved sekvensering av 16S rDNA. ReBoot, en tversgående algoritme, kombinert med en samling av similaritetsmål, ble brukt for å finner mønstre av sameksistens og gjensidig utelukkelse. I tillegg ble en tidsseriebasert framgangsmåte prøvd ut, der forsøksdata ble tilpasset den generaliserte Lotka-Volterra-likningen, for å se om dynamikken i samfunnene kunne reproduseres.

Interaksjonene som ble funnet gjennom ReBoot, var for det meste positive. Similaritetsmålene klarte å strukturere OTUene som var dominerende under forskjellige seleksjonsregimer selv om resultatene fra ReBoot ikke var noe bedre på dette området enn direkte anvendelse av similaritetsmålene. De predikerte interaksjonsnettverkene viste en struktur med klynger av nært beslektede OTUer med positive interaksjoner mellom seg, og med negative interaksjoner mellom klyngene. Dette funnet var mest sannsynlig en følge av at OTUene i hver klynge hadde samme miljøpreferanser og forekom derfor i de samme prøvene. Av den grunn samsvarte sannsynligvis ikke sameksistensmønsteret overens med den faktiske interaksjonsstrukturen i samfunnene.

Det ble vist at datasettene hadde tidsutviklinger som indikerte deterministisk oppførsel og konvergens mot en stabil tilstand. Dessuten fungerte algoritmen godt til å predikere Lotka-Volterra-koeffisientene til et lite, simulert samfunn. Imidlertid ga modelleringen med Lotka-Volterra ingen fornuftige resultater på dataene fra forsøkene. Blant ga kryss-validering ingen optimal verdi for tilpasningsparametrene, og de predikerte tidsseriene samsvarte ikke med de reelle tidsseriene fra forsøkene. Sannsynligvis skyldtes problemene for lang tid mellom hver prøve, kanskje også at antallet OTUer var for høyt.

Table of Contents

List of Figures

List of Tables

Nomenclature

gLV	generalized Lotka-Volterra- A deterministic community model
OTU	Operational Taxonomical Unit- A taxonomical group defined on a pragmatic(operational) crierion
PCA	Principal Component Analysis- Method for reducing the variability in data to the fewest number of dimensions as possible
PCoA	Principal Coordinate Analysis- PCA-like approach based on a similarity matrix instead of raw data
PCR	Polymerase Chain Reaction- Method for amplifying a specific DNA fragment with known end sequences

Mathematical notation

The following notation is used throughout the thesis:

- $\mathbf{x}$: Vector $\begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix}$
- $\mathbf{0}$: The zero vector (all elements equal to zero)
- $\mathbf{1}$: The one vector (all elements equal to 1)
- $\underbrace{X}_{n \times m}$: Matrix (upper-case) with n rows and m columns
- $X^\intercal$: Matrix transpose (row and columns interchanged)
- $(X)_{ij}$ Element in row i and column j in matrix X
- $\sigma(\mathbf{x})$: Permutation (reordering) of the elements in the vector $\mathbf{x}$

- $\begin{bmatrix}\mathbf{x_1} & \mathbf{x_2} & \dots & \mathbf{x_m}\end{bmatrix}$: Matrix having the vectors $\mathbf{x_1}, \mathbf{x_2}, \dots, \mathbf{x_m}$ as columns
- $\operatorname{diag}(X)$: The vector containing the elements on the main diagonal of X (matrix X is assumed to be square)
- $\operatorname{diag}(\mathbf{x})$: The diagonal matrix containing the elements of $\mathbf{x}$ on its main diagonal (and all other entries to be zero)
- $\operatorname{bin}(\mathbf{x})$: The binary encoding $\mathbf{x}^*$ of $\mathbf{x}$, this is for each of the elements: $x_i^* = \begin{cases}1 & x_i \neq 0\\ 0 & x_i = 0\end{cases}$
- $R_{\mathbf{x}}$: The vector containing the ranks of each element in $\mathbf{x}$ in increasing order. Ranks are averaged if more than one element has the same value.
- $\operatorname{Tr}(X)$: Trace of a quadratic matrix X (sum of elements on the main diagonal)
- $\|\mathbf{x}\|_2$: The Euclidean norm of the vector $\mathbf{x}$, equal to $\sqrt{\mathbf{x}^\intercal\mathbf{x}}$
- $\arg\max_x f(x)$: The value(s) of x, for which the function f attains its maximum value.
- $\arg\min_x f(x)$: The value(s) of x, for which the function f attains its minimum value.
- $\ln(x)$: The natural logarithm

Reference to implementation

Whenever an implementation of a method by an `R`-package is used or discussed, it is cited on the form:
`package::function(options='option')`

Chapter 1

Introduction

During the recent years, modern high-throughput sequencing technology has enabled identification of entire microbial communities[1–3]. Doing so has the promise of finding out how the microbial communities are affected by its environment. Furthermore, this had also initiated an extensive research into how bacteria interact with each other.

1.1 Types of interactions

Bacteria and other microorganisms do not live in isolation. They affect the life of each other through interactions[1]. Interactions between organisms can be classified according to their effect on each of the two actors. In this respect, we talk about negative, neutral and positive interactions. For instance, if two microorganisms cooperate and have a positive effect on each other, we call it mutulism. On the other hand, competition, such as for food resources, provides an disadvantage for both actors, so it is a negative interaction. An interaction need not be symmetric[4], for instance often one species may benefit from the presence of the other, whereas the other species has no advantage of this (commensalism). Also, interactions can be context specific, there might be that bacteria are not interacting during lack of nutrients, or attain a mutualistic cross-feeding behavior when nutrients are present. See Figure 1.1 for a graphical overview of the different kinds of interactions.

1.2 Practical applications

Understanding of the interaction and interplay between microbes is essential for studying community assembly processes. In turn, understanding how microbial communities assemble provides great insight into the questions of which microorganism are present under specific conditions and in which abundances. Microbial

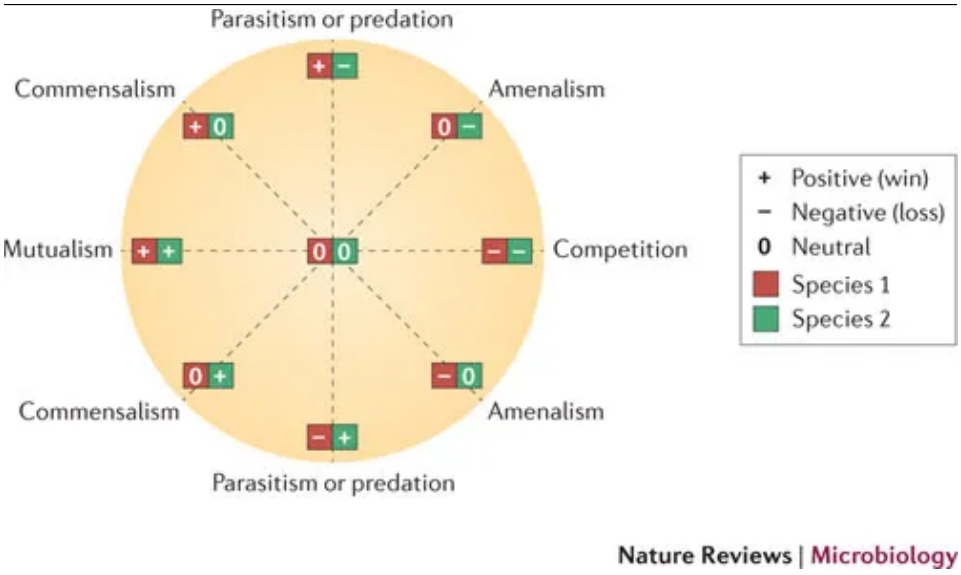


Figure 1.1: Overview of the different kinds of interactions[1]

processes shape the ecosystem and the health of all living macroorganisms. Thus, knowing how to utilize microbial interactions may have give us tools to manipulate environmental processes to our benefits or to provide good health for humans and animals. The human gut microbiota is probably the most studied subfield of microbial ecology up to present. The state of the gut of the host may influence the health condition of the entire body[5] and a disbiotic gut may lead to diseases such as inflammatory bowel disease[6], obesity[7] and allergy[8].

Among the same lines, more attention has been drawn to the microbes' role in survival and growth of larva of marine fish[9, 10]. Contrary to the traditional viewpoint that infectious diseases are caused by specific pathogens, also non-specific opportunistic bacteria may cause detrimental effects for the fish if the microbiota surrounding the fish or residing inside the gut, is brought out of balance. This includes poor growth, higher stress levels and increased mortality. Management to provide healthy microbial community for the fish can thus make aquaculture more profitable and more acceptable by the public. As a consequence, understanding how microbes interplay with each other and their hosts is therefore a major concern in the aquaculture industry.

Table 1.1: Usual properties of r- and K-strategists. Adopted from [10–12]

Trait	**r-strategist**	**K-strategist**
Maximum growth rate	High	Low
Affinity for substrate	Low	High
Competative ability at resource limitations	Low	High
Conditions favoring selection	Pulse feeding, abrupt change of environment	Continues feeding, stable environment
Stability of community	Low, community prone to collapse	High
rrn operon copies	Many	One or a few

1.3 Selection regimes

Andrew and Harris[11] first brought up the concept of r- and K-strategists to explain life strategies of microbes, although these concepts originally were developed in macro-ecology. The terms r and K originate from the logistic equation for population growth:

$$\frac{\mathrm{d}\,N}{\mathrm{d}\,t} = Nr\left(1 - \frac{N}{K}\right), \tag{1.1}$$

where N is the size of the population. According to this equation, coefficient r can be interpreted as the maximum growth rate, whereas the coefficient K is referred to as the carrying capacity. Even though the logistic equation is largely abandoned as a quantitative mean as describing r- and K-selection[11], the underlying qualitative description remains. An r-strategist is optimized for rapid growth when resources are in surplus at the expense of being able to compete for resources in crowed environments. Hence, r-strategists are most usually found in unstable communities dominated by changes. On the other hand, K-strategists are adapted for growth on limited resources, but rarely show high growth rates even when resources are abundant. They tend to dominate stable environments and can live in higher densities than the r-strategists. A summary of the common properties of K- and r-strategists are shown in Table 1.1. As pointed out by Vadstein *et al.*[9, 10], most common bacteria being know as detrimental for fish health are r-strategists. Therefore, Vadstein *et al.*[9] states that K-selection in aquaculture systems might be a good strategy for increasing viability and growth for fish at early life stages.

1.4 How is the microbiome composition determined?

The 16S ribosomal RNA gene[1] is the most common marker used in community microbiome studies. It is universally distributed, has both variable and conserved regions and follows (with a few exceptions) the phylogenetic linage. It has been utilized for a long time for determining phylogeny of microorganisms in pure cultures[13]. Furthermore, molecular techniques such as tRFLP and DGGE have been used for fingerprinting entire communities. However, identifying the relative abundance and identity of all microorganisms being present in communities has awaited recent developments in sequencing technology[2].

After DNA extraction of the community, the S16 target sequence is amplified using broad coverage primers, designed to amplify all bacteria equally. These amplicons are then sequenced on a high throughput platform such as Ilumina miSeq[14]. After sequencing, the individual reads must be organized or clustered in a meaningful way, based on the similarity of the reads. Traditionally, a similarity threshold of 97% has been set for discriminating bacterial species. However, selecting a similarity threshold for the smallest taxonomical unit to consider, is no trivial task. No clear species boundaries exist for bacteria, and in some cases, a threshold of 99% is required to distinguish bacteria believed to be different species. In order to mitigate the problem of fuzzy species boundaries, it is common to work with *Operational Taxonomical Units*(OTUs) based on a custom, self-made criterion[15]. From to the number of reads for each OTU in a sample, the relative abundance for each OTU is estimated, creating an abundance profile. Most often, studies of microbial datasets involve many samples, in which case the number of reads or relative abundances are aggregated into an OTU table.

1.5 Recent developments

Inferring microbial interactions from real 16S rDNA surveys is far from straightforward[1, 3, 16]. The available data are usually the relative abundances of microbes, sometimes combined with measurements of total cell count and other ecological parameters. Among the most basic methods are the ones based on correlations or dissimilarities. In addition, more refined methods specially adopted for relative abundances[17] exist, in addition to regression[18] and probablistic graph models[3]. Also, dynamical modelling has been attempted[19]. The latter approach proved particularly useful as it enabled finding a mechanism which prevents the growth of *Clostridium difficile* in the gut[20]. However, up to date we lack a coherent theory of microbial interactions and many of the proposed methods have turned out to have critical shortcomings[16, 21]. Neither do we know much about the structures and mechanisms shaping the microbial communities.

[1]For short we refer to this as 16S rDNA, referring to the fact that this is the DNA sequence coding for the 16S rRNA

1.6 Project aims

Lab reactor experiments are easier to control and involve less complicating factors than samples form natural ecosystems. This makes it easier to study the different aspects of microbial dynamics and community assembly. We will use data from three reactor experiments (see Section 2.1) as a basis for our study.

In order to gain insight into the nature and structure of the interaction patters in bacterial communities, we will use two different approaches. The ReBoot method introduced by Faust *et al.*[22] will be used to predict interactions from patterns of co-occurrences and co-exclusions. By this method, we seek to answer the following questions:

1. How are microbial communities structured?
2. Do our results correspond to real ecological interactions or are there confounding factors?
3. What is the most sound way of inferring ecological interactions?

As our data in part are time series of developing microbial communities, we also attempt using a dynamic deterministic model suited to *predict* the microbial community. For this, we choose the generalized Lotka-Volterra (gLV) model. In this context, we want to answer the following questions:

4. Do microbial communities follow a specific path based on external selection pressure or are dynamics dominated by stochastic effects?
5. Can dynamics in microbial communities be described, explained, reproduced and predicted?

Chapter 2

Background and theory

2.1 Reactor experiments

All experiments used in this thesis were conducted by the Analysis and Control of Microbial Systems Group (ACMS) at NTNU and are listed in Table 2.1 and described in further detail below. The general outline for the experiments is as follows:

Seawater with bacteria was used as inoculum for batch or chemostat reactors. Samples were taken at specific days as described for the three experiments below. For each sample, DNA was extracted using a standard kit and kept frozen until the end of the experiment. The V3 and V4 regions of the 16S-RNA gene were amplified by PCR with broad coverage primers. Thereafter, the PCR products were labelled with unique (for each sample) barcodes by a second PCR reaction, the samples were pooled and sent to sequencing on one Illumina MiSeq lane at the Norwegian Sequencing center. The raw reads were processed by the USEARCH pipeline[23], where reads with similarity higher than than 97% were clustered into OTUs. Assignment of taxonomy to the OTUs was made comparing the representative sequence to the RDP reference dataset (version 16)[24]. The output of such a workflow is an OTU table, showing the number of reads in each of the samples. For more details on the experimental procedure or data processing, please consult Gundersen[12] or Solberg[25].

We will briefly describe each of the experiments:

Seawater Batch reactor experiment over 24 days with seawater from surface and 90 meter depth, samples being taken on day 1, 2, 17 and 24. At the start of the experiment, the reactors was subjected to r-selection after a nutrient pulse, while the reactors were assumed to be K-selected the two last sampling days. Even though the full results of the experiment has not yet been published, data from

Table 2.1: Experiments used in this thesis

Name used in thesis	Number of samples	Number of reactors	Total number of OTUs	Number of OTUs when filtering at 10^{-4} mean abundance
Seawater	16	4	1353	390
Biofilm	96	9	608	122
Selection-switch	203	12	1302[1]	165

this experiment are used in figure 5 in Vadstein *el al.*[9].

Biofilm Chemostat experiment over 12 weeks carried out as a part of the master thesis of Erlend Hafsten[25]. The reactors had different numbers of biofilm carriers (TR1: 136, TR2: 70, TR3: 0). Samples were taken from the water (W) and the biofilm (C) at week 1,2, 4, 6, 8, 10 and 12. The target of the experiment was to identify the interactions between biofilm and planctonic communities.

Selection-switch Master thesis of Madeleine Stenshorne Gundersen, in collaboration with Ian Morelan[12]. Chemostat experiment carried out over 50 days, were samples from the water community were taken at day 1, 2, 4, 6, 8, 12, 16, 20, 24, 28, 29, 30, 32, 36, 40, 44 and 50. The reactors were divided into high (H) and low (L) carrying capacity. Before day 29, one group of reactors was r-selected (nutrient pulse) and the other group was K-selected (continious nutrient supply). At the 29, the selection regimes were switched such that r-selected reactors became K-selected and vice versa. The main goal of the experiment was to find out whether selection could eliminate potential opportunistic r-strategists.

2.2 Nature of the data

The number of raw read counts from the sequencing platform does *not* correspond to the cell count in the sample due to the PCR amplification. Indeed, the experimental procedures have been designed to provide the same total number of reads for all samples (normalization). Hence, the counts are only valid as indications of the *relative* abundances within the same sample. For this reason, the data are normalized to relative abundances before analysis. Further complications discussed below might cause the estimate of the relative abundances to be inaccurate. We will now list up known difficulties with 16S rDNA datasets:

[1] OTUs having a positive abundance in the relevant samples

- Using relative abundances alone leads to unrecoverable loss of data and erroneous results, as discussed by Faust *et al.*[22] and Gloor *et al.*[26]. For instance, in a community of only two OTUs, the two will be perfectly anti-correlated, making it impossible to draw inferences about their interactions. The best way of getting to know the *absolute* number of each OTU in each sample is scaling by the total cell count determined by other methods such as flow cytometry or q-PCR.

- The broad coverage primers are made to amplify the rDNA of all known bacteria. Ideally, this should amplify the rDNA of each bacterium equally. However, there will be bacteria where the universal primers do not match that well, yielding a lower (or in the worst case no) PCR output than expected. Note that this error should be consistent over all samples and experiments[27, 28].

- Bacteria have different numbers of Ribosomal RNA (rrn) operons containing the S16-RNA gene. Typically, slow-growing K-strategists have one such operon, whereas r-strategists may have over ten rrn operons enabling them to grow fast. Due to this fact, some bacteria are over-represented in the results [29]. As for the PCR selectivity, this is an error being consistent over samples.

- Datasets of microbial communities are *sparse*. Only a few OTUs dominate the data and are found in most samples. The majority of the OTUs are rare and detected in a few samples only. As a result, the OTU tables are filled with lots of zeros, which is an obstacle for further analysis. Even though an OTU is registered as having zero abundance in a sample, it may not be completely missing, but simply have an abundance below the detection limit[17].

- An OTU table typically contains hundreds or thousands of OTUs, often far more than the number of samples. The so-called "curse of dimensionality" reflect that data with so many dimensions are difficult to analyse by conventional means, especially if few data points are provided.

- The results are prone to sequencing errors and incomplete taxonomy databases. For instance, chimera occur when a read is composed on DNA from two different OTUs. The USEARCH pipeline have algorithms to remove those[30], but some errors most certainly remain. Also, many OTUs, especially the rare ones, cannot be determined on a fine taxonomical level. This can be due to incomplete reference taxonomical databases or be artefacts of sequencing errors.

- 16S rDNA microbial datasets contain many OTUs being rare. They may result from sequencing errors, chimera, DNA contamination or simply be real OTUs being present in the samples in low abundance. This may contribute

to a large number of OTUs, making it harder to catch the biological signal and slows down computations. Therefore rare OTUs should be filtered out prior to analysis[16, 31].

2.3 Ordinations

Data of microbial communities are multidimensional by nature. This poses a great challenge: How can be perceive and interpret this many dimensions when living in a world of only three space dimensions? The idea often being used in this setting is to represent the data in a fewer number of dimensions while retaining as much as possible of the underlying information. We will now introduce two such methods: Principal Component Analysis (PCA) and Principal Coordinates Analysis (PCoA)

Definition

Given a matrix $\underbrace{A}_{n \times m}$, the loadings for the first principal component $\mathbf{w}_1$ is the one maximizing the variability $\|A\mathbf{w}_1\|_2^2 = (A\mathbf{w}_1)^\intercal (A\mathbf{w}_1)$ under the constraint $\|\mathbf{w}_1\|_2^2 = \mathbf{w}_1^\intercal \mathbf{w}_1 = 1$. The entity $\mathbf{z}_1 = A\mathbf{w}_1$ is then called the score vector for the first principal component and serves as a first approximation of the variability between the columns. Imagine that A contains the abundances of the OTUs with samples in rows and OTUs in columns. Then the first principal component is an approximation of the community structure represented by A in one dimension.

Further principal components can be extracted in sequence by finding the loadings explaining the most of the variability, while being orthogonal to the past principal components, the details are beyond the scope for this thesis[32]. In practice, we plot the two first principal components. This corresponds to projecting the n-dimensional space into the two-dimensional plane explaining as much of the variability of the data as possible.

PCoA (Principal Coordiantes Analysis) use the same underlying idea as PCA, but in this case, the PCA ordination is made from a matrix of similarities or dissimilarities between the data points.

Proportion of variance explained

When investigating the structure of a dataset with PCA or PCoA, we often plot the results in two dimensions. Hence, we want the first two principal components to explain as much of the variability as possible. The proportion of variance explained by principal component i with scores $\mathbf{z}_i$ is[32]:

$$\frac{\|\mathbf{z}_i\|_2^2}{\mathrm{Tr}\,(A^\intercal A)} \tag{2.1}$$

By definition, the proportion of variance explained decreases monotonically with i. The proportion of variance explained should be printed in the axes of a PCA plot as this reflect of much of the overall structure in the data which is explained by the plot. If the plot only explains a small fraction of the variability, the usefulness of the plot should be taken into question.

2.4 Similarity measures

Definition

There does not exist any single definition on what a similarity measure is. The given definition used in this thesis, is the our own operational definition:

Given two real vectors $\mathbf{x}$ and $\mathbf{y}$ of the same length n, an unsigned similarity measure is defined as a function $f : \mathbb{R}^n \times \mathbb{R}^n \to [0, 1]$, satisfying:

$$f(\mathbf{x}, \mathbf{x}) = 1 \quad \text{for all} \quad \mathbf{x} \neq \mathbf{0} \quad \text{(Any vector is totally similar to itself)} \tag{2.2}$$

$$f(\mathbf{x}, \mathbf{y}) = f(\mathbf{y}, \mathbf{x}) \quad \text{for all} \quad \mathbf{x}, \mathbf{y} \quad \text{(Similarity is a symmetric relation)} \tag{2.3}$$

$$f\left(\sigma(\mathbf{x}), \sigma(\mathbf{y})\right) = f(\mathbf{x}, \mathbf{y}) \quad \text{for all} \quad \mathbf{x}, \mathbf{y}$$
$$\text{and any permutation} \quad \sigma \quad \text{(The ordering of data points does not matter)} \tag{2.4}$$

Here, 0 is interpreted as total dissimilarity and 1 is interpreted as total similarity.

Similarly, we can define a signed similarity measure as $f : \mathbb{R}^n \times \mathbb{R}^n \to [-1, 1]$, which satisfies the conditions above. Here, the interpretation of 1 is total similarity, 0 is non-similarity and -1 is negative similarity.

We can extend this definition to matrices, this is, given a matrix

$$\underbrace{X}_{n \times m} = \left[\mathbf{x_1} \quad \mathbf{x_2} \quad \dots \quad \mathbf{x_m}\right], \tag{2.5}$$

we define $f(X) = \underbrace{S}_{m \times m}$ as follows:

$$(S)_{ij} = f\left(\mathbf{x_i}, \mathbf{x_j}\right) \tag{2.6}$$

Equations (2.2) to (2.4) thus imply in respective order:

$$\operatorname{diag}(S) = \mathbf{1} \quad \text{if all columns vectors in } X \text{ are non-zero} \tag{2.7}$$

$$S^{\mathsf{T}} = S \tag{2.8}$$

$$f\left(\sigma_R(X)\right) = f(X) = S \quad \text{for any row permutation} \quad \sigma_R \tag{2.9}$$

Comments:

As we will see, the concept of similarity measures in this thesis is a rather broad one. For instance, this includes what is perceived in literature as correlations (for

instance Pearson and Spearman), and also what is considered normalized distances (or dissimilarities), such as Bray-Curtis and Jaccard. These two are often described with different terminology and might have other conventions, such as whether the vectors are taken to be row or columns in the data matrix. For this project however, we will neglect these differences and rather group the measures according to the criteria listed below. Also, note that the triangle inequality: $f(\mathbf{x}, \mathbf{y}) + f(\mathbf{y}, \mathbf{z}) \geq f(\mathbf{x}, \mathbf{z})$ is *not* a part of our criteria. Some of our similarity measures will obey this inequality, whereas other will not.

Subtypes of similarity measures

We further provide our own operational classification of the different kinds of similarity measures:

- Presence-absence: Given two vectors $\mathbf{x}$ and $\mathbf{y}$, a presence-absence metric will only consider which elements in $\mathbf{x}$ and $\mathbf{y}$ are non-zero. For the non-zero-elements, the magnitude of the numbers does not matter. In our notation, it means:

$$f\left(\text{bin}(\mathbf{x}), \text{bin}(\mathbf{y})\right) = f(\mathbf{x}, \mathbf{y}) \tag{2.10}$$

- Non-parametric: Such similarity measures are based on the ranking of the elements in the individual vectors, but not their magnitude. Stated by formulas, we obtain:

$$f\left(R_{\mathbf{x}}, R_{\mathbf{y}}\right) = f(\mathbf{x}, \mathbf{y}) \tag{2.11}$$

  Often, care should be taken when considering tied ranks, in which case corrections often are made.

- Parametric: Given $\mathbf{x}, \mathbf{y}$, the metric is a continuous function of the two vectors and hence, the magnitude of the elements matter. We thus get the defining equation:

$$\lim_{\substack{\mathbf{x}\to\mathbf{x}_0 \\ \mathbf{y}\to\mathbf{y}_0}} f(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}_0, \mathbf{y}_0) \tag{2.12}$$

Robustness to noise and other considerations

We note that presense-absence metrics are very sensible to noise as the slightest perturbation can change the result dramatically. This is indeed troublesome when there is noise in the measurements or limited resolution, this is: Might the microbe be missing from the sample or is it present and its abundance below the detection threshold? For situations where the species inventory is well-characterized, presence-absence can useful, but for analysing microbial communities, we will not pay them much attention.

Non-parametric measures can be said to be more robust to noise as small perturbations often do not change the ranks and hence the statistic. However, in cases with many tied or close observations, especially relevant is the case with many zeros, small perturbations might have a huge impact. Parametric measures might be easier to deal with in this respect as they change continuously with the added noise. However, the patterns they capture might often be narrower. For instance, the Pearson correlations is parametric and measures linear relationship. However, often we are more interested in general monotonous relationships, in which cases the Spearman correlation is more suited[33].

Similarity measures used in this thesis

ff For this thesis, a selection of similarity measures were selected and used in the analysis. These are listed in Table 2.2.

2.4.1 Adding of noise

In order to assess the stability of similarity measures with uncertain data, noise can be added. OTU tables have a limited resolution (determined by the total number of reads). Hence a zero in the dataset does not mean that the abundance of the OTU necessary is zero. More often this is due to the fact that the OTU is present below the detection level[1]. Thus, the rationale of adding noise has been to turn down the significance of interactions being due to the sparsity and discreteness of the data. Significant correlations found are less to be trusted if a little noise make them disappear.

2.5 The ReBoot pipeline

The ReBoot (Permutation-Renormalization and Bootstrap) pipeline was originally presented by Faust *et al.*[22] and is illustrated in Figure 2.1. The inputs to the ReBoot algorithm is an OTU table and a similarity measure. Compared to taking the raw correlations of the relative abundances, the ReBoot approach adds some refinements:

- The similarity scores are found through bootstraping, making the method more robust to spurious correlations in addition to estimating the standard deviation of the estimate[45]
- It has a way of assigning the statistical significance to each interaction pairs. This is done through comparison of the bootstrap distribution with the so-called *null distribution*

[2]With negative numbers, this statistic can be negative, but we will not consider negative abundances in this work

Table 2.2: Similarity measures used in this thesis

Name	Type	Comments	References
Pearson correlation	Parametric, signed	Measures linear correlation	[34]
Spearman correlation	Non-parametric, signed	The Pearson correlation between ranks	[33, 35]
Kendall's tau	Non-parametric, signed	Based on the number of inversions in ranks	[36]
Bray-Curtis	Parametric, unsigned	Commonly used in ecology	[37, 38]
Jaccard index	Presence-absence, unsigned	Calculated as intersection over union	[37, 38]
Generalized Jaccard index	Parametric, unsigned	Also called Ruzicka similarity	[39]
nc.score	Non-parametric, signed	Based on patterns of co-exclusion and co-existence	[40]
Squared Euclidean similarity	Parametric, unsigned	Normalized Euclidean distance	[41]
Mutual information	Parametric, unsigned	Based on information theory, requires discretizing prior to calculation	[42, 43]
Cosine similarity	Parametric, unsigned[2]	Geometrically interpreted as the cosine of the angle between the two vectors	[37, 44]

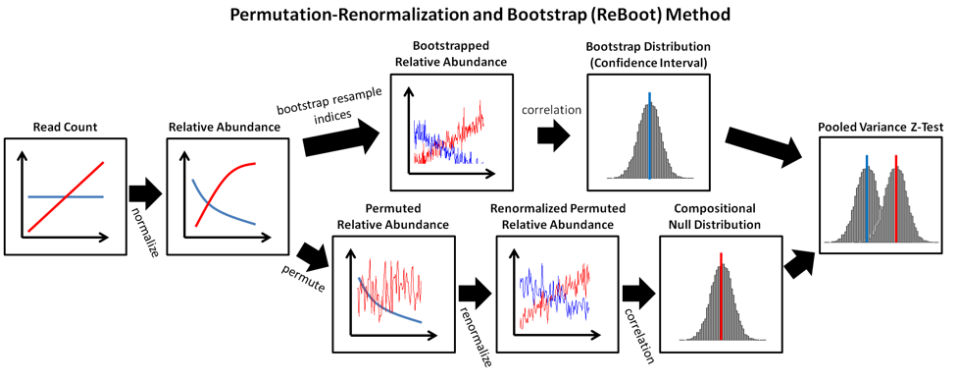


Figure 2.1: Overview of the ReBoot approach[22]

2.5.1 Calculating correlations and null distribution by bootstrapping and permuation

The samples from the dataset are drawn with replacement to form another dataset having the same number of samples, but possibly with some of the original samples repeated. From this bootstrap selection, the similarity measure computes all pairwise correlations and stores them in a matrix. This is repeated B times[3], forming a bootstrap distribution for each pairs of OTUs. The null distribution for each pair of OTUs is created by permuting the abundances of one of the OTUs, finding the correlation score between the OTUs and repeat the procedure N times[4]. In order to compensate for the fact that permutation breaks the compositionality of the data (samples do no longer sum to one), which may lead to spurious correlations, the data are renormalized after permutation.

2.5.2 Comparing the two distributions

The null distribution can be interpreted as the correlation values to expect if there was no association between the OTUs. This is compared to the bootstrap distribution to see whether the association is by chance or not. Let $\bar{X}_{ij}$ and $\bar{Y}_{ij}$ be the means of the bootstrap and null distribution, respectively of the correlation between OTU i and OTU j. Furthermore, let $\widehat{\text{Var}}\left[X_{ij}\right]$ and $\widehat{\text{Var}}\left[y_{ij}\right]$ be their respective sample variances. The observed z-statistic is then computed as:

$$z_{ij} = \frac{\bar{X}_{ij} - \bar{Y}_{ij}}{\sqrt{\frac{1}{2}\left(\widehat{\text{Var}}\left[X_{ij}\right] + \widehat{\text{Var}}\left[Y_{ij}\right]\right)}}, \tag{2.13}$$

[3] B is taken to be a number high enough to provide stability
[4] N is in practise taken to be equal to B

which then is assumed to have a standard normal distribution[5]. According to the normal distribution, the p-value is computed as

$$p_{ij} = 2 \cdot \Phi\left(-|z_{ij}|\right), \tag{2.14}$$

where Φ is the cumulative probability function of the standard normal distribution. Note that this p-value only applies this pair of OTUs and does not take into account all the other possible associations between OTUs, in which case we encounter the problem of multiple testing. To accommodate this, each computed p-value is corrected by the Benjamini-Hochberg-Yekutieli procedure[46], producing q-values.

2.6 Network theory

A collection of *nodes* or *vertices*, connected by *links* or *edges* is called a *network*. The concept of network is generic, so the nodes can be whatever entity we are studying such as individual persons, computers, genes or microbial OTUs. Similarly, the links can refer to any possible connection between the nodes, this may be conversations between people, gene regulations or interaction between bacteria[47].

A subset of nodes in a graph having stronger interconnections among themselves than the rest of the network, is called a *community*. A community may thus correspond to a subgroup of nodes having something in common or are interacting in a special way. There exist numerous formal definitions of what a community is, as well as methods for detecting the communities. Hence, the result of a community finding procedure depends on criteria and algorithms being used[47].

The walktrap community detection algorithm presented by Pons *et al.*[48] uses random walks to provide a basis for aggregative community detection. The underlying idea is that short random walks starting from one node has a higher likelihood of staying inside its community. From random walks of a fixed length t, the transition probabilities P_{ij} (probability of ending up at node j given start at node i) are estimated. These numbers are further used to calculate distances between nodes. From these distances, Ward clustering is used, resulting in a hierarchical tree. Each cut of this tree corresponds to a partitioning of the overall network. The partition which maximizes the so-called modularity score is finally chosen as the partition capturing the modularity the best.

[5]The presented formula is taken from the source code of `ccrepe`. We would consider the formula

$$z_{ij} = \frac{\bar{X}_{ij} - \bar{Y}_{ij}}{\sqrt{\frac{1}{B}\widehat{\mathrm{Var}}\left[X_{ij}\right] + \frac{1}{N}\widehat{\mathrm{Var}}\left[Y_{ij}\right]}}$$

to the most statistically sound.

2.7 The Lotka-Volterra modelling approach

2.7.1 Introduction

This approach is based on the generalized Lotka-Volterra[49] equation, being widely used in ecology. Faust `et al.`[1] suggested using a gLV model for studying microbial community dynamics, whereupon Buffie *et al.*[20] used it to predict and identify which bacteria inhibit the human pathogen *Clostridium difficile*. The equation models the community as a system of differential equation. Imagine we have a system of n OTUs, labelled $x_1, x_2, \ldots, x_n$. The system is then being modelled as:

$$\frac{\mathrm{d}\, x_i}{\mathrm{d}\, t} = x_i \left(a_{i0} + \sum_{j=1}^{n} a_{ij} x_j \right) \text{ for } i = 1, \ldots, n. \tag{2.15}$$

For $n = 1$, the equation reduces to the logistic equation. The coefficient a_{i0} can be interpreted as the maximum growth rate of the OTU when it is alone, whereas the coefficients a_{ij} are interpreted as the interaction between the OTUs. a_{ii} is called the *self interaction* coefficient. Note that $a_{ij} \neq a_{ji}$ in general (also if we exclude the trivial case when $i = 0$ or $j = 0$), so unlike the ReBoot approach, we can model asymmetrical effects between OTUs.

2.7.2 Creating linear systems

We will turn on how to find the coefficients from experimental data. Let us label the time points of sampling $t_0, t_1, \ldots, t_N$ for the rest of this section. For a system with n OTUs, there are in total $n\,(n+1)$ coefficients to be fitted. We will present a method (with two variations) to reduce the inference problem to systems of linear equations. The main inspiration for the procedure comes from Kloppers and Greeff[50] and Stein *et al.*[19], where the latter adopts the procedure to microbial data sets. The general idea is as follows:

For each OTU i, we express the differences in abundance of OTU i in two consecutive timepoints in terms of the abundances of all other OTUs. Performed over all consecutive pairs of timepoints, we obtain an equation system where the coefficients affecting the presence of OTU i are the solutions. This procedure is repeated for all OTUs to predict.

Integral method

The *integral method* is based on rewriting Equation (2.15) as:

$$x_i\,(t_{k+1}) - x_i\,(t_k) = \int_{t_k}^{t_{k+1}} \frac{\mathrm{d}\, x_i(t)}{\mathrm{d}\, t}\, \mathrm{d}t = \int_{t_k}^{t_{k+1}} x_i(t) \left(a_{i0} + \sum_{j=1}^{n} a_{ij} x_j(t) \right) \mathrm{d}t. \tag{2.16}$$

$x_i(t_{k+1})$ and $x_i(t_k)$ are available from the data. The right side of the system is approximated by the trapezoidal rule, hence:

$$x_i(t_{k+1}) - x_i(t_k) = \int_{t_k}^{t_{k+1}} x_i(t)\left(a_{i0} + \sum_{j=1}^{n} a_{ij}x_j(t)\right)\mathrm{d}t \approx$$
$$\frac{t_{k+1}-t_k}{2}\left(x_i(t_{k+1})\left(a_{i0} + \sum_{j=1}^{n} a_{ij}x_j(t_{k+1})\right) + x_i(t_k)\left(a_{i0} + \sum_{j=1}^{n} a_{ij}x_j(t_k)\right)\right) =$$
$$\frac{(t_{k+1}-t_k)\left(x_i(t_{k+1}) + x_i(t_k)\right)}{2}\cdot a_{i0} + \sum_{j=1}^{n}\frac{(t_{k+1}-t_k)\left(x_i(t_{k+1})\,x_j(t_{k+1}) + x_i(t_k)\,x_j(t_k)\right)}{2}\cdot a_{ij} \tag{2.17}$$

Note that this applies for $k = 0, 1, \ldots, N-1$. Putting this together all for values of k yields the linear system:

$$\underbrace{\begin{bmatrix} d_{i,0} \\ d_{i,1} \\ \vdots \\ d_{i,N-1} \end{bmatrix}}_{\mathbf{d}_i} = \underbrace{\begin{bmatrix} \bar{x}_{i,0} & \bar{x}_{i,0}\bar{x}_{1,0} & \ldots & \bar{x}_{i,0}\bar{x}_{n,0} \\ \bar{x}_{i,1} & \bar{x}_{i,1}\bar{x}_{1,1} & \ldots & \bar{x}_{i,1}\bar{x}_{n,1} \\ \vdots & & & \\ \bar{x}_{i,N-1} & \bar{x}_{i,N-1}\bar{x}_{1,N-1} & \ldots & \bar{x}_{i,N-1}\bar{x}_{n,N-1} \end{bmatrix}}_{X_i} \cdot \underbrace{\begin{bmatrix} a_{i,0} \\ a_{i,1} \\ \vdots \\ a_{i,n} \end{bmatrix}}_{\mathbf{a}_i}, \tag{2.18}$$

where $d_{i,k} = x_i(t_{k+1}) - x_i(t_k)$, $\bar{x}_{i,k} = \frac{(t_{k+1}-t_k)\left(x_i(t_{k+1})+x_i(t_k)\right)}{2}$ and $\bar{x}_{i,k}\bar{x}_{1,k} = \frac{(t_{k+1}-t_k)\left(x_i(t_{k+1})x_j(t_{k+1})+x_i(t_k)x_j(t_k)\right)}{2}$.

Log-integral method

The *log-integral* method is a variation of the integral method, where Equation (2.15) is divided by x_i prior to integration. According to Kloppers and Greeff[50], this is an improvement over the integral method:

$$\ln\left(x_i(t_{k+1})\right) - \ln\left(x_i(t_k)\right) = \int_{t_k}^{t_{k+1}} \frac{1}{x_i(t)}\cdot\frac{\mathrm{d}\,x_i(t)}{\mathrm{d}\,t}\,\mathrm{d}t = \int_{t_k}^{t_{k+1}}\left(a_{i0} + \sum_{j=1}^{n} a_{ij}x_j(t)\right)\mathrm{d}t \approx$$
$$(t_{k+1}-t_k)\,a_{i0} + \sum_{j=1}^{n}\frac{(t_{k+1}+t_k)\left(x_j(t_{k+1}) + x_j(t_k)\right)}{2}a_{ij}, \tag{2.19}$$

which again can be converted into a linear system:

$$\underbrace{\begin{bmatrix} l_{i,0} \\ l_{i,1} \\ \vdots \\ l_{i,N-1} \end{bmatrix}}_{\mathbf{l}_i} = \underbrace{\begin{bmatrix} t_1 - t_0 & \bar{x}_{1,0} & \dots & \bar{x}_{n,0} \\ t_2 - t_1 & \bar{x}_{1,1} & \dots & \bar{x}_{n,1} \\ \vdots & & & \\ t_N - t_{N-1} & \bar{x}_{1,N-1} & \dots & \bar{x}_{n,N-1} \end{bmatrix}}_{X} \cdot \underbrace{\begin{bmatrix} a_{i,0} \\ a_{i,1} \\ \vdots \\ a_{i,n} \end{bmatrix}}_{\mathbf{a}_i}, \tag{2.20}$$

where $l_{i,k} = \ln\left(x_i\left(t_{k+1}\right)\right) - \ln\left(x_i\left(t_k\right)\right)$ and $\bar{x}_{k,j} = \frac{\left(t_{k+1}+t_k\right)\left(x_j\left(t_{k+1}\right)+x_j\left(t_k\right)\right)}{2}$. Notice that the matrix X in Equation (2.20) does not depend on i.

2.7.3 Solving the linear systems

If the linear system inferred from the integral method or the log-integral method has a number of unknowns being equal or less than the number of equations as assumed by Kloppers and Greeff[50], the system

$$X_i \mathbf{a}_i = \mathbf{d}_i \tag{2.21}$$

can be solved by the methods of least squares, this is

$$\left(X_i^\intercal X_i\right) \mathbf{a}_i = X_i^\intercal \mathbf{d}_i, \tag{2.22}$$

yielding the solution

$$\mathbf{a}_i = \left(X_i^\intercal X_i\right)^{-1} X_i^\intercal \mathbf{d}_i. \tag{2.23}$$

For microbial datasets, we often get the problem that the number of OTUs (unknowns) is *greater* than the number of equations. In this case, we need another approach. This is provided by Stein *et al.*[19] and is called Tikhonov or *ridge regularization*[51]. For each i, the optimal solution $\mathbf{a}_i^*$ is the one minimizing the cost function:

$$C\left(\mathbf{a}_i;\, \lambda_{\text{self}},\, \lambda_{\text{interaction}}\right) = \left\|X_i \mathbf{a}_i - \mathbf{d}_i\right\|_2^2 + \lambda_{\text{self}} \cdot a_{i,0}^2 + \lambda_{\text{interaction}} \sum_{j=1}^{n} a_{i,j}^2. \tag{2.24}$$

The rationale behind this approach this to get a reasonable good fit to the data (first term), while at the same time try to lower the magnitude of the growth rates and interactions (second and third term respectively). The optimal solution has the closed form:

$$\mathbf{a}_i^* = \arg\min_{\mathbf{a}_i} C\left(\mathbf{a}_i;\, \lambda_{\text{self}},\, \lambda_{\text{interaction}}\right) = \left(X_i X_i^\intercal + D_\lambda\right)^{-1} X_i^\intercal \mathbf{d}_i, \tag{2.25}$$

where $D_\lambda = \text{diag}\left(\lambda_{\text{self}}, \underbrace{\lambda_{\text{interaction}}, \lambda_{\text{interaction}}, \ldots, \lambda_{\text{interaction}}}_{n \text{ times}}\right)$.

The tuning parameters λ_{self} and $\lambda_{\text{interaction}}$ need to be estimated, which is done through cross-validation[19].

2.7.4 Prediction of system

After tuning parameters have been found and used for inferring the coefficients, the estimated coefficients can be inserted into Equation (2.15). Unfortunately, no analytical solution exist unless $n = 1$. However, given the initial abundances, the equation can be solved numerically as an initial value problem. The final result is then a prediction of the dynamics of the microbial community given the initial abundances.

2.7.5 Special considerations

As presented earlier, microbial datasets are often sparse with a lot of zero abundances. For the integral method, we see from Equation (2.17) that a zero still makes the equation valid. However, if both $x_i(t_k)$ and $x_i(t_{k+1})$ are zero, the equation corresponding to these timesteps, reduces to the trivial equation where all coefficients are zero, as well as the right side. This captures the intuitive idea that we do not gain any information about an OTU having zero abundance.

For the log-integral method, the situation is slightly worse. The logarithm of zero is undefined, so equations where $x_i(t_k)$ *or* $x_i(t_{k+1})$ is zero must be thrown away or the zero abundance be replaced with a small positive number.

According to Stein *et al.*[19], the approach assumes the data to be absolute abundances. Hence, the relative abundances in the OTU table should be scaled by measurements of total cell count through methods such as qPCR or flow cytometry.

2.8 Comparison of the two main approaches

Whereas both the ReBoot procedure and Lotka-Volterra approach both aim to uncover the interactions of the OTUs, they do so in different ways. ReBoot is a so-called cross-sectional approach where samples are pooled together, meaning that they are all treated equally with respect to each other. The output of the procedure is a list of correlations found to be significant, with their similarity score and their q-value. Note that these similarity scores are symmetric, meaning that the magnitude and sign of an interaction is equally strong for both OTUs taking part in it. However, we know that many ecological interactions are asymmetric[52], but the ReBoot procedure has no way of accounting for this fact. Moreover, the results from ReBoot must largely be taken on faith, as we usually have no good way

of knowing whether the associations correspond to real ecological interactions. The Lotka-Volterra approach on other hand, is a time series based method, meaning that we have to feed the pipeline with some kind of time series in order to make it work. On the other hand, cross-sectional methods such as ReBoot can deal with any kind of microbial data. However, the Lotka-Volterra approach has some benefits that the ReBoot has not. It deals with indirect effects more easily, allow asymmetric interactions and since it produces coefficients in a differential equation model, its results can easily be assessed by predicting time series and comparing it to the reference.

Chapter 3

Materials and methods

The starting point of the analysis was the OTU tables. For simplifying and structuring the workflow, we created a custom `R`-packed named `micInt`. A graphical summary of the overall workflow is shown in Figure 3.1

3.1 Preprocessing

3.1.1 Filtering and normalization

Prior to analysis, all data was filtered according to the overall mean abundance. This is, all OTUs with mean abundance (over all samples) less than 10^{-4} were removed. The choice of the number 10^{-4} was a result of trial and error experimenting with different cutoffs. For the selection-switch experiment, absolute abundances were estimated from the relative ones by multiplying with the bacterial density, determined by flow cytometry. The relative abundances from each of the three experiments were then renormalized after filtering by scaling the OTU abundances per sample such that they summed to one, retaining the compositionality.

3.1.2 Subdivision prior to ReBoot pipeline and generation of PCoA plots

The seawater experiment contains only 16 samples, so it was decided not to split it up prior to the analysis. For the biofilm experiment, the data were grouped in two different ways:

- By source of sample: Water (W) or biofilm carrier (C)
- By treatment: Treatment 1 (TR1) with 136 carriers per reactor, treatment 2 (TR2) with 70 carriers per reactor or treatment 3 (TR3) wih 0 biofilm carriers per reactor

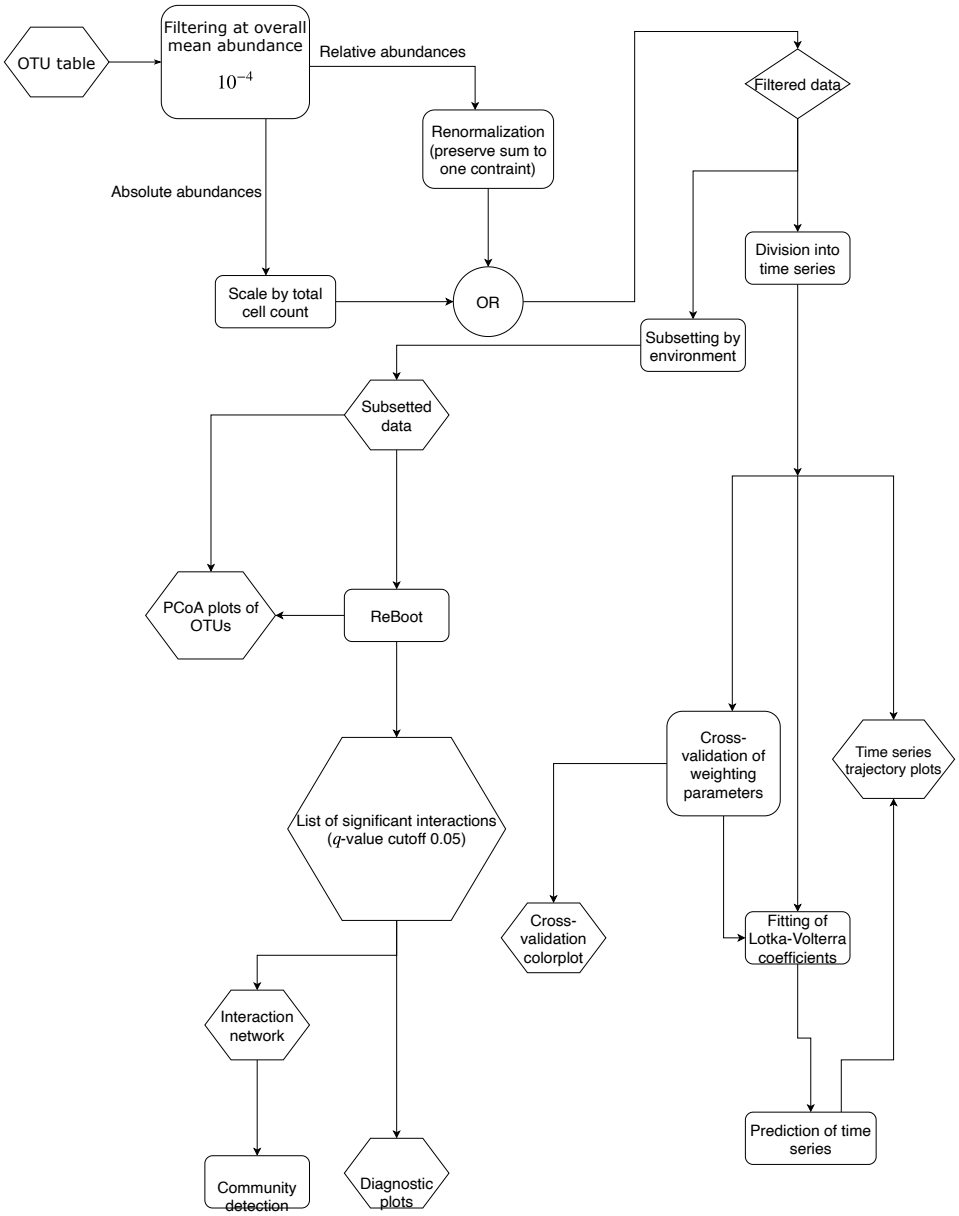


Figure 3.1: Overview of the workflow in the thesis

As a result, the groupings were overlapping, meaning that all treatments have both water and carrier samples, except TR3 where carrier samples were absent. In theory, the samples could be further subdivided such that information about the source *and* treatment was captured. However, as we need many samples to get reasonable results, this was not done.

For the selection-switch experiment, the data were subdivided into each combination of the present selection selection regime (r or K) and carrying capacity (high (H) and low (L)). The samples from day 1, 2, 29, 30 and 32 were filtered away as these data represent transient dynamics (start of experiment and change of selection regime), an observation being pointed out by Gundersen[12]. Also, failure of filtering away the transient days, produced PCoA plots were the patterns were less observable.

3.1.3 Generation of time series for time trajetory plots and Lotka-Volterra pipeline

For the seawater experiment, the number of samples per reactor was way too small and the samples too far apart in time to employ the Lotka-Volterra pipeline. However, the other experiments did provide time series long and dense enough such to try out the approach. Please note though that the biofilm experiment does not provide absolute abundances, so the relative abundances were used instead.

For making the time trajectory plots, all time points were included and the samples from the selection-switch experiment were grouped according to the initial selection regime[1] and the nutrient concentration. The biofilm was subsetted according to source of sample *and* by treatment, opposed to the subdivision conducted in the ReBoot pipeline.

For further work on the Lotka-Volterra pipeline, the subdivision was done as for the ReBoot pipeline. However, the biofilm experiment was still subdivided by source of sample and treatment. After subdivision, the data from each reactor served as a time series.

3.2 The pipeline generating and modifying similarity measures

All of the basic similarity measures presented in Table 2.2 are implemented by `micInt:: similarity_measures`. Furthermore, some of the similarity measures are modified by adding noise or by mean scaling. In those cases, we still used the base similarity measure in addition to the modified one. See Table 3.1 for which modifications were applied to which similarity measures.

[1] The selection regime for each reactors changes halfway in the experiment

Table 3.1: Implementation and modification of similarity measures applied in thesis

Similarity measure	Implementation	Noise added?	Mean scaled?	Reason for not applying all modifications
Pearson correlation	`stats::cor(method='pearson')`	✓	✗	Mean-scaling factor disappears in calculation
Spearman correlation	`stats::cor(method='spearman')`	✓	✗	Mean scaling makes no difference to non-parametric measures
Kendall's tau	`stats::cor(method='kendall')`	✓	✗	Mean scaling makes no difference to non-parametric measures
Bray-Curtis	`vegan::vegdist(method='bray')`	✓	✓	
Generalized Jaccard index	`vegan::designdist(method='1-J/(A+B-J)', terms='minimum')`	✓	✓	
nc.score	`ccrepe::nc.score`	✓	✗	Mean scaling makes no difference to non-parametric measures
Squared Euclidean similarity	`vegan::designdist(method='((A+B-2*J)/P)/(1+(A+B-2*J)/P)', terms='quadratic')`	✓	✗	Mean-scaling factor disappears in calculation
Jaccard index	`vegan::vegdist(method='jaccard', binary=T)`	✗	✗	For presence-absence measures, adding noise does not make sense, neither does mean scaling make any difference
Mutual information	`infotheo::mutinformation`	✓	✗	Mean scaling makes no difference to non-parametric measures
Cosine similarity	`vegan::designdist(method='J/sqrt(A*B)', terms='quadratic')`	✓	✗	Mean-scaling factor disappears in calculation

3.2.1 Addition of noise

We defined a level of noise as γ. Then, adhering to the notation in Section 2.4, given a similarity measure f, we make a stochastic noisified similarity measure f^* by defining:

$$f^*(\mathbf{x}, \mathbf{y}) = f\left(\mathbf{x} + \boldsymbol{\epsilon}_{\boldsymbol{x}}, \mathbf{y} + \boldsymbol{\epsilon}_{\boldsymbol{y}}\right), \tag{3.1}$$

where $\boldsymbol{\epsilon}_{\boldsymbol{x}}$ and $\boldsymbol{\epsilon}_{\boldsymbol{y}}$ are independent and identically distributed stochastic vectors. We use two different forms of noise:

- Normally distributed noise adding componentwise a random variable $\epsilon \sim N\left(0, \gamma^2\right)$ to the vector of abundances.
- Uniform noise adding componentwise a random variable $\epsilon \sim \text{Uniform}\left(-\gamma, +\gamma\right)$ to the vector of abundances.

In other words: The noise level γ is the standard deviation for the normally distributed noise and the range of the interval for the uniformly distributed noise. The similarity measures with noise added will have the postfixes `normal` and `uniform`, for normally and uniformly distributed noise, respectively.

For the implementation in `micInt::noisify`, there are two additional aspects to the procedure which are worth to note:

- The definition in Equation (3.1) allows for the abundance to be negative after adding noise. As we do not want to deal with negative abundances, we solve the problem by taking the elementwise absolute value after adding noise. Hence, we get a bounce-back effect when an OTU abundance falls below zero.
- For the presence-absence similarity measures, adding noise makes no sense, so adding of noise is skipped in these cases.

3.2.2 Mean scaling

The idea behind mean scaling is the fact that some similarity measures consider OTUs being present at different abundances to be non-similar even though their abundances are perfectly correlated. For example, consider OTU B to live from secondary metabolites from OTU A. The abundance of OTU B is almost exactly proportional to the abundance of OTU A, but the abundance of OTU B is one order of magnitude lower due the low production and energy yield of the secondary metabolite. There is an actual interaction between the two OTUs, but some similarity measures may not reflect this.

In order to resolve this issue, we try to scale the abundance vectors by their respective means, this is: Given a similarity measure f, we create a new one g by defining:

$$g(\mathbf{x}, \mathbf{y}) = f\left(\mathbf{x}/\bar{x}, \mathbf{y}/\bar{y}\right), \tag{3.2}$$

where $\bar{x}$ and $\bar{y}$ denote the means of the two abundance vectors. This procedure is carried out by `micInt::noisify`.

Mean scaling only applies to some similarity measures. For absence-presence and non-parametric measures, this approach makes no difference at all. Likewise, certain parametric similarity measures such as cosine and Pearson have normalization terms canceling the effect of mean scaling. As shown in Table 3.1, only Squared Euclidean, Bray-Curtis and Generalized Jaccard index are applicable for the mean scaling, in which case the mean scaling is indicated by the postfix `mean_scaled` in the name of the similarity measure.

3.2.3 Chaining

Addition of noise and mean scaling can be chained, obtaining measures which are both noisified and mean scaled. As the noise is intended to refer to the original data, addition of noise is done before performing mean scaling. However, the way the measures are wrapped means that the functions to do so are called in the opposite order. In order to get all variations of the similarity measures available in `micInt`, one can type `sim_measures % > % mean_scale % > % noisify`.

3.3 ReBoot pipeline

The base ReBoot functionality is implemented in the `R`-package `ccrepe`[40] and is available as a Cytoscape app[53]. The pipelined approach used in this thesis consists of the following steps:

1. Filter and normalize dataset
2. Resample pairwise correlations by bootstrapping
3. Create null-distributions based on random permutations of the microbes
4. Compare resampling and null distribution
5. Find the correlations being significant above a certain threshold
6. Create tables and networks showing the interaction structure

Item 1 is already captured in Section 3.1.1, whereas Items 2 to 4 were already implemented in `ccrepe::ccrepe` and are explained in Section 2.5. In between these points, there are some implementational aspects to be mentioned:

- The analysis were conducted for all available similarity measures in parallel (see table 3.1 for the available combinations), making the results from one similarity measure independent from the others. By contrast, in Faust *et. al.*[22], the association networks for the different similarity measures were merged before the significance was calculated.

- Large parts of the analysis were conducted by the wrapper function `micInt::runAnalysis` around `ccrepe`
- Renormalization of absolute abundances does not make sense. Therefore, the original `ccrepe` package was modified, adding the option to turn renormalization off. Disabling of renormalization was only used when analyzing the absolute abundances from the selection-switch experiment. Even when the renormalization was turned off, the bootstrapping was enabled.
- In order to dynamically account for the discreteness and uncertainty in the OTU table X, the lowest non-zero entry $x_{\min}$ was extracted. This was further used to determine the noise level γ to be added to the similarity measures:

$$\gamma = s \cdot x_{\min}, \tag{3.3}$$

  where s is called the *magnitude factor* and set to 10 for the all analysis presented. In practice, $x_{\min}$ corresponded to an abundance of one read. Given that each sample had 10000 to 100000 reads, the magnitude factor of $s = 10$ ensured low levels of noise which should not have any major effect on the OTUs having high abundance in the sample.

3.3.1 Creating interaction tables

All possible interactions were filtered by its q-value given a critical value q_{crit}. Hence, the association between OTU i and OTU j was considered significant if $q_{ij} < q_{crit}$. For our purposes, we set $q_{crit} = 0.05$, meaning that the expected number of false discoveries (false positive) to the total number of reported positives is less than 0.05[46]. From the associations found to be significant, a table was created. We commonly refer to these significant associations as *significant interactions*. This does *not* mean that they actually reflect any casual interactions, nor that we strongly believe that this is the case. Rather, it is a convenient naming.

3.3.2 Diagnostic plots

Three types of diagnostic plots were created from the interaction tables, all implemented by `micInt::autoplot.interaction_table`:

- A plot showing the number of significant interactions found for each OTU versus its mean abundance. The rationale behind this plot is that rare OTUs are expected to be listed with a smaller number of interactions than the more common ones. This is due to the fact that the signal for a common OTU is clearer and less influenced by noise, making its interactions more significant. Hence, this plot is used as a diagnostic tool is order to investigate this hypothesis.

- An abundance-product plot. The abundance product is defined as the product of the mean abundances of two OTUs (versions for median and maximum abundances are implemented, but not used). For each significant interaction being found, the abundance product to the two involved OTUs is plotted on the y-axis, whereas the x-axis corresponds to the q-value. Again, the purpose of the plot is to check whether the most interactions are between the abundant OTUs, as we would expect.

- Plot showing the similarity score versus the q-value for each significant interaction. The rationale behind this plot is the fact that the relationship between similarity score and the assigned significance is not necessary monotonous using the ReBoot approach.

3.3.3 Creation of networks

From the tables of interactions, networks were made for each similarity measure for the 200 most significant interactions and plotted using the `R`-package `igraph`[54]. Also, for illustrational purposes, a network containing all significant interactions was plotted. The links were colored according to the sign of the interaction.

In addition, a special network was made, consisting of the 200 most significant interactions using Spearman correlation with normally distributed noise. Only the edges corresponding to positive interactions were initially added to a network. Interaction clusters[2] in this network were detected by the walktrap algorithm (`igraph::cluster_walktrap(steps = 4)`)[48, 54] and labeled in the graph. Later, the edges corresponding to negative interactions were added to the network. Also, a phylogenetic tree being generated from the selection-switch experiment[12] was pruned in order to only retain the OTUs being present in the network. This phylogenetic tree was plotted using the package `ape`[55] and the interactions clusters, in addition to the taxonomy of class level were labeled.

3.4 Creation of PCoA plots

Each available similarity measure (as listed in Table 3.1) was used on the filtered OTU tables, creating a matrix of pairwise similarities between the OTUs[3]. A PCoA ordination was created from this matrix using the build-in `R`-function `stats::prcomp` and the result was plotted. In the same manner, PCoA ordination plots were made from the similarity scores obtained from the ReBoot procedure, in which case the matrix of the similarity scores were fed directly into `stats::prcomp`.

[2] This is the same as network communities, a different naming is used to avoid confusion with microbial communities

[3] PCA or PCoA ordination plots of *samples* are common in microbial ecology, but the topic of this section is PCoA plots of OTUs. This means that we transpose the OTU table compared to the common situation.

3.4.1 Special features included in the plots

In order to connect the OTU PCoA plots to the properties of the samples, we introduce three novel concepts:

r-proportion

In the selection-switch and seawater experiments, the samples can be partitioned into two groups; the samples which are under r-selection at sampling and the ones that are uder K-selection at sampling. Let us assume we have in total N samples, where n_r and n_K of them are under r- and K-selection respectively. We create an heuristic of which OTUs thrive in the r-selected environments by defining the *r-proportion* as the ratio of the OTU's mean abundance in the r-selected samples to its overall mean abundance. This is, the r-proportion of OTU i is given by:

$$\pi_i = \frac{\frac{1}{n_r} \sum_{j \in I_r} x_{i,j}}{\frac{1}{N} \sum_{j=1}^{N} x_{i,j}}, \tag{3.4}$$

where $x_{i,j}$ is the abundance of OTU i in sample j and I_r is the subset of the indices $1, 2, \ldots, N$ which correspond to the r-selected samples.

π_i always lie between zero (not present in the r-selected samples) and $\frac{N}{n_r}$ (only presents in the r-selected samples). Our assumption is that OTUs with large r-proportion are likely r-strategists, whereas the OTUs with a low r-proportion are more likely K-strategists.

Biofilm proportion

The biofilm experiment has samples from two sources; the biofilm on the carriers and in the surrounding water. As a heuristic for bacteria specializing in colonizing the biofilm, we introduce the *biofilm proportion*. This is defined the same way as the r-proportion in Equation (3.4) with the only difference that biofilm samples are considered r-selected samples and the water samples K-selected. According to this definition, we believe that OTUs specializing in growing in the biofilm should have a large biofilm proportion, whereas the OTUs specializing in growing in planctonic environments should have a correspondingly low biofilm proportion.

Favorite treatment

The biofilm experiment has three different treatments. We want to include in the plot where each OTU has the highest mean abundance. Hence, for each OTU i and treatment j, we compute the mean abundance $\bar{x}_{i,j}$ in this treatment. The favorite

treatment is then the treatment where OTU i has the highest mean abundance, this is:

$$f_i = \arg\max_j \bar{x}_{i,j}. \tag{3.5}$$

3.5 Lotka-Volterra approach

3.5.1 Time trajectory plots

The time trajectory plots are PCoA ordinations of the Bray-Curtis similarities between the samples (not the OTUs) based on the time series. The procedure is implemented in `micInt::plot_trajectory`. For each subsetting of the data, the time trajectories from each reactor were plotted together. To compare trajectory plots for different subsettings, an overall ordination was made from all time series and the time trajectories for each subsetting superimposed on this ordination.

3.5.2 Generation of equation systems

The linear systems were created by the log-integral method as described in Section 2.7.2, using `micInt::integralSystem(kind='log_integral')`. Equations for the selection-switch experiment were created only for the absolute data, as we do not think using relative abundances would add any additional value.

Equations corresponding to zero abundances on the left side of Equation (2.19) were removed as replacing the zero abundance with a pseudo-count as done by Freidman and Elm[17] was considered questionable in our opinion. The approach of removing equation provides further complications, the matrix X in Equation (2.20) is no longer the same for all OTUs. Had the matrix been the same for all OTUs as in Stein *et. al.*[19], obtaining all coefficients would have been computationally much faster.

Fitting the coefficients for a single time series only, provides too little information given the large number of coefficients when many OTUs are present. Hence, the data were multiplexed, where equations from the biological replicates (time series run under identical conditions) were stacked on top of each other.

3.5.3 Generation of artificial time series

A community of two OTUs, named x and y, was defined according to the gLV model:

$$\frac{\mathrm{d}\,x}{\mathrm{d}\,t} = x\left(\mu_x + a_{xx}x + a_{xy}y\right) \tag{3.6}$$

$$\frac{\mathrm{d}\,y}{\mathrm{d}\,t} = y\left(\mu_y + a_{yx}x + a_{yy}y\right), \tag{3.7}$$

where the coefficients were given by: $\boldsymbol{\mu} = \begin{bmatrix}\mu_x \\ \mu_y\end{bmatrix} = \begin{bmatrix}2 \\ 1\end{bmatrix}$ and $A = \begin{bmatrix}a_{xx} & a_{xy} \\ a_{yx} & a_{yy}\end{bmatrix} = \begin{bmatrix}0.5 & -1 \\ 1 & -1\end{bmatrix}$. The coefficients were at purpose selected in order to obtain a stable focus at $\begin{bmatrix}x \\ y\end{bmatrix} = \begin{bmatrix}2 \\ 3\end{bmatrix}$. From these equations and parameters, simulations of time series were run and samples were taken at regularly spaced time points. Thereafter, the Lotka-Volterra approach was applied to estimate the parameters from the time series. Each simulation ran from $t = 0$ to $t = 10$. The time between samplings varied geometrically through $10^{-2}, 10^{-1.8}, 10^{-1.6}, \ldots, 10^{0.2}, 10^{0.4}$, and the number of time series being multiplexed was varied through 1 and 10. The simulations were run on each combination of these parameters. Initial abundances for each time series were picked uniformly and independently from the rectangle with corners $\begin{bmatrix}0 \\ 0\end{bmatrix}$ and $\begin{bmatrix}4 \\ 6\end{bmatrix}$, this is: From zero to the double of the steady state abundances. Due to the fact that we in all cases had more equations than unknowns, we set both regularization parameters to zero, meaning that the linear systems were solved by least squares.

After the inferred coefficients $\hat{\boldsymbol{\mu}}$ and $\hat{A}$ were calculated, the mean of the absolute componentwise errors for $\boldsymbol{\mu}$ and A were reported and plotted:

$$d_{\boldsymbol{\mu}} = \frac{\left|\widehat{\mu_x} - \mu_x\right| + \left|\widehat{\mu_y} - \mu_y\right|}{2} \tag{3.8}$$

$$d_A = \frac{\left|\widehat{a_{xx}} - a_{xx}\right| + \left|\widehat{a_{xy}} - a_{xy}\right| + \left|\widehat{a_{yx}} - a_{yx}\right| + \left|\widehat{a_{yy}} - a_{yy}\right|}{4} \tag{3.9}$$

In order to obtain more stable estimates, the values of $d_{\boldsymbol{\mu}}$ and $d_{\boldsymbol{\mu}}$ were averaged over 10 replicates.

3.5.4 Cross-validation

Cross-validation was performed according to Algorithm 1, implemented in `micInt::cv.LV`. Leave-one-out cross-validation inspired by Hastie *et. al.*[56] was used, this means that the number of folds was equal to the number of time series. Root mean squared error (RMSE) was used to assess the cross-validation error. The cross-validating of the parameters λ_{self} and $\lambda_{\text{interactions}}$ was done over a rectangular grid, with geometrically progressing weights. For the selection-switch experiment, the weights for λ_{self} were $10^{-3}, 10^{-2.9}, 10^{-2.8}, \ldots, 10^{2.9}, 10^{3.0}$. All weight combinations are shown in table 3.2.

In order to ensure that the chosen weight parameter values were reasonable, a colorplot of the RMSE for each parameter value was created. In this manner, it can for instance be detected if the parameter values lied on the boundary of the search area, in which case the search area should be extended. If performed properly, the

Table 3.2: Regularization weights used in Lotka-Volterra cross-validation

Experiment	Tuning parameter	Smallest weight	Ratio	Largest weight
Selection-switch	λ_{self}	10^{-3}	$10^{0.1}$	10^{-3}
Selection-switch	$\lambda_{\text{interaction}}$	10^{-1}	$10^{0.1}$	10^{10}
Biofilm	λ_{self}	10^{-3}	$10^{0.05}$	10^{-3}
Biofilm	$\lambda_{\text{interaction}}$	10^{-1}	$10^{0.05}$	10^{10}

optimal cross-validation colorplot should report the optimal weight combination as a minimum well inside the plot.

3.5.5 Predicting the communities

After the coefficients were fitted, the equations were solved using the `deSolve`[57] library implementing an `lsoda` solver. The initial value problem had the same starting condition as one of the time series in the subdivision, this is the reference time series. The predicted and reference time series were shown in a time trajectory plot made from the all time series in the subdivision and the predicted time series. Also, note that the time series to predict also was used in the fitting procedure, which is in general a bad idea because this may lead to overfitting (the algorithm does correctly reproduce the training data, but fails to predict test data not fed into the fitting algorithm).

Algorithm 1 K-fold cross validation to find regularization parameters

Require:

- List (w_i) of length W: The regularization weights to consider
- List $\left(s_j\right)$ of length S: The different time series to consider, all containing data of the same n OTUs
- Integer K: The number of folds

Create a sampling vector g by repeating the sequence $1, \ldots, K$ such that a vector of S elements is obtained
Create fold vector f by sampling without replacement from g
Create arrays $\left(\overline{MAE}_i\right)$ and $\left(\overline{RMSE}_i\right)$ corresponding to mean absolute error and root mean squared error, respectively for each of the parameter combinations
for $i \leftarrow 1$ to W **do**
    Create matrices $RMSE$ and MAE of dimension $K \times n$
    Create regularization matrix D_λ from the regularization weights w_i
    for $j \leftarrow 1$ to K **do**
        for $k \leftarrow 1$ to n **do**
            Create X and $\mathbf{d}$ for OTU k by log-integral method, using observations not in fold j (use time series s_k if $g_k \neq j$)
            Create $\widetilde{X}$ and $\widetilde{\mathbf{d}}$ for OTU k by log-integral method, using observations in fold j (use time series s_k if $g_k = j$)
            $\mathbf{a}^* \leftarrow \left(XX^\intercal + D_\lambda\right)^{-1} X^\intercal \mathbf{d}$
            Let m be the number of rows in X and $\mathbf{d}$
            $\mathbf{v} \leftarrow \widetilde{X}\mathbf{a}^* - \widetilde{\mathbf{d}}$
            $MAE_{j,k} \leftarrow \frac{1}{m} \sum_{l=1}^{n} |v_l|$
            $RMSE_{j,k} \leftarrow \sqrt{\frac{1}{m} \sum_{l=1}^{n} v_l^2}$
    $\overline{MAE}_i \leftarrow \frac{1}{K \cdot n} \sum_{j=1}^{K} \sum_{k=1}^{n} MAE_{j,k}$
    $\overline{RMSE}_i \leftarrow \sqrt{\frac{1}{K \cdot n} \sum_{j=1}^{K} \sum_{k=1}^{n} RMSE_{j,k}^2}$
return $\overline{MAE}$ and $\overline{RMSE}$

Chapter 4

Results

In this chapter, we will present the results according to the following outline:

- Bacterial density graph for the selection-switch experiment
- Results from ReBoot procedure: Tables, diagnostic plots and networks
- PCoA plots of OTUs based on both raw abundances and ReBoot similarity scores
- Results from Lotka-Volterra approach: Trajectory plots, inference accuracy on a simulated community, cross-validation colorplots and predictions of time series

4.1 Cell count in the selection-switch experiment

The selection-switch experiment is the only one where the bacterial density is determined, by flow cytometry. Even though the cell count by ifself is not the primary focus in this thesis, it is nevertheless useful to study as it helps explaining the differences between absolute and relative abundances. From Figure 4.1, we see that the bacterial density can change by a factor of two in less than a day, suggesting that just using the relative abundances might give misleading conclusions about the dynamics.

Figure 4.1: Bacterial density during the selection-switch experiment (each line represents a reactor), determined by flow cytometry. Notice the rapid changes in density after the selection regime was switched at day 29.

4.2 Interactions identified by ReBoot approach

4.2.1 Comparison of similarity measures

For each combination of dataset, subdivision and simililarity measure, we report the number of significant correlations found and the proportion of them shown to be negative. An example showing the results for the full selection-switch experiment with absolute abundances is shown in Table 4.1. For subsets of the selection-switch experiment with absolute abundances, the selection-switch experiment with relative abundances and the other experiments, the results are reported in Appendix A.

We first observe that mean-scaled similarity measures did not give any significant interactions. Because of this, we conclude that such similarity measures have no or little power of inferring interactions, and we will not consider mean scaling for the rest of this thesis. Also, the squared Euclidean similarity detected none or only a few significant interactions, regardless of modifications. On the other hand, the non-parametric measures (Spearman correlation, Kendall's tau and nc.score) gave more significant interactions than the other similarity measures in most cases. However, most of the interactions found by the non-parametric similarity measures disappeared when applying noise. This effect of noise was the most extreme on the small seawater experiment and the subdivisions of the larger experiments. The latter observation is as expected because larger datasets should give more robust conclusions. The parametric similarity measures Bray-Curtis, cosine, generalized Jaccard and Pearson yielded fewer interactions than the non-parametric similarity measures, but these interactions were much more robust to noise. This observation contradicts the former view that Spearman correlation is more robust than Pearson correlation[58]. However, we have an explanation for this finding:

The microbial datasets have many zeros in them. As the non-parametric similarity measures are based solely on the ranks of the observations, we think that shared zeros (samples where both OTUs have zero abundance) will make the abundance vectors seem more similar and this may lead to a high similarity score. Indeed, the non-parametric similarity measures were the ones reporting the highest number of significant interactions. Shared zeros causing abundance vectors to appear more similar, is in general an undesired artifact. This is because we do not expect bacteria to be more similar due to features they do not share[1]. Adding noise to the data would distort the pattern of common zeros. Also, in this case, the hypothesis appears to be correct, as most significant interactions disappear after applying noise. By the same argument we can explain why the parametric similarity measures reported fewer, but more robust interactions: Shared zeros does not make the abundance vector more similar and adding low noise levels would have small effects as parametric similarity measures respond continuously to its input.

Notice that any positive noise level γ would theoretically disrupt the pattern

[1] A stupid analogy could explain this reasoning: Both human and stones cannot fly, lack wings, beaks and feathers, but this shared lack of features does not make humans more alike stones.

of shared zeros. Testing the effect of noise was only done for a fixed noise level. It would be interesting to assess how different levels of noise would decrease the number of significant interactions. According to our hypothesis of shared zeros, the reported number of significant interactions for the non-parametric similarity measures would then drop instantly after adding noise, and then show a slower decline at higher noise levels.

Another difference between the parametric and non-parametric similarity measures is the sign of the interactions. Generally, the Pearson correlation yielded a lower proportion of negative interactions than the signed non-parametric similarity measures. Also, when applying noise to non-parametric similarity measures, the percentage of negative interactions decreased. The cosine similarity is in practice unsigned for our purposes as all abundances are positive. We did not find any major difference between the proportion of negative interactions in the biofilm and water samples for the biofilm experiment. Also, for the selection-switch dataset there was no consistent trend that r-selected reactors had more or less negative interactions than the K-selected reactors.

To some degree, there were differences between absolute and relative abundances. When replacing absolute abundances with relative ones for the selection-switch experiment, we generally obtained more significant interactions for the parametric similarity measures, whereas the difference for the non-parametric similarity measures was small. However, the inferred interactions from the absolute data were more robust to noise. There was also a tendency to obtain more negative interactions from relative data, but the difference was less visible.

Table 4.1: Performance of the different similarity measures on the overall absolute data from the selection switch experiment

Name of similarity measure	Signed	Type of similarity measure	Number of significant interactions	Proportion of negative interactions
bray_curtis	FALSE	parametric	628	
bray_curtis_normal	FALSE	parametric	561	
bray_curtis_scaled	FALSE	parametric	0	
bray_curtis_scaled_normal	FALSE	parametric	0	
bray_curtis_scaled_uniform	FALSE	parametric	0	
bray_curtis_uniform	FALSE	parametric	556	
cosine	TRUE	parametric	519	0.00
cosine_normal	TRUE	parametric	517	0.00
cosine_uniform	TRUE	parametric	520	0.00
generalized_jaccard_index	FALSE	parametric	543	
generalized_jaccard_index_normal	FALSE	parametric	493	
generalized_jaccard_index_scaled	FALSE	parametric	0	
generalized_jaccard_index_scaled_normal	FALSE	parametric	0	
generalized_jaccard_index_scaled_uniform	FALSE	parametric	0	
generalized_jaccard_index_uniform	FALSE	parametric	503	
jaccard_index	FALSE	presence-absence	857	
kendall	TRUE	non-parametric	3354	0.38
kendall_normal	TRUE	non-parametric	1782	0.34
kendall_uniform	TRUE	non-parametric	1772	0.34
mutual_information	FALSE	non-parametric	2735	
mutual_information_normal	FALSE	non-parametric	1097	
mutual_information_uniform	FALSE	non-parametric	1091	
nc_score	TRUE	non-parametric	3348	0.38
nc_score_normal	TRUE	non-parametric	1753	0.33
nc_score_uniform	TRUE	non-parametric	1752	0.34
pearson	TRUE	parametric	527	0.20
pearson_normal	TRUE	parametric	516	0.19
pearson_uniform	TRUE	parametric	516	0.19
spearman	TRUE	non-parametric	3333	0.38
spearman_normal	TRUE	non-parametric	1787	0.33
spearman_uniform	TRUE	non-parametric	1791	0.34
squared_euclidean	FALSE	parametric	0	
squared_euclidean_normal	FALSE	parametric	0	
squared_euclidean_scaled	FALSE	parametric	0	
squared_euclidean_scaled_normal	FALSE	parametric	0	
squared_euclidean_scaled_uniform	FALSE	parametric	0	
squared_euclidean_uniform	FALSE	parametric	0	

For our further analysis, we selected the Pearson correlation, the Spearman correlation, the cosine similarity and the Bray-Curtis similarity. In our experience, these four similarity measures combined are for the most part representative for the results of the other similarity measures.

4.2.2 Diagnostic plots

For the selected similarity measures, we made diagnostic plots showing:

- Abundance product versus q-value for each interaction. This is shown in Figure 4.3 for the selection-switch experiment with absolute abundances. For the other datasets, consult Appendix B.
- The number of interactions for each OTU versus its mean abundance. This is shown in Figure 4.3 for the selection-switch experiment with absolute abundances and in Appendix B for the other datasets.
- The similarity score versus q-value for each significant interaction. This is found for in Figure 4.4 for the selection-switch experiment with absolute abundances. For the other datasets, see Appendix B.

We note that for the biofilm and seawater experiments, there does not seem to be a clear trend between mean overall abundance and the number of interactions in which the OTU is involved. For the selection-switch experiment however, we observed that the OTUs which are very rare, had few or none significant interactions. For the more abundant OTUs, this pattern is missing. In the same manner, we saw no clear evidence that a low q-value correlates with a high abundance product. This applied to all dataset and similarity measures. One outlier in the abundance product plot from the seawater experiment (Figure B.1) is worth to note. It has both the highest abundance product and the lowest q-value. This interaction is the one between the two dominant *Vibrio* OTUs present in the experiment. From our discussion of signal-to-noise ratio in Section 3.3.2, our findings are unexpected. These strange patterns might be an indication that the ReBoot pipeline treats OTUs with low and high abundance almost equally. If so, this is an undesired feature. OTUs with low abundance are more sensitive to noise and perturbations, yielding more uncertain results. Hence, we may propose that the ReBoot algorithm does not capture the real aspects of significance well enough. We could of course use statistical methods to assess the patterns present in the abundance product plots and the plots showing number of significant interactions. However, we think such analysis would be little informative, as very weak associations could result in statistically significant results without having any biological significance. We can also observe from the abundance product plots that most significant interactions have a q-value close to the cutoff of 0.05. This observation is as expected, as we would assume interactions with low q-values to be rarer.

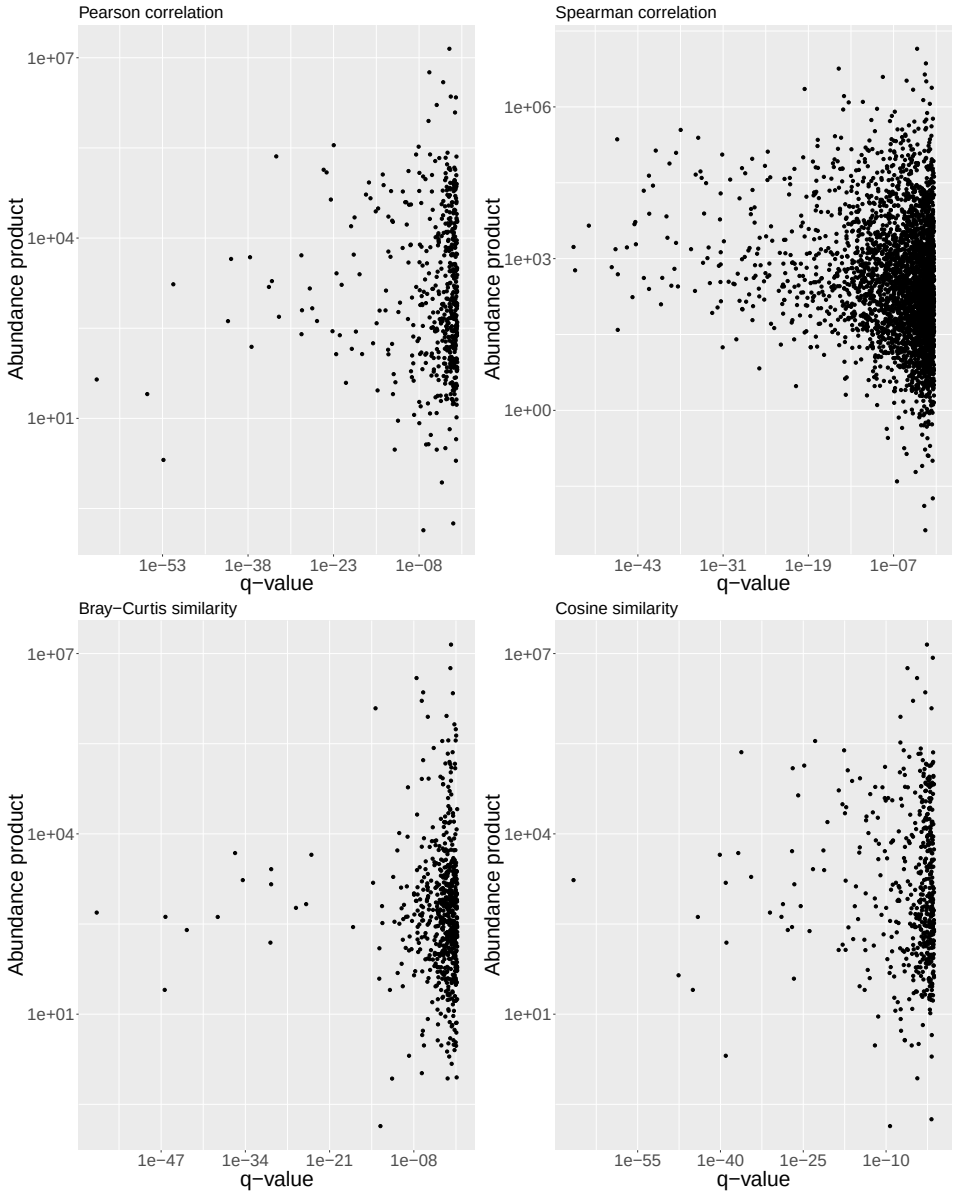


Figure 4.2: Abundance product versus q-values for each significant interaction in the selection-switch experiment with absolute data

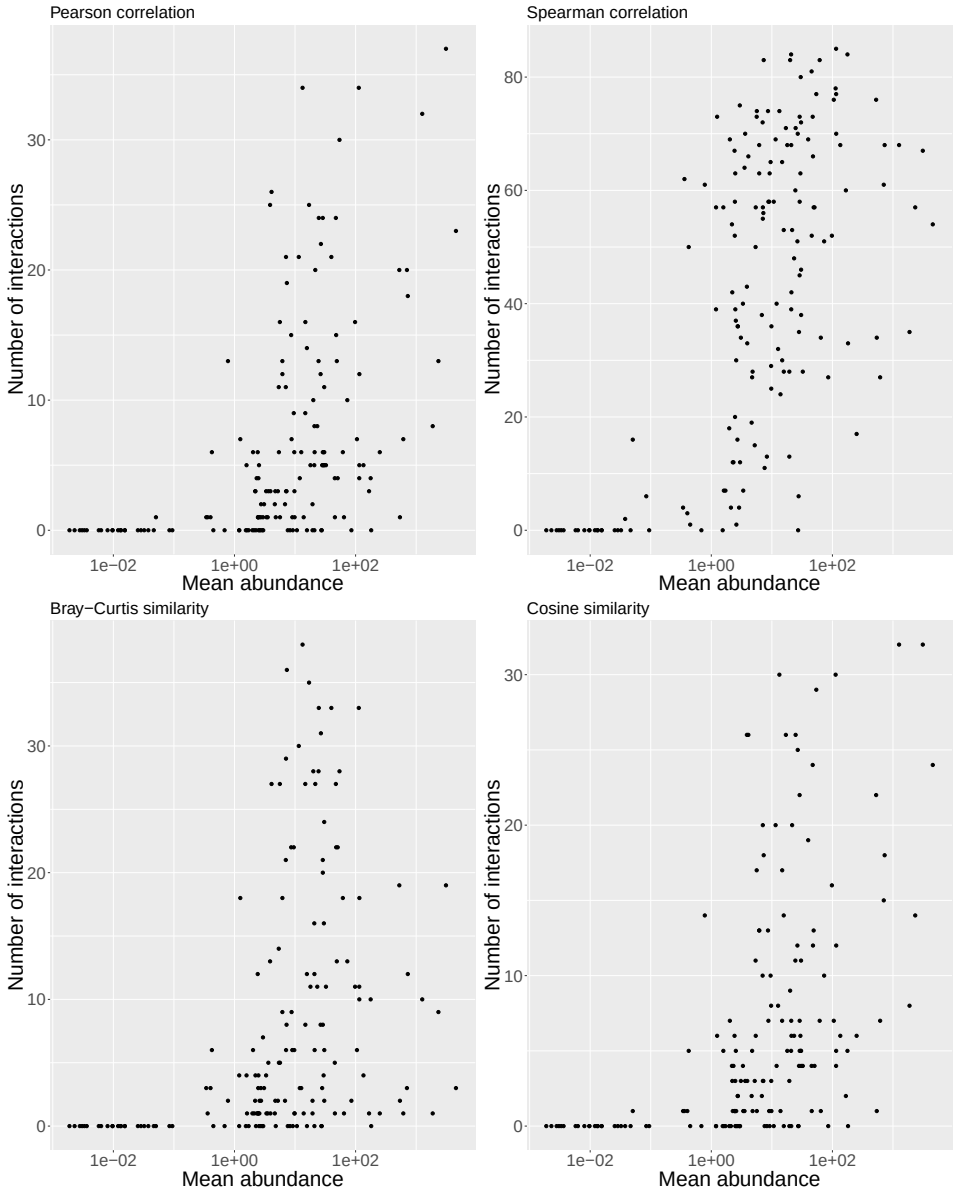


Figure 4.3: Number of significant interactions versus overall mean abundance for each OTU in the selection-switch experiment with absolute data

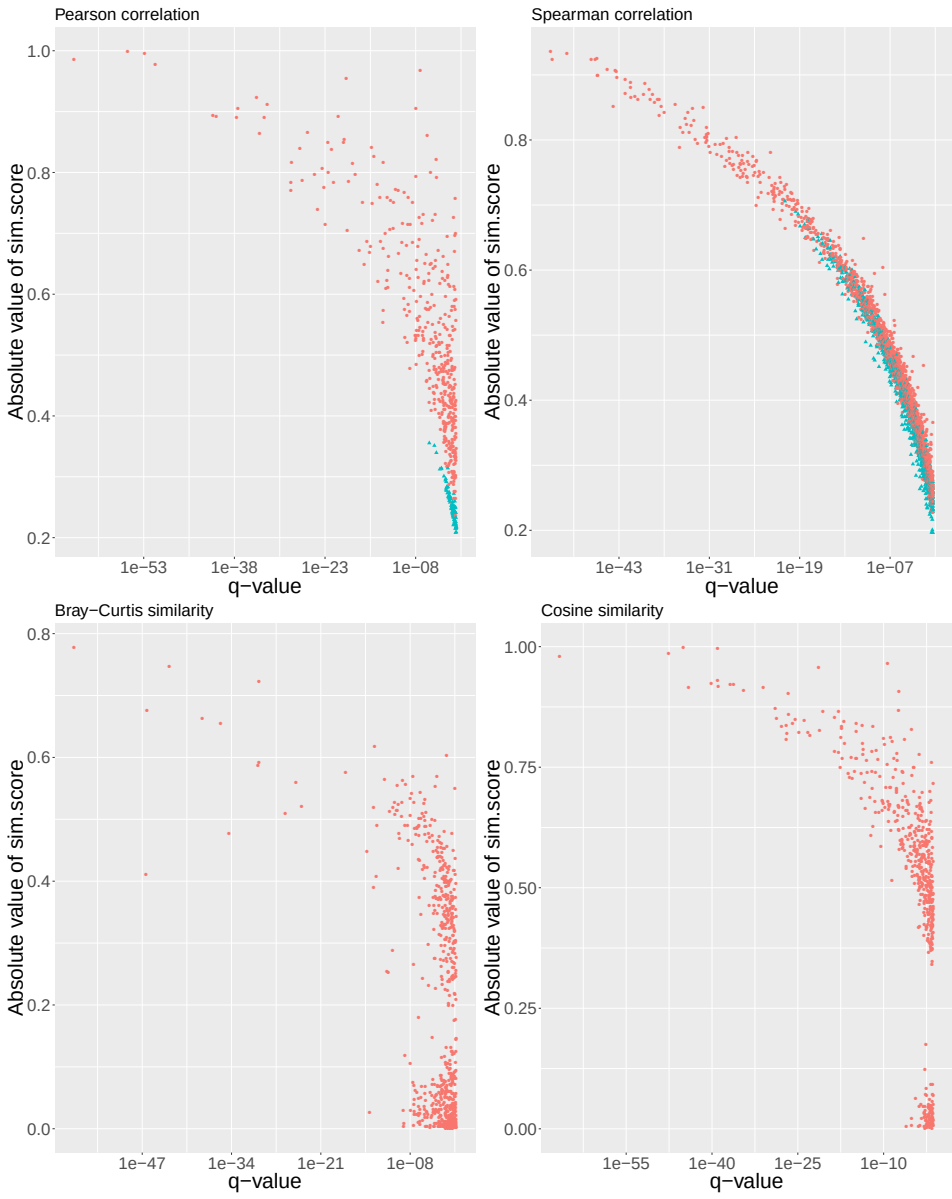


Figure 4.4: Similarity scores versus q-values for each significant interaction in the selection-switch experiment with absolute data. Red circles indicate positive interactions, whereas blue triangles indicate negative interactions.

For the most part, we would expect the correlations with the most extreme (furthest away from zero) similarity scores to be regarded as the most significant. However, due to the approach of assessing the significance from a pooled Z-test, the association does not need to be monotonous. The Spearman correlation did indeed provide almost an exact monotonous relationship between similarity scores and q-values for the selection-switch experiment and the biofilm experiment. The pattern was present, but weaker for the parametric Pearson, Bray-Curtis and cosine similarities on data from the selection-switch experiment and the biofilm experiment. However, in the small seawater dataset, the correspondence between similarity scores and q-values was poor for all similarity measures. The negative interactions were generally not among the interactions having the lowest q-value or highest absolute similarity score value. The only exception was the seawater dataset with Spearman correlation, where there were negative interactions among the most significant ones. Hence, it appears that the positive interactions are more pronounced than the negative ones.

The Bray-Curtis and cosine similarities reported significant interactions having similarity scores close to zero, whereas the Pearson and Spearman correlations gave almost no interactions having a similarity score less then 0.2. We would not expect a similarity score close to zero to correspond to any significant interaction, so this finding is indeed strange. Comparing the diagnostic plots from the selection-switch experiment with absolute and relative abundances, we can hardly see any difference.

Note the strange fact that some of the OTUs in Figure B.8 have mean abundances *less* than 10^{-4}. This is due to the fact that the filtering of OTUs at mean abundance 10^{-4} was done *prior* to removing the samples from day 1, 2, 29, 30 and 32.

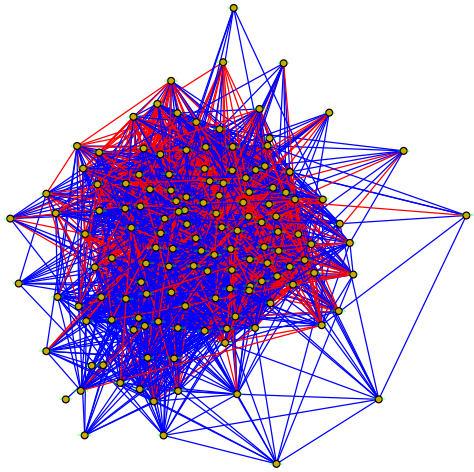


Figure 4.5: Network from the selection-switch experiment with absolute abundances using Spearman correlation. All significant interaction are retained. Blue edges correspond to positive interactions, whereas red edges correspond to negative interactions.

4.2.3 Networks of interactions

We created network of networks of the interactions from the ReBoot results. Keeping as many edges as in Figure 4.5 resulted in a hairball difficult to interpret. Subsequently, we restricted ourself to visualize only the 200 most significant edges (ranked by q-value). In addition to our four selected similarity measures, we also view the result of the noised versions of the similarity measures (with normally distributed noise). This is to evaluate the robustness of the networks. For the selection-switch experiment with absolute abundances, the results are shown in Figure 4.6, while the other networks reside in Appendix C.

We first observe the general trend that the networks have clusters of positively interacting OTUs, whereas negative interactions form preferentially between clusters. Given the same data, the same OTU clusters are in many cases recognizable across the similarity measures. Indeed, when taking the intersection (not shown) of the networks in Figure 4.6, the resulting network has 49 edges. This observation strengthen our findings that the similarity measures give consistent results. By visual inspections, none of the networks seem be result of uniformly random linkage (Erdős-Rényi model,), small-world (Watt-Strogatz) or scale-free (Barabási-Albert). Noise has a distorting effect on the networks, but the differences are for the most part minor and the different parts of the networks are generally recognizable. The

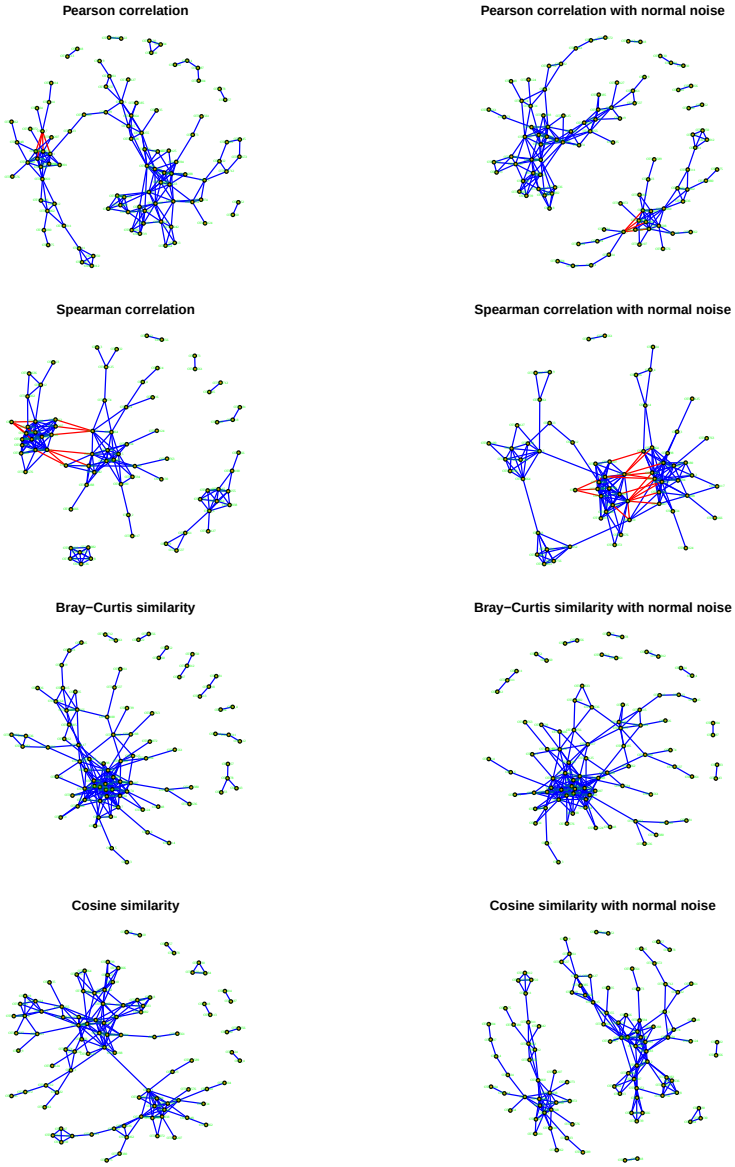


Figure 4.6: Network of significant interactions for the selection-switch experiment with absolute data. Blue edges correspond to positive interactions, whereas red edges correspond to negative interactions.

only major exception is when noise is applied to the Spearman correlation on the seawater experiment, where almost all significant interactions disappear. For the selection-switch experiment, the networks for the relative and absolute abundances are quite similar for the Spearman correlation and Bray-Curtis similarity. For the cosine similarity and Pearson correlation though, we observe that the networks created with absolute data create (almost) disjoint clusters, whereas for the relative data the two parts of the networks are more interconnected by negative interactions. These differences are less visible in networks showing all significant edges (not shown).

For the selection-switch experiment with absolute data, we detected two major interaction clusters in the networks. In order to investigate these further, we used the network provided by the Spearman similarity with normally distributed noise, kept only the positive interactions, detected interaction clusters and finally added the negative interactions. The result is shown in Figure 4.7. We found community 1 and 5 to be the most prominent as they have positive interactions within, but negative interactions across. We next created a phylogenetic tree of the OTUs being present in the network, shown in Figure 4.8. From this, we notice that the interaction clusters correspond quite well with the taxonomies. According to Gundersen[12], the phylogentic branches of most OTUs in community 1 (*Alphaproteobacteria* and *Flavobacteria*) are generally K-strategists, whereas the *Gammaproteobacteria* OTUs dominating community 5 are for the most part r-strategists.

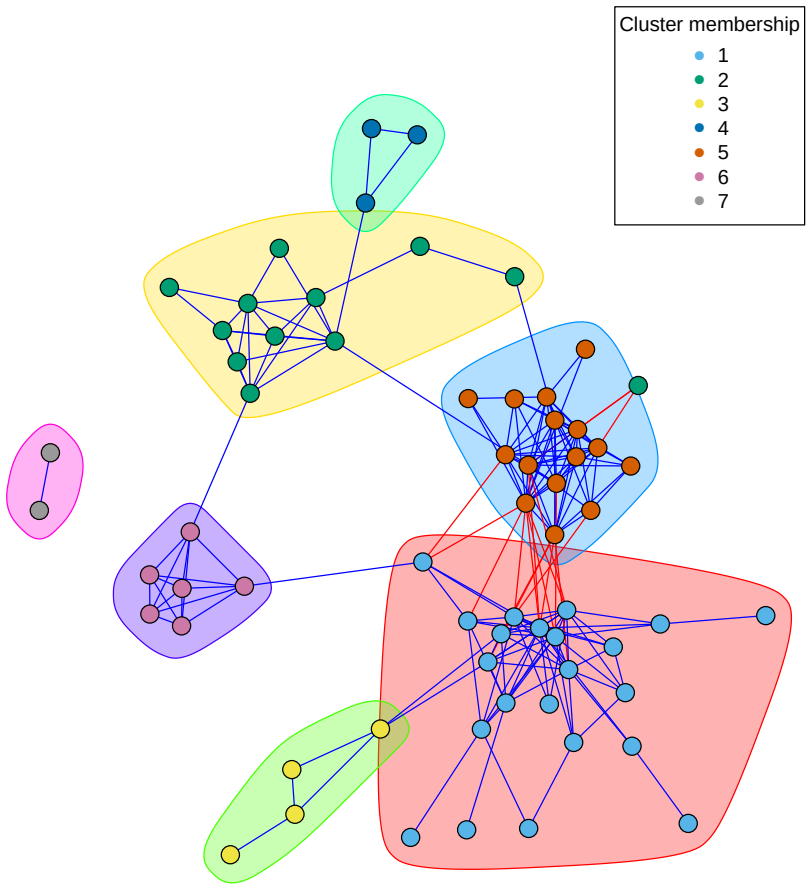


Figure 4.7: Community-labeled interaction network using absolute data from the selection-switch experiment and the Spearman corrrelation with added normally distributed noise. Blue edges correspond to positive interactions, whereas red edges correspond to negative interactions. The legend sorts the interaction clusters by the node color, not the background community marking.

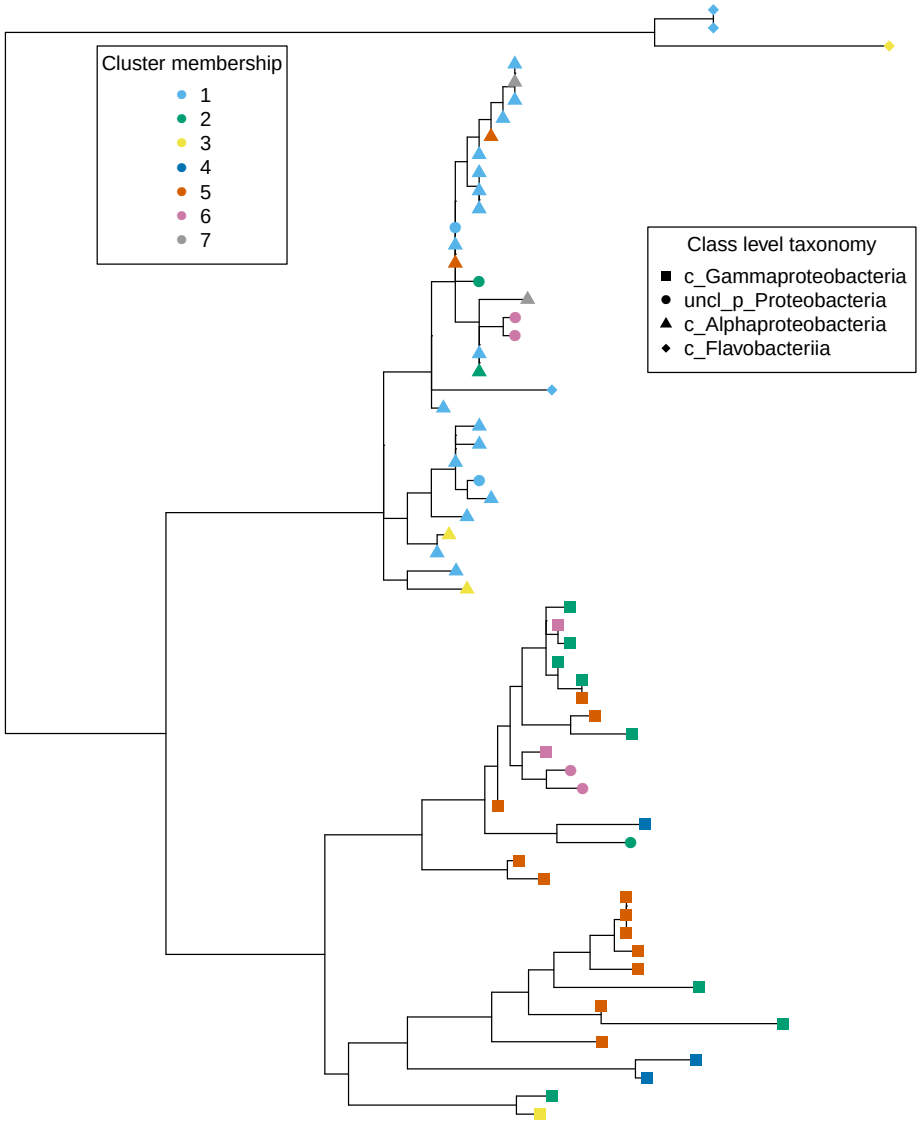


Figure 4.8: Phylogenetic tree of the OTUs shown as nodes in Figure 4.7. The coloring corresponds to the interaction clusters in the network. Also, the taxonomy on class level is included as shapes.

4.2.4 PCoA plots

To assess how the similarity measures structure the community, we present PCoA ordinations of the OTUs. For this procedure, the full experiments were used, not subdivisions. However for the selection-switch experiment, the samples taken at day 1, 2, 29, 30 and 32 were still filtered away. If the samples taken at these transient days not were removed, the OTUs dominating under r- and K-selection do not separate well in the plots.

We first made the ordinations based on the pairwise similarity scores of the filtered OTU table without applying the ReBoot procedure. The results are shown in Figure 4.9 for absolute abundances from the selection-switch experiment and in Appendix D for the other datasets.

Next, we compared these results with ordinations based on the similarity scores from ReBoot. These are shown in figure Figure 4.10 for absolute abundances from the selection-switch experiment. The rest of the plots are in Appendix D.

We would guess that the r- and K-strategists being present in the seawater experiment and the selection-switch experiment would have different signatures, allowing PCoA plots to distinguish them. Indeed, this is the case for both the seawater and selection-switch experiments. For the biofilm experiment, the separation between bacteria mainly found on biofilm carriers and those mainly found in water is not that easy to spot.

The choice of similarity measure did have an effect, the parametric Pearson and cosine similarities had the greatest success of separating the points r- and K-strategists, whereas the Bray-Curtis similarity was poorer to provide this separation. The generalized Jaccard index, the categorical Jaccard index, squared Euclidean similarity and mutual information coefficient (plots are omitted in this thesis) were even worse in this regard. For the biofilm experiment only the Pearson and cosine similarities did suffice in providing a clear distinction between OTUs dominating in water and on biofilm carriers. We also recognize several of the latter similarity measures to give few or none significant interactions, suggesting they are poor of capturing the structure resident in the data.

Notice that for the selection-switch experiment, the absolute and relative abundances resulted in very similar plots. Likewise, there is hardly any large difference between the ordination plots made with the ReBoot pipeline and those made without it, even though points do shift a bit and the plots get rotated. However, for the seawater experiment we may though see a slightly better separation between r- and K-strategists applying the ReBoot results. This may be attributed to the fact that the dataset is small and thus can benefit from the extra robustness provided through bootstrapping.

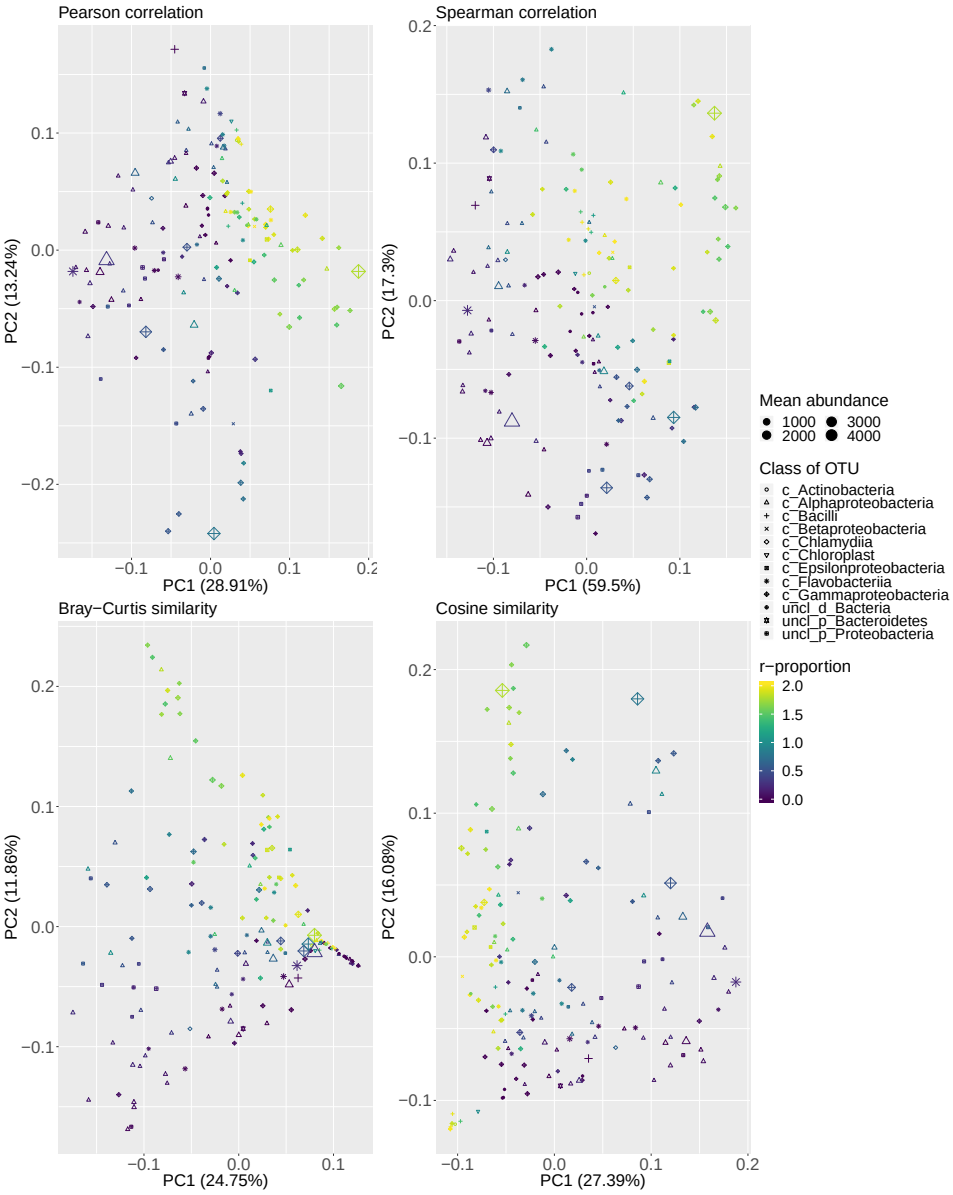


Figure 4.9: PCoA ordination plots for the absolute data from the selection-switch experiment without the using ReBoot pipeline. The r-proportion is described in Section 3.4.1. In this plot, the taxonomies of the OTUs on class level are included.

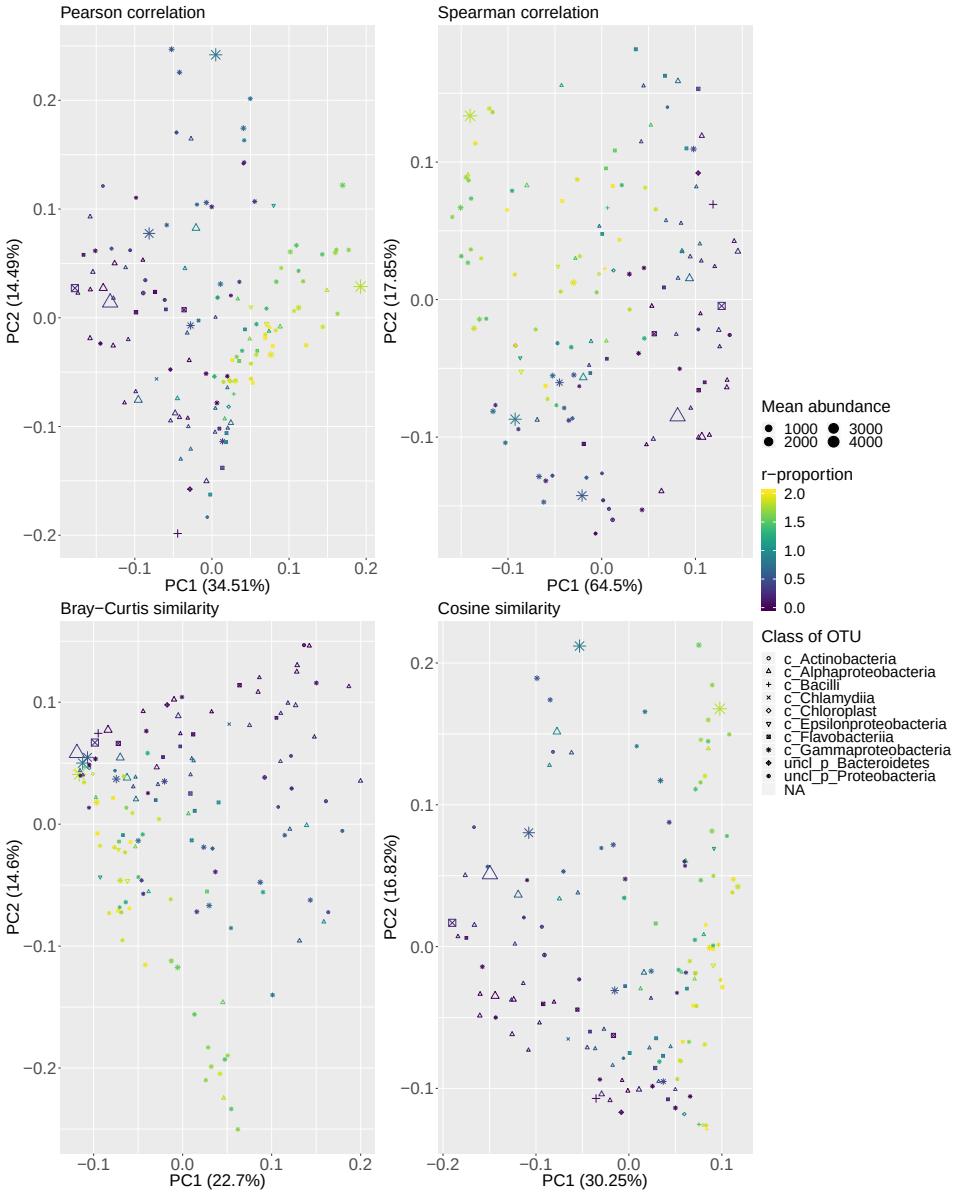


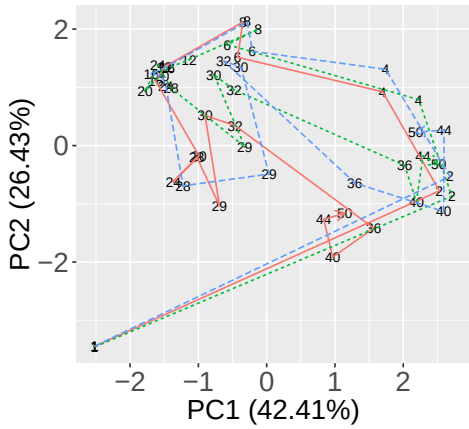
Figure 4.10: PCoA ordination plots for the absolute data from the selection-switch experiment using the ReBoot pipeline. The r-proportion is explained in Section 3.4.1. In this plot, the taxonomies of the OTUs on class level are included.

4.3 Population dynamics identified by Lotka-Volterra modeling

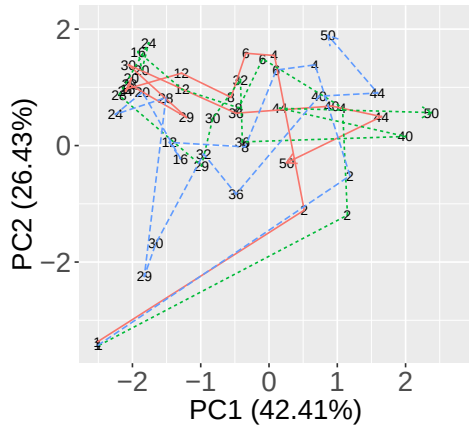
4.3.1 Trajectory plots

We start by showing trajectory plots for the time series in order to get an impression of how the communities evolve over time. For the selection-switch experiment, the results for absolute abundances are shown in Figure 4.11. Consult Appendix E for the rest of the plots.

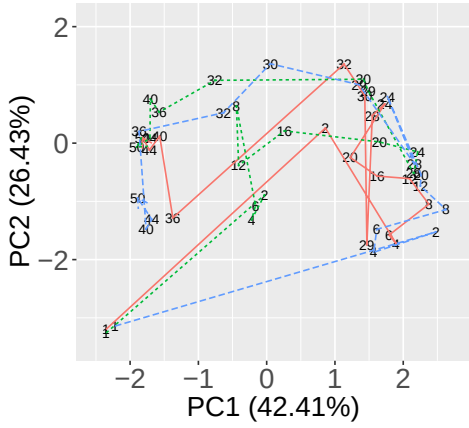
For the biofilm experiment(Figure E.1), the water time series end up in somewhat the same location in the ordination plot, suggesting convergence towards a more stable community. The same tendency can be spotted for the biofilm carriers even though time series `Carrier6` is diverging from this pattern. The selection-switch experiment shows an even clearer sign of deterministic behavior, where the independent time series follow trajectories close to each other under K-selection. Moreover, communities under K-selection do converge to the same area in the ordination plots, regardless of the starting conditions. Under r-selection however, the trajectories show greater variation and less consistency. However, we observe a tendency for r-selected communities to stay in the same area in the plots. As for the PCoA plots, diagnostic plots and interaction networks, the differences between time series trajectories for relative data and absolute data seem to agree about the main trends.



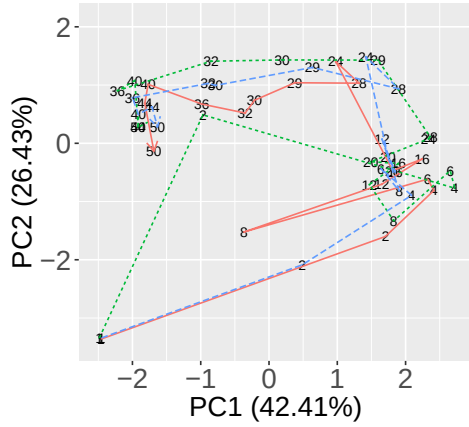
(a) K-selected community before day 29, r-selected community after day 29, with high carrying capacity. The time series plotted are: KRH1 (red, solid), KRH2 (green, short dashes) and KRH3 (blue, long dashes).



(b) K-selected community before day 29, r-selected community after day 29, with low carrying capacity. The time series plotted are: KRL1 (red, solid), KRL2 (green, short dashes) and KRL3 (blue, long dashes).



(c) r-selected community before day 29, K-selected community after day 29, with high carrying capacity. The time series plotted are: RKH1 (red, solid), RKH2 (green, short dashes) and RKH3 (blue, long dashes).



(d) r-selected community before day 29, K-selected community after day 29, with low carrying capacity. The time series plotted are: RKL1 (red, solid), RKL2 (green, short dashes) and RKL3 (blue, long dashes).

Figure 4.11: Trajectory plots for the selection-switch experiment with absolute abundances. The data are ordinated by PCoA using Bray-Curtis similarity. The text labels on the points correspond to the day of sampling. All time series in the overall figure were used to make the ordination. Later, the time series stemming from identical selection regimes were superimposed on each individual subfigure.

4.3.2 Simulated community

In order to assess the interference algorithm, we simulated an artificial community of two OTUs as described in Section 3.5.3. It follows the gLV equations and has the true parameters $\boldsymbol{\mu} = \begin{bmatrix}\mu_x \\ \mu_y\end{bmatrix} = \begin{bmatrix}2 \\ 1\end{bmatrix}$ and $A = \begin{bmatrix}a_{xx} & a_{xy} \\ a_{yx} & a_{yy}\end{bmatrix} = \begin{bmatrix}0.5 & -1 \\ 1 & -1\end{bmatrix}$. A sample phase plane plot is shown in Figure 4.12 and is included for the sole purpose of giving the reader an impression of how the system behaves. Based on the simulations, samples were taken and our inference algorithm was used to estimate the gLV coefficients from the artificial data(see Section 3.5.3 for more details). Finally, the estimated coefficients were compared to the real ones. The results are shown in Figure 4.13.

Because the inference algorithms consider the difference in OTU abundances between consecutive time points, more frequent sampling should increase the accuracy of the predictions. The observations from the plot agree with this expectation. Having time between samples corresponding to approximately one order of magnitude lower than the maximal growth rates turned out to perform decently (time step 0.1, approximate absolute error 10^{-2}), but with time steps ten times larger, the accuracy was unreliable. Even though the integral and log-integral method both provided reliable results at frequent sampling rates, the log-integral method always yielded better results, agreeing with the statement made by Kloppers and Greeff[50] that the log-integral method is an improvement over the integral method.

The effect of multiplexing time series had the opposite effect of what was expected, adding more data into the fitting decreased accuracy, an effect being most pronounced at frequent sampling rates. This observation is strange given the fact that more information is available with more time series. One possible explanation is that with more data, numerical instability could be an issue.

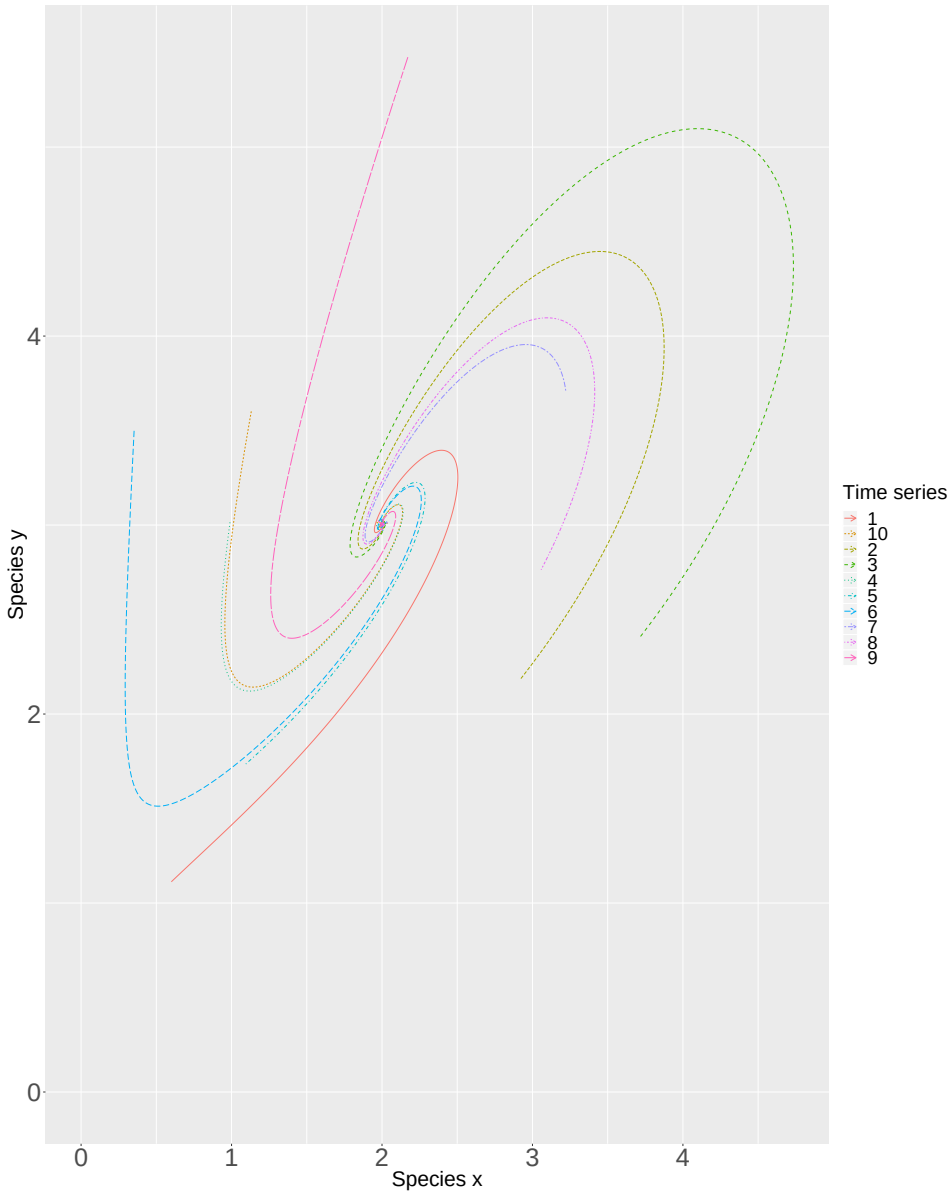


Figure 4.12: Sample phase-plane plot of simulated time series

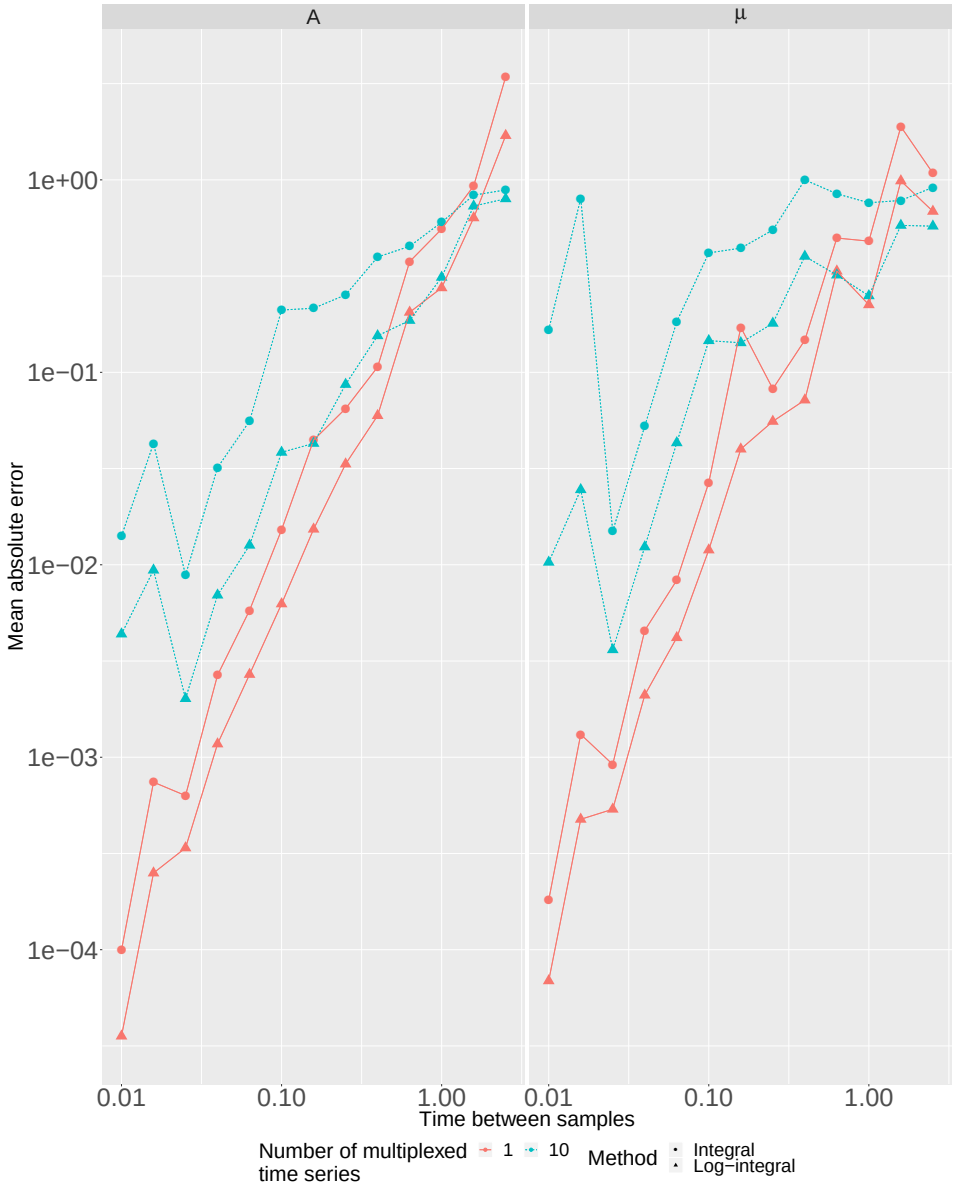


Figure 4.13: Accuracy of predicting Lotka-Volterra coefficients from the artificial community. The y-axis corresponds to the mean of the absolute componentwise error of A and $\boldsymbol{\mu}$, respectively. The numbers are averaged over 10 replicates.

4.3.3 Cross-validation

Our real datasets contain more OTUs than samples, so we had to fit regularization parameters in order to infer the Lotka-Volterra coefficients. This was done by cross-validation as described in Section 3.5.4 and the results are summarized in colorplots. For the selection-switch experiment, the results are shown in Figure 4.14, while they are shown for the biofilm dataset in Appendix F. Also, a more stringent filtering of OTUs at 10^{-3} mean abundance was done for the selection-switch experiment and used as basis for another cross-validation colorplot, shown in Appendix F.

For the plots from the selection-switch experiment, the cross-validation error decreases with the increasing $\lambda_{\text{interaction}}$, no matter how big it is. For the really large values ($10^8 - 10^{10}$) of $\lambda_{\text{interaction}}$, the variations in cross-validation error are nevertheless small. This applies both to the normal and stringent filtration. Hence, we cannot find any minimum. For the biofilm experiment however, we find something that resembles a minimum in most of the plots, but at the same time we observe that the regularization parameters have minimal effect on the cross-validation errors.

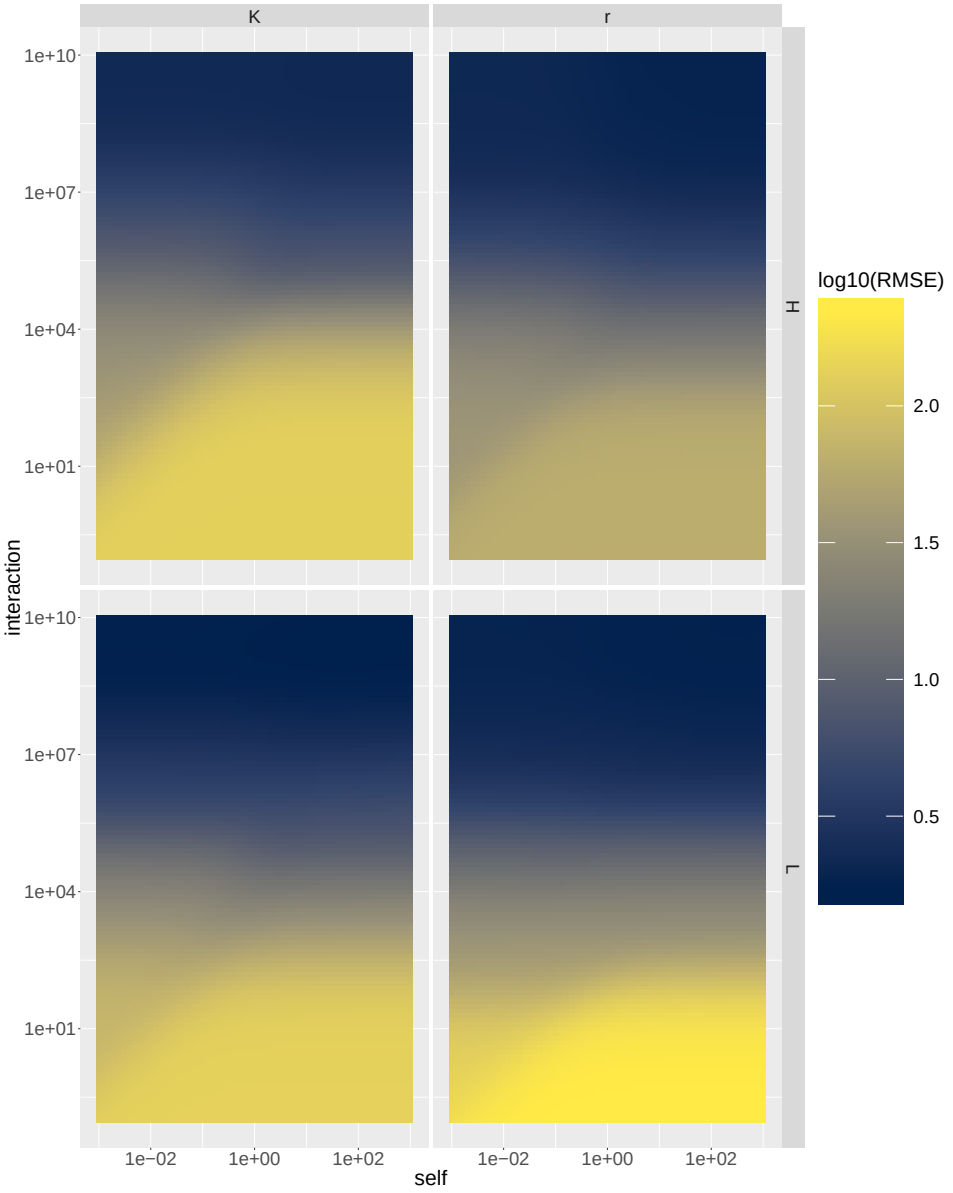


Figure 4.14: Cross-validation results for the selection switch-dataset with ordinary filtering at mean relative abundance 10^{-4}. The color reported corresponds to base-10 logarithm of the Root Mean Squared Error. The plots are shown for each combination of present selection regime (in columns) and nutrient supply (in rows).

4.3.4 Predicting the communities

After the regularization parameters were selected for each time series, the systems were solved for each subdivision, obtaining the Lotka-Volterra coefficients. For each subdivision (having its own fitted set of coefficients), one time series was used as a reference for a predicted time series, having the same initial condition. For the biofilm experiment, the predicted time series diverged (exploded toward infinity), so they could not be shown. For the selection-switch experiment with ordinary filtering, the trajectory plots of the predicted and reference time series are shown in Figure 4.15. For the stringent filtering, consult Appendix G

We observe that the predicted time series do not resemble the actual time series at all, even though the starting point is the same. We also notice that the predicted time series possess much less dynamics (are more stationary) than the actual ones. This is indeed not strange given that optimal values of $\lambda_{\text{interaction}}$ selected from the colorplots are very large. High values of the regularization parameters shrink the coefficients, resulting in coefficients having a small magnitude[51]. Quite the opposite seems to be the case for the biofilm experiment where the predicted time series diverge. Note however that the Lotka-Volterra modeling approach was never intended to be applied for relative abundances[19], and the predictions for the biofilm dataset were not normalized to account for the sum-to-one constraint present in the data.

The failure to find optimal tuning parameters and predicting the communities, may come from various sources. From Section 4.3.2, we learn that our inference algorithm should work for small systems of OTUs when the sampling rate is frequent enough. However, our real datasets have samples further apart than recommended. The generation time of bacteria grown in the lab might be as low as tens of minutes, but the generation time of wild bacteria in nature remains largely unknown[59]. However, we might assume that the generation time for many bacteria in our experiments is one day or less, given how rapidly the cell count in Figure 4.1 fluctuates. Given the fact that the samples from the selection-switch experiment were generally taken four days apart (the sampling was denser at the start and right after the selection switch), we thereby conclude that the time resolution was insufficient in order to make any reasonable inferences. For the biofilm experiment, which was in general sampled every two weeks, the situation is even worse. We also suspect suspect that the high numbers of OTUs present in our datasets poses additional challenge for our algorithms to predict the communities accurately, but we do not have any simulations exploring this effect.

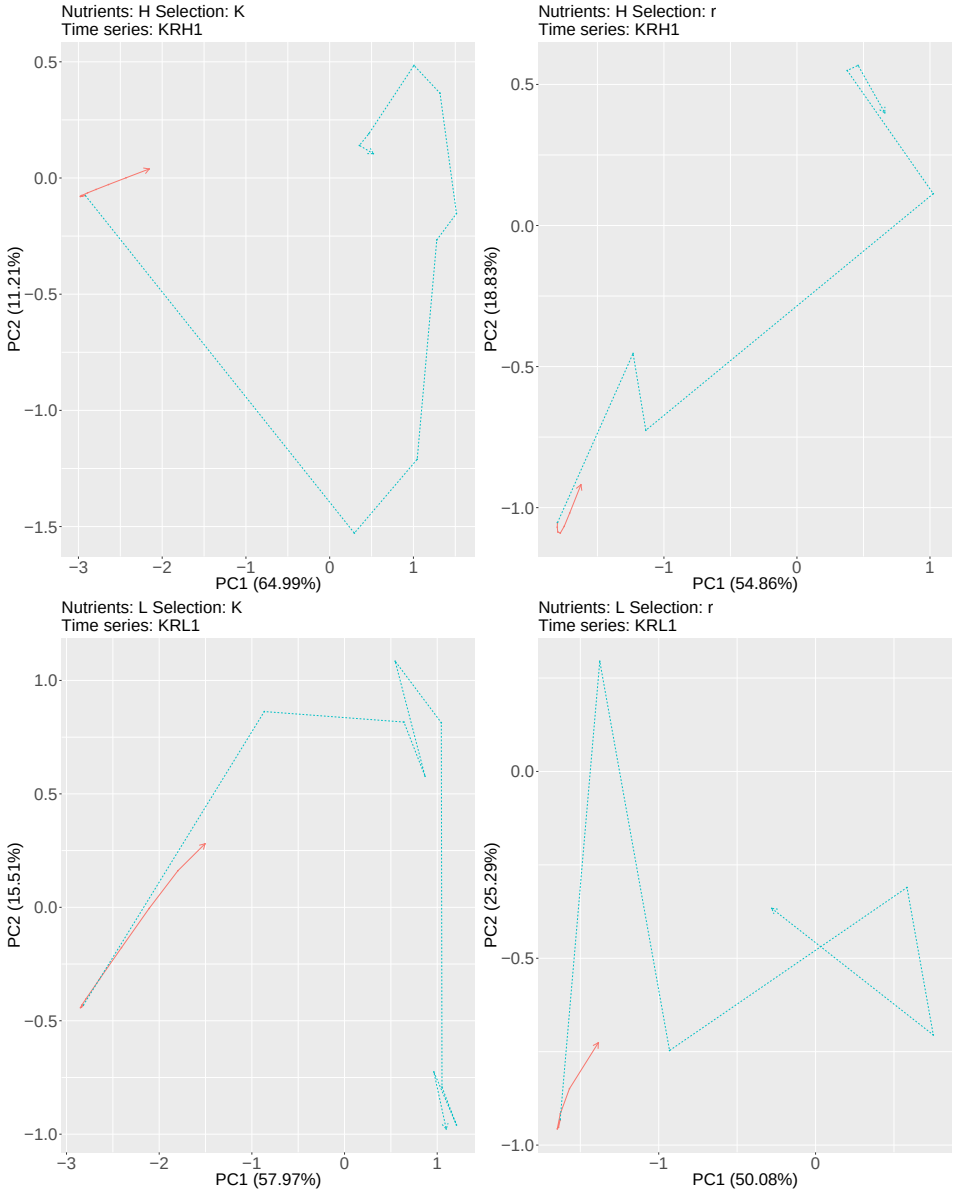


Figure 4.15: Predicted (determined through inferred gLV coefficients, shown as red solid line) and reference (actual, shown as blue dashed line) time series from the selection-switch experiment with ordinary filtering at mean relative abundance 10^{-3}. The time series are shown in PCoA ordinations using Bray-Curtis similarity. The ordinations are based on all time series of each subdivision.

Chapter 5

Discussion

Reviewing our results, we want to answer our research questions:

1. How are microbial communities structured?
2. Do our results correspond to real ecological interactions or are there confounding factors?
3. What is the most sound way of inferring ecological interactions?
4. Do microbial communities follow a specific path based on external selection pressure or are dynamics dominated by stochastic effects?
5. Can dynamics in microbial communities be described, explained, reproduced and predicted?

Finally, we will discuss how further work can be done in order to provide better answers to the questions.

5.1 Interactions of microbial communities

If we were to interpret our results naively, our brief answer to Question 1 could be: "Microbial communities are dominated by positive interactions within clusters of related bacteria, with negative interactions between the clusters." However, correlation does not imply causality, as pointed out by Fisher and Mehta[18]: Closely related bacteria may have similar environmental preferences (niches) and would for that reason appear in high abundance in the same samples. Indeed, Dorman *et al.*[4] warn that phylogenetic signals may give results resembling those of biotic interactions. Competitive exclusion could eventually remove OTUs with the same niche[60], but in our case it is likely that these processes occur at a far longer

timescales than observed. Alternatively, the differences in fitness between OTUs with the same niche might be too small to be of any importance, according to Hubbel's neutral theory[61, 62].

The counterintuitive diagnostic plots also provide a good reason to take the results with a pinch of salt. Furthermore, the dominance of positive interactions is contrary to literature as Coyte *et al.*[63] tell that a microbial community should be dominated by negative interactions in order to be stable. In addition, Foster and Bell[64] claim that negative interactions dominate seawater communities based on studies of co-cultures. We hereby conclude that our results do not correspond to causal interactions, but are merely artifacts of OTUs having the same environmental preferences, thus answering Question 2. According to the same line, a wiser answer to Question 1 could be: "The co-occurence patterns are dominated by the interaction clusters of OTUs, but the nature of the real interaction structure remains unknown." While the co-occurance pattern by itself can be useful to study, we must acknowledge what it tells us and what it does not.

Based on our previous discussion, we have no other clear answer to Question 3 other than saying that *our* approach is not sound to infer ecological interactions. A good method of inferring ecological interactions has to distinguish the co-occurrence patterns from the underlying causal relationships. In literature ([1, 16, 26, 31]), it is reported that compositional data may easily create spurious correlations. This is the reason why methods (included ReBoot), aiming to mitigate this challenge, have been developed. However, from our own results we have seen that replacing relative abundances with absolute ones did only have minor effects on the disappointing results. Therefore, microbial ecologists should be warned that distinguishing the co-occurrence patterns from the real ecological interactions is still a challenge even if the problem of compositionality is resolved.

Are there then better cross-sectional algorithms than ReBoot to find ecological interactions? According to Weiss *et al.*[16], the ReBoot approach *does not* perform particularly well compared to its competitors. In any case, no inference tool studied by Weiss *et al.*[16] did perform particularly well on realistic data, suggesting the need for even better methods. Hirano and Takemoto[21] argue that tools designed to deal with compositional data, such as sparCC[17] and LSA[65], do not perform any better than classical Pearson and Spearman correlation. From our own results, we have seen that the refinements of the ReBoot approach did not provide more structured PCoA plots than direct application of the similarity measures. In our view, the more sophisticated methods of finding ecological interactions may work well on some simulated datasets[16, 31], but not perform any better than the simplest methods when faced with datasets having another structure[21].

5.2 Dynamics of microbial communities

We realize that if the true answer to Question 4 is: "Microbial communities are shaped entirely by stochastic effects", then there is no hope to predict how a microbial community will develop. However, we found clear patterns of determinism of the dynamics in our data, even though some stochastic effects were present. Studies in literature have come of contradictory conclusions regarding the importance of stochastic effects[66]. Hence, it seems more reasonable to view the degree of determinism as dependent of the community in question. Faust *et al.*[67] and Dini-Andreote *et al.*[68] recognize this and provide computational frameworks to measure and characterize the random effects.

Considering Question 5, we have gained a satisfying description of the dynamics, considering the patterns in the time trajectories. However, understanding why the time series develop the way they do, is a harder question. Some qualitative explanations might be useful, but do not provide the full answer: Convergence to almost the same state under the same selection pressure can be explained by the fact that some bacteria are better to compete than others under the specific selective pressure. Also, the observation that the K-selected reactors in the selection-switch experiment showed more uniform and deterministic behavior than the r-selected time series is no surprise as K-selected communties are believed to be more stable[9, 12]. The latter observation may also imply that K-selected communities are easier to control as applying the necessary external environment will most likely make the community converge towards the desired state.

A deeper understanding of what happens in the microbial communities is yet beyond reach. We hoped that the Lotka-Volterra modeling should give us deeper insight into how microbial communities evolve over time. However, this procedure gave no sensible results. Therefore, considering Question 5, we did not succeed in reproducing and predicting the dynamics of the microbial communities. As commented in Section 4.3.4, the failure is likely due to low time resolution. We acknowledge that Stein *et al.*[19] present predictions being qualitatively close to the reference, even though the time between sampling is longer (one sample per day) than we would recommend from our own simulations. Reasonable success of predicting time series was also reported by Kloppers and Greeff[50][1]. However, Stein *et al.*[19] and Kloppers and Greeff[50] considered communities consisting of far fewer entities than in our case. Hence, the high number of OTUs might have been a more important cause for our failure than the time between sampling points. Indeed, Bucci *et al.*[69] recommend that a study modeling N OTUs, should have at least $\frac{N^2}{2}$ data points. For the selection-switch experiment with filtering of OTUs at 10^{-4} mean abundance, this could require 13 613 samples, 68 times the actual number. Based on our discussion, we suggest that any lab-reactor experiment

[1]Note that this paper considered marked shares of companies and was in no way concerned with microbial communities

aiming to predict the Lotka-Volterra coefficients, should have more samples taken closer in time, maybe once per hour. In order to realize this, novel methods of automatic sampling must be developed and used.

5.3 Suggestions for further work

In our opinion, methods based on time series (such as the gLV equation) have a larger potential to provide further insight into microbial processes than cross-sectional methods such as ReBoot. Fisher and Mehta[18] and Bucci *et al.*[69] in particular support this view. The reasons why we think this is the case are:

- Having the ability to predict time series provides an objective measure of how good the algorithm is. For cross-sectional methods, there is more difficult to find a good criterion for goodness of fit.
- Co-occurrence patterns may be caused by entirely different processes than the interactions between microbes. Time series based methods may be hampered by this challenge too. However, we think still this will be less of a problem for time series as the external environment is usually more or less the same between consecutive samples, whereas in cross-sectional analysis, all samples pooled together are (usually) treated equally.
- Time series data are less prone to be affected by indirect interactions as delayed effects are visible if the time resolution is small enough.

Continuing on a gLV-based approach is likely to require refinements. We have already discussed the need for higher sampling rates, but we will also probably need more sophisticated ways of finding the coefficients. Our current approach makes an all-to-all model of the interactions. However, there might be likely that the interaction pattern might be explained by a more sparse interaction network, in which case the results could be more accurate, robust to noise and easier to interpret. LIMITS[18] (Learning Interactions from MIcrobial Time Series) uses an approach similar to the log-integral method, but the way it solves the linear system is different. Instead of solving the system by least squares, it uses sparse linear regression. Among the same lines, MDSINE (Microbial Dynamical Systems INference Engine) extends the work of Stein *et al.*[19], adding three more regularization methods:

- Maximum likelihood constrained ridge regression (MLCRR). This works as the ridge regularization used in this thesis, but adds the *a priori* constraint that all maximal growth rates are positive and self-interaction coefficients are negative.
- Bayesian adaptive lasso (BAL). This is a special ℓ^1 regularization method based on a Bayesian approach.

- Bayesian variable selection (BVS). Here, the interactions are picked directly in a sort of variable selection.

According to the paper, utilizing these refined algorithms, especially those based on Bayesian approaches should give more precise predictions than the original method presented by Stein *et al.*[19]. Neural network models have been used on microbial datasets[70], but to our knowledge, such methods have so far only been used to predict phenotypes of samples based on the microbial profile. Predicting the dynamics of a microbiome by a neural network model would be interesting to try as neural network algorithms work in completely different way than the gLV approach. Recurrent neural networks are commonly used to reproduce and classify temporal dynamics, so we suggest giving it a try in further work.

In this thesis, we have considered the OTUs to be discrete entities, but remember that they are distinguished at 97% similarity of the 16S rDNA marker sequence, which may look like an arbitrary criterion. Therefore, it might be more correct to think of a microbial community as a phylogenetic continuum rather than a composition of discrete taxonomic units. Our findings support this view as we found the co-occurrence patterns to match fairly well with the phylogenetic identity. From this viewpoint, we can use the phylogeny as a predictor for the interactions and instead treat each individual read as the smallest unit. Hierarchical Modelling of Species Communities (HMSC)[71], being a Joint Species Distribution Model(JSDM), takes into account the assumption that phylogenetically related species should have similar traits and phenotypes. Moreover, this procedure can easily incorporate external ecological data and even tell which factors are the most important for the abundances of each species(variance partitioning). We believe that such a procedure could make the predictions more robust and accurate as information is re-used across the OTU boundaries. Bjork *et al.*[72] has indeed tried out a JSMD approach for analyzing microbial datasets, but the analysis were more focused on the effect of the host, instead of the interactions between the microbes themselves. However, adopting the existing HMSC framework to suit our needs to predict microbial time series, is likely to require some effort.

Chapter 6

Conclusion and outlook

The datasets studied in this thesis clearly showed certain interesting structures, but our efforts to determine the underlying interactions largely failed. The ReBoot approach did give a good insight into the co-occurrence patterns. This allowed us for instance to separate the OTUs dominating in r- and K-selected environments. However, phylogenetically closely related OTUs often appear together in clusters with strong internal positive associations. We do not think this pattern correspond to the real ecological interactions between the bacteria. Rather, we think they appear together because they have the same niche and therefore are present in the same samples, in which the community is shaped by external variables. Even though the similarity measures provided had different characteristics, most of them seem to agree on major aspects of the co-occurrence patterns. Among the same lines, random noise and transforming absolute abundances to relative ones did make detectable differences, but the main structures of the clustered association networks were still retained.

Also, the temporal development of the time series showed interesting dynamics, especially the K-selected reactors in the selection-switch dataset. Yet again, our attempts of inferring Lotka-Volterra coefficients and predicting the system did not succeed. While we know that the algorithm works on small artificial datasets with rapid sampling rates, using the same approach gave no reasonable results on our own real datasets. We attribute this disappointing discovery to the high number of OTUs and poor time resolution.

Whereas some tools have turned out to provide valuable insight into the interactions betweens microbes in specific studies, none of the current methods has proved to provide reliable and accurate results for a large variety of simulated and real world data. We believe that the most promising approaches for detecting microbial interactions depend on microbial time series. In this case, it will require good datasets with frequent sampling.

If we would come to the point where we understood how microorganisms in-

teract among themselves and with each other, and know how we as humans could shape and control these interactions, the rewards would be large. Aquaculture industry could provide a stable K-selected community, excluding pathogens and opportunists. Doctors could more easily treat and prevent diseases related to malfunctioning gut microbiota, included, but not restricted to, infections, allergy and obesity. Process industry could have communities of microorganisms faithfully carrying out the desired tasks in perfect concert, while being cost-efficient and environmental friendly. However, understanding the functioning and dynamics of microbial communities is still in its infancy and it will probably require much more research to get a proper understanding of how microbes interact.

Bibliography

1. Faust, K. & Raes, J. Microbial interactions: from networks to models. *Nat. Rev. Microbiol.* **10,** 538–550. ISSN: 1740-1526 (Aug. 2012).
2. Tringe, S. G. *et al.* Comparative metagenomics of microbial communities. *Science* **308,** 554–557. ISSN: 0036-8075 (Apr. 2005).
3. Layeghifard, M., Hwang, D. M. & Guttman, D. S. Disentangling Interactions in the Microbiome: A Network Perspective. *Trends Microbiol.* **25,** 217–228. ISSN: 0966-842X (Mar. 2017).
4. Dormann, C. F. *et al.* Biotic interactions in species distribution modelling: 10 questions to guide interpretation and avoid false conclusions. *Glob. Ecol. Biogeogr.* **27,** 1004–1016. ISSN: 1466-822X (Sept. 2018).
5. Turnbaugh, P. J. *et al.* The Human Microbiome Project. *Nature* **449,** 804–810. ISSN: 0028-0836 (Oct. 2007).
6. McIlroy, J., Ianiro, G., Mukhopadhya, I., Hansen, R. & Hold, G. L. Review article: the gut microbiome in inflammatory bowel diseaseavenues for microbial management. *Aliment. Pharmacol. Ther.* **47,** 26–42. ISSN: 0269-2813 (Jan. 2018).
7. Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457,** 480–U7. ISSN: 0028-0836 (Jan. 2009).
8. Pascal, M. *et al.* Microbiome and Allergic Diseases. *Front. Immunol.* **9.** ISSN: 1664-3224 (July 2018).
9. Vadstein, O., Attramadal, K. J. K., Bakke, I. & Olsen, Y. K-Selection as Microbial Community Management Strategy: A Method for Improved Viability of Larvae in Aquaculture. *Front. Microbiol.* **9.** ISSN: 1664-302X (Nov. 2018).
10. Vadstein, O. *et al. A Strategy To Obtain Microbial Control During Larval Development Of Marine Fish* English. in *Fish Farming Technology* (ed Reinertsen, H and Dahle, LA and Jorgensen, L and Tvinnereim, K) 1St International Conf On Fish Farming Technology, Trondheim, Norway, Aug 09-12, 1993 (A A Balkema, Rotterdam, 1993), 69–75. ISBN: 90-5410-326-4.

11. Andrews, J. H. & Harris, R. F. r-selection and k-selection and microbial ecology. *Adv. Microbial Ecol.* **9,** 99–147. ISSN: 0147-4863 (1986).

12. Gundersen, M. S. *The effect of r- and K-selection on planktonic microbial community characteristics in lab-scale bioreactors* MA thesis (NTNU Norwegian University of Science and Technology, May 2019). `http://hdl.handle.net/11250/2621728`.

13. Woese, C. R. & Fox, G. E. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci U S A* **74,** 5088–90 (Nov. 1977).

14. Kennedy, K., Hall, M. W., Lynch, M. D. J., Moreno-Hagelsieb, G. & Neufeld, J. D. Evaluating Bias of Illumina-Based Bacterial 16S rRNA Gene Profiles. *Appl. Environ. Microbiol.* **80,** 5717–5722. ISSN: 0099-2240 (Sept. 2014).

15. Nguyen, N.-P., Warnow, T., Pop, M. & White, B. A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity. *npj Biofilms Microbiomes* **2.** ISSN: 2055-5008 (2016).

16. Weiss, S. *et al.* Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J.* **10,** 1669–1681. ISSN: 1751-7362 (July 2016).

17. Friedman, J. & Alm, E. J. Inferring Correlation Networks from Genomic Survey Data. *PLOS Computational Biology* **8,** 1–11. `https://doi.org/10.1371/journal.pcbi.1002687` (Sept. 2012).

18. Fisher, C. K. & Mehta, P. Identifying Keystone Species in the Human Gut Microbiome from Metagenomic Timeseries Using Sparse Linear Regression. *PLoS One* **9.** ISSN: 1932-6203 (July 2014).

19. Stein, R. R. *et al.* Ecological Modeling from Time-Series Inference: Insight into Dynamics and Stability of Intestinal Microbiota. *PLoS Comput. Biol.* **9.** ISSN: 1553-7358 (desember 2013).

20. Buffie, C. G. *et al.* Precision microbiome reconstitution restores bile acid mediated resistance to Clostridium difficile. *Nature* **517,** 205–U207. ISSN: 0028-0836 (Jan. 2015).

21. Hirano, H. & Takemoto, K. Difficulty in inferring microbial community structure based on co-occurrence network approaches. *BMC Bioinformatics* **20.** ISSN: 1471-2105 (June 2019).

22. Faust, K. *et al.* Microbial Co-occurrence Relationships in the Human Microbiome. *PLoS Comput. Biol.* **8.** ISSN: 1553-734X (July 2012).

23. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26,** 2460–2461. ISSN: 1367-4803 (Oct. 2010).

24. Cole, J. R. *et al.* Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* **42,** D633–D642. ISSN: 0305-1048 (Jan. 2014).

25. Solberg, E. H. *Capacity of a flow-through biofilter system to secure K-selection and microbial stability in the out-flowing water* MA thesis (NTNU Norwegian University of Science and Technology, Sept. 2018). http://hdl.handle.net/11250/2615549.

26. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome Datasets Are Compositional: And This Is Not Optional. *Front. Microbiol.* **8.** ISSN: 1664-302X (Nov. 2017).

27. Polz, M. F. & Cavanaugh, C. M. Bias in template-to-product ratios in multitemplate PCR. *Appl. Environ. Microbiol.* **64,** 3724–3730. ISSN: 0099-2240 (Oct. 1998).

28. Case, R. J. *et al.* Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. *Appl. Environ. Microbiol.* **73,** 278–288. ISSN: 0099-2240 (Jan. 2007).

29. Crosby, L. D. & Criddle, C. S. Understanding bias in microbial community analysis techniques due to rrn operon copy number heterogeneity. *Biotechniques* **34,** 790–+. ISSN: 0736-6205 (Apr. 2003).

30. Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C. & Knight, R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27,** 2194–2200. ISSN: 1367-4803 (Aug. 2011).

31. Berry, D. & Widder, S. Deciphering microbial interactions and detecting keystone species with co-occurrence networks. *Front. Microbiol.* **5.** ISSN: 1664-302X (mai 2014).

32. James, G., Witten, D., Hastie, T. & Tibshirani, R. *Unsupervised Learning* in *An Introduction to Statistical Learning: with Applications in R* 373–418 (Springer New York, New York, NY, 2013). ISBN: 978-1-4614-7138-7. https://doi.org/10.1007/978-1-4614-7138-7_10.

33. Hauke, J. & Kossowski, T. Comparison of Values of Pearson's and Spearman's Correlation Coefficients on the Same Sets of Data. *Quaestiones Geographicae* **30.** https://content.sciendo.com/view/journals/quageo/30/2/article-p87.xml (2011).

34. *Correlation Coefficient* in *The Concise Encyclopedia of Statistics* 115–119 (Springer New York, New York, NY, 2008). ISBN: 978-0-387-32833-1. https://doi.org/10.1007/978-0-387-32833-1_83.

35. *Spearman Rank Correlation Coefficient* in *The Concise Encyclopedia of Statistics* 502–505 (Springer New York, New York, NY, 2008). ISBN: 978-0-387-32833-1. https://doi.org/10.1007/978-0-387-32833-1_379.

36. *Kendall Rank Correlation Coefficient* in *The Concise Encyclopedia of Statistics* 278–281 (Springer New York, New York, NY, 2008). ISBN: 978-0-387-32833-1. https://doi.org/10.1007/978-0-387-32833-1_211.

37. Oksanen, J. *et al. vegan: Community Ecology Package* R package version 2.4-6 (2018). https://CRAN.R-project.org/package=vegan.

38. Greenacre, M. & Primicerio, R. *Chapter 5 Measures of distance between samples: non-Euclidean* in *Multivariate Analysis of Ecological Data* 61–73 (Fundación BBVA, Plaza de San Nicolás, 4. 48005 Bilbao, 2013). ISBN: 978-84-92937-50-9. https://www.fbbva.es/wp-content/uploads/2017/05/dat/DE_2013_multivariate.pdf.

39. Cha, S.-H. Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions. *International journal of mathematical models and methods in applied sciences* **1,** 300–307 (2017).

40. Schwager, E., Bielski, C. & Weingart, G. *ccrepe: ccrepe_and_nc.score* R package version 1.14.0 (2014). http://bioconductor.org/packages/release/bioc/html/ccrepe.html.

41. Greenacre, M. & Primicerio, R. *Chapter 4 Measures of distance between samples: Euclidean* in *Multivariate Analysis of Ecological Data* 47–59 (Fundación BBVA, Plaza de San Nicolás, 4. 48005 Bilbao, 2013). ISBN: 978-84-92937-50-9. https://www.fbbva.es/wp-content/uploads/2017/05/dat/DE_2013_multivariate.pdf.

42. Cover, T. M. & Thomas, J. A. *Introduction and Preview* in *Elements of Information Theory* 1–12 (John Wiley & Sons, Inc., 2005). ISBN: 9780471748823. http://dx.doi.org/10.1002/047174882X.ch1.

43. Meyer, P. E. *infotheo: Information-Theoretic Measures* R package version 1.2.0 (2014). https://CRAN.R-project.org/package=infotheo.

44. Sidorov, G., Gelbukh, A., Gómez-Adorno, H. & Pinto, D. Soft Similarity and Soft Cosine Measure: Similarity of Features in Vector Space Model. *Computación y Sistemas* **18,** 491–504. http://www.cys.cic.ipn.mx/ojs/index.php/CyS/article/view/2043 (2014).

45. *Bootstrap* in *The Concise Encyclopedia of Statistics* 51–54 (Springer New York, New York, NY, 2008). ISBN: 978-0-387-32833-1. https://doi.org/10.1007/978-0-387-32833-1_40.

46. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **29,** 1165–1188. ISSN: 0090-5364 (Aug. 2001).

47. Barabási, A.-L. *et al. Network science* (Cambridge university press, 2016).

48. Pons, P. *et al.* Computing communities in large networks using random walks. *LECT NOTES COMPUT SC* **3733,** 284–293. ISSN: 0302-9743 (2005).

49. Lotka, A. J. Analytical Note on Certain Rhythmic Relations in Organic Systems. *Proceedings of the National Academy of Sciences* **6,** 410–415. ISSN: 0027-8424. eprint: http://www.pnas.org/content/6/7/410.full.pdf. http://www.pnas.org/content/6/7/410 (1920).

50. Kloppers, P. H. & Greeff, J. C. Lotka-Volterra model parameter estimation using experiential data. *Appl. Math. Comput.* **224,** 817–825. ISSN: 0096-3003 (Nov. 2013).

51. Hastie, T., Tibshirani, R. & Friedman, J. *Linear Methods for Regression* in *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* 43–99 (Springer New York, New York, NY, 2009). ISBN: 978-0-387-84858-7. https://doi.org/10.1007/978-0-387-84858-7_3.

52. Cazelles, K., Araujo, M. B., Mouquet, N. & Gravel, D. A theory for species co-occurrence in interaction networks. *Theor. Ecol.* **9,** 39–48. ISSN: 1874-1738 (Mar. 2016).

53. Faust, K. & Raes, J. CoNet app: inference of biological association networks using Cytoscape. *F1000Res* **5,** 1519 (2016).

54. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal* **Complex Systems,** 1695. http://igraph.org (2006).

55. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20,** 289–290. ISSN: 1367-4803 (Jan. 2004).

56. Hastie, T., Tibshirani, R. & Friedman, J. *Model Assessment and Selection* in *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* 219–259 (Springer New York, New York, NY, 2009). ISBN: 978-0-387-84858-7. https://doi.org/10.1007/978-0-387-84858-7_7.

57. Soetaert, K., Petzoldt, T. & Setzer, R. W. Solving Differential Equations in R: Package deSolve. *J. Stat. Softw.* **33,** 1–25. ISSN: 1548-7660 (Feb. 2010).

58. Croux, C. & Dehon, C. Robustness versus Efficiency for Nonparametric Correlation Measures. eng. *IDEAS Working Paper Series from RePEc.* http://search.proquest.com/docview/1698253432/ (2008).

59. Gibson, B., Wilson, D. J., Feil, E. & Eyre-Walker, A. The distribution of bacterial doubling times in the wild. *Proc Biol Sci* **285** (June 2018).

60. Hardin, G. The competitive exclusion principle. *Science* **131,** 1292–7 (Apr. 1960).

61. Hubbell, S. P. *The unified neutral theory of biodiversity and biogeography* eng. ISBN: 0691021295 (Princeton University Press, Princeton, N.J, 2001).

62. Rosindell, J., Hubbell, S. P. & Etienne, R. S. The Unified Neutral Theory of Biodiversity and Biogeography at Age Ten. *Trends Ecol. Evol.* **26,** 340–348. ISSN: 0169-5347 (July 2011).

63. Coyte, K. Z., Schluter, J. & Foster, K. R. The ecology of the microbiome: Networks, competition, and stability. *Science* **350,** 663–666. ISSN: 0036-8075 (Nov. 2015).

64. Foster, K. R. & Bell, T. Competition, Not Cooperation, Dominates Interactions among Culturable Microbial Species. *Curr. Biol.* **22,** 1845–1850. ISSN: 0960-9822 (Oct. 2012).

65. Ruan, Q. *et al.* Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors. *Bioinformatics* **22,** 2532–2538. ISSN: 1367-4803 (Oct. 2006).

66. Faust, K., Lahti, L., Gonze, D., de Vos, W. M. & Raes, J. Metagenomics meets time series analysis: unraveling microbial community dynamics. *Curr. Opin. Microbiol.* **25,** 56–66. ISSN: 1369-5274 (June 2015).

67. Faust, K. *et al.* Signatures of ecological processes in microbial community time series. *Microbiome* **6.** ISSN: 2049-2618 (June 2018).

68. Dini-Andreote, F., Stegen, J. C., van Elsas, J. D. & Salles, J. F. Disentangling mechanisms that mediate the balance between stochastic and deterministic processes in microbial succession. *Proc. Natl. Acad. Sci. U. S. A.* **112,** E1326–E1332. ISSN: 0027-8424 (Mar. 2015).

69. Bucci, V. *et al.* MDSINE: Microbial Dynamical Systems INference Engine for microbiome time-series analyses. *Genome Biol.* **17.** ISSN: 1474-760X (June 2016).

70. Reiman, D., Metwally, A. & Dai, Y. Using convolutional neural networks to explore the microbiome. *Conf Proc IEEE Eng Med Biol Soc* **2017,** 4269–4272 (July 2017).

71. Ovaskainen, O. *et al.* How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecology Letters* **20,** 561–576. `https://onlinelibrary.wiley.com/doi/abs/10.1111/ele.12757` (Mar. 2017).

72. Bjork, J. R., Hui, F. K. C., O'Hara, R. B. & Montoya, J. M. Uncovering the drivers of host-associated microbiota with joint species distribution modelling. *Mol. Ecol.* **27,** 2714–2724. ISSN: 0962-1083 (June 2018).

Appendix A

Yields of the different similarity measures

Here, we present the outcomes for the ReBoot procedure for the different similarity measures as in Section 4.2.1. For the relative data from the selection-switch experiment, the results are in Tables A.1 to A.5. For the selection-switch experiment, the similar results for absolute data are shown in Tables 4.1 and A.6 to A.9. The results from the seawater experiment are shown in Table A.10, while the results from the biofilm experiment are shown in Tables A.11 to A.16.

Table A.1: Performance of the different similarity measures using relative data from the K_H subset of the selection-switch experiment

Name of similarity measure	Signed	Type of similarity measure	Number of significant interactions	Proportion of negative interactions
bray_curtis	FALSE	parametric	103	
bray_curtis_normal	FALSE	parametric	70	
bray_curtis_scaled	FALSE	parametric	0	
bray_curtis_scaled_normal	FALSE	parametric	0	
bray_curtis_scaled_uniform	FALSE	parametric	1	
bray_curtis_uniform	FALSE	parametric	74	
cosine	TRUE	parametric	249	0.00
cosine_normal	TRUE	parametric	204	0.00
cosine_uniform	TRUE	parametric	199	0.00
generalized_jaccard_index	FALSE	parametric	64	
generalized_jaccard_index_normal	FALSE	parametric	51	
generalized_jaccard_index_scaled	FALSE	parametric	0	
generalized_jaccard_index_scaled_normal	FALSE	parametric	0	
generalized_jaccard_index_scaled_uniform	FALSE	parametric	1	
generalized_jaccard_index_uniform	FALSE	parametric	54	
jaccard_index	FALSE	presence-absence	66	
kendall	TRUE	non-parametric	1421	0.38
kendall_normal	TRUE	non-parametric	203	0.29
kendall_uniform	TRUE	non-parametric	197	0.28
mutual_information	FALSE	non-parametric	121	
mutual_information_normal	FALSE	non-parametric	13	
mutual_information_uniform	FALSE	non-parametric	16	
nc_score	TRUE	non-parametric	1238	0.35
nc_score_normal	TRUE	non-parametric	170	0.26
nc_score_uniform	TRUE	non-parametric	165	0.25
pearson	TRUE	parametric	236	0.02
pearson_normal	TRUE	parametric	167	0.01
pearson_uniform	TRUE	parametric	165	0.01
spearman	TRUE	non-parametric	1384	0.37
spearman_normal	TRUE	non-parametric	251	0.29
spearman_uniform	TRUE	non-parametric	243	0.28
squared_euclidean	FALSE	parametric	0	
squared_euclidean_normal	FALSE	parametric	0	
squared_euclidean_scaled	FALSE	parametric	0	
squared_euclidean_scaled_normal	FALSE	parametric	0	
squared_euclidean_scaled_uniform	FALSE	parametric	0	
squared_euclidean_uniform	FALSE	parametric	0	

Table A.2: Performance of the different similarity measures using relative data from the K_L subset of the selection-switch experiment

Name of similarity measure	Signed	Type of similarity measure	Number of significant interactions	Proportion of negative interactions
bray_curtis	FALSE	parametric	175	
bray_curtis_normal	FALSE	parametric	142	
bray_curtis_scaled	FALSE	parametric	0	
bray_curtis_scaled_normal	FALSE	parametric	0	
bray_curtis_scaled_uniform	FALSE	parametric	0	
bray_curtis_uniform	FALSE	parametric	139	
cosine	TRUE	parametric	447	0.00
cosine_normal	TRUE	parametric	339	0.00
cosine_uniform	TRUE	parametric	323	0.00
generalized_jaccard_index	FALSE	parametric	152	
generalized_jaccard_index_normal	FALSE	parametric	121	
generalized_jaccard_index_scaled	FALSE	parametric	0	
generalized_jaccard_index_scaled_normal	FALSE	parametric	0	
generalized_jaccard_index_scaled_uniform	FALSE	parametric	0	
generalized_jaccard_index_uniform	FALSE	parametric	127	
jaccard_index	FALSE	presence-absence	76	
kendall	TRUE	non-parametric	1143	0.29
kendall_normal	TRUE	non-parametric	173	0.23
kendall_uniform	TRUE	non-parametric	181	0.24
mutual_information	FALSE	non-parametric	175	
mutual_information_normal	FALSE	non-parametric	35	
mutual_information_uniform	FALSE	non-parametric	32	
nc_score	TRUE	non-parametric	1042	0.29
nc_score_normal	TRUE	non-parametric	161	0.24
nc_score_uniform	TRUE	non-parametric	147	0.23
pearson	TRUE	parametric	440	0.03
pearson_normal	TRUE	parametric	278	0.03
pearson_uniform	TRUE	parametric	274	0.03
spearman	TRUE	non-parametric	1189	0.30
spearman_normal	TRUE	non-parametric	221	0.27
spearman_uniform	TRUE	non-parametric	226	0.28
squared_euclidean	FALSE	parametric	0	
squared_euclidean_normal	FALSE	parametric	2	
squared_euclidean_scaled	FALSE	parametric	0	
squared_euclidean_scaled_normal	FALSE	parametric	0	
squared_euclidean_scaled_uniform	FALSE	parametric	0	
squared_euclidean_uniform	FALSE	parametric	2	

Table A.3: Performance of the different similarity measures on the overall relative data from the selection switch experiment

Name of similarity measure	Signed	Type of similarity measure	Number of significant interactions	Proportion of negative interactions
bray_curtis	FALSE	parametric	525	
bray_curtis_normal	FALSE	parametric	445	
bray_curtis_scaled	FALSE	parametric	0	
bray_curtis_scaled_normal	FALSE	parametric	0	
bray_curtis_scaled_uniform	FALSE	parametric	0	
bray_curtis_uniform	FALSE	parametric	438	
cosine	TRUE	parametric	363	0.00
cosine_normal	TRUE	parametric	328	0.00
cosine_uniform	TRUE	parametric	340	0.00
generalized_jaccard_index	FALSE	parametric	460	
generalized_jaccard_index_normal	FALSE	parametric	381	
generalized_jaccard_index_scaled	FALSE	parametric	0	
generalized_jaccard_index_scaled_normal	FALSE	parametric	0	
generalized_jaccard_index_scaled_uniform	FALSE	parametric	0	
generalized_jaccard_index_uniform	FALSE	parametric	392	
jaccard_index	FALSE	presence-absence	870	
kendall	TRUE	non-parametric	3272	0.40
kendall_normal	TRUE	non-parametric	1398	0.41
kendall_uniform	TRUE	non-parametric	1394	0.41
mutual_information	FALSE	non-parametric	2610	
mutual_information_normal	FALSE	non-parametric	849	
mutual_information_uniform	FALSE	non-parametric	836	
nc_score	TRUE	non-parametric	3251	0.40
nc_score_normal	TRUE	non-parametric	1374	0.40
nc_score_uniform	TRUE	non-parametric	1396	0.41
pearson	TRUE	parametric	466	0.40
pearson_normal	TRUE	parametric	399	0.42
pearson_uniform	TRUE	parametric	418	0.41
spearman	TRUE	non-parametric	3269	0.40
spearman_normal	TRUE	non-parametric	1417	0.40
spearman_uniform	TRUE	non-parametric	1421	0.41
squared_euclidean	FALSE	parametric	3	
squared_euclidean_normal	FALSE	parametric	3	
squared_euclidean_scaled	FALSE	parametric	0	
squared_euclidean_scaled_normal	FALSE	parametric	0	
squared_euclidean_scaled_uniform	FALSE	parametric	0	
squared_euclidean_uniform	FALSE	parametric	3	

Table A.4: Performance of the different similarity measures using relative data from the r_H subset of the selection-switch experiment

Name of similarity measure	Signed	Type of similarity measure	Number of significant interactions	Proportion of negative interactions
bray_curtis	FALSE	parametric	17	
bray_curtis_normal	FALSE	parametric	13	
bray_curtis_scaled	FALSE	parametric	0	
bray_curtis_scaled_normal	FALSE	parametric	0	
bray_curtis_scaled_uniform	FALSE	parametric	0	
bray_curtis_uniform	FALSE	parametric	14	
cosine	TRUE	parametric	98	0.00
cosine_normal	TRUE	parametric	102	0.00
cosine_uniform	TRUE	parametric	94	0.00
generalized_jaccard_index	FALSE	parametric	10	
generalized_jaccard_index_normal	FALSE	parametric	13	
generalized_jaccard_index_scaled	FALSE	parametric	0	
generalized_jaccard_index_scaled_normal	FALSE	parametric	0	
generalized_jaccard_index_scaled_uniform	FALSE	parametric	0	
generalized_jaccard_index_uniform	FALSE	parametric	12	
jaccard_index	FALSE	presence-absence	121	
kendall	TRUE	non-parametric	628	0.19
kendall_normal	TRUE	non-parametric	56	0.07
kendall_uniform	TRUE	non-parametric	55	0.07
mutual_information	FALSE	non-parametric	63	
mutual_information_normal	FALSE	non-parametric	7	
mutual_information_uniform	FALSE	non-parametric	4	
nc_score	TRUE	non-parametric	546	0.16
nc_score_normal	TRUE	non-parametric	42	0.05
nc_score_uniform	TRUE	non-parametric	43	0.05
pearson	TRUE	parametric	79	0.00
pearson_normal	TRUE	parametric	50	0.00
pearson_uniform	TRUE	parametric	48	0.00
spearman	TRUE	non-parametric	615	0.20
spearman_normal	TRUE	non-parametric	58	0.05
spearman_uniform	TRUE	non-parametric	63	0.08
squared_euclidean	FALSE	parametric	0	
squared_euclidean_normal	FALSE	parametric	0	
squared_euclidean_scaled	FALSE	parametric	0	
squared_euclidean_scaled_normal	FALSE	parametric	0	
squared_euclidean_scaled_uniform	FALSE	parametric	0	
squared_euclidean_uniform	FALSE	parametric	0	

Table A.5: Performance of the different similarity measures using relative data from the r_L subset of the selection-switch experiment

Name of similarity measure	Signed	Type of similarity measure	Number of significant interactions	Proportion of negative interactions
bray_curtis	FALSE	parametric	40	
bray_curtis_normal	FALSE	parametric	31	
bray_curtis_scaled	FALSE	parametric	0	
bray_curtis_scaled_normal	FALSE	parametric	0	
bray_curtis_scaled_uniform	FALSE	parametric	0	
bray_curtis_uniform	FALSE	parametric	30	
cosine	TRUE	parametric	177	0.00
cosine_normal	TRUE	parametric	143	0.00
cosine_uniform	TRUE	parametric	152	0.00
generalized_jaccard_index	FALSE	parametric	27	
generalized_jaccard_index_normal	FALSE	parametric	26	
generalized_jaccard_index_scaled	FALSE	parametric	0	
generalized_jaccard_index_scaled_normal	FALSE	parametric	0	
generalized_jaccard_index_scaled_uniform	FALSE	parametric	0	
generalized_jaccard_index_uniform	FALSE	parametric	26	
jaccard_index	FALSE	presence-absence	45	
kendall	TRUE	non-parametric	693	0.34
kendall_normal	TRUE	non-parametric	96	0.22
kendall_uniform	TRUE	non-parametric	94	0.19
mutual_information	FALSE	non-parametric	44	
mutual_information_normal	FALSE	non-parametric	6	
mutual_information_uniform	FALSE	non-parametric	6	
nc_score	TRUE	non-parametric	580	0.32
nc_score_normal	TRUE	non-parametric	81	0.20
nc_score_uniform	TRUE	non-parametric	79	0.18
pearson	TRUE	parametric	158	0.09
pearson_normal	TRUE	parametric	98	0.07
pearson_uniform	TRUE	parametric	99	0.07
spearman	TRUE	non-parametric	700	0.35
spearman_normal	TRUE	non-parametric	105	0.20
spearman_uniform	TRUE	non-parametric	108	0.21
squared_euclidean	FALSE	parametric	0	
squared_euclidean_normal	FALSE	parametric	0	
squared_euclidean_scaled	FALSE	parametric	0	
squared_euclidean_scaled_normal	FALSE	parametric	0	
squared_euclidean_scaled_uniform	FALSE	parametric	0	
squared_euclidean_uniform	FALSE	parametric	0	

Table A.6: Performance of the different similarity measures using absolute data from the K_H subset of the selection-switch experiment

Name of similarity measure	Signed	Type of similarity measure	Number of significant interactions	Proportion of negative interactions
bray_curtis	FALSE	parametric	162	
bray_curtis_normal	FALSE	parametric	131	
bray_curtis_scaled	FALSE	parametric	0	
bray_curtis_scaled_normal	FALSE	parametric	0	
bray_curtis_scaled_uniform	FALSE	parametric	0	
bray_curtis_uniform	FALSE	parametric	129	
cosine	TRUE	parametric	340	0.00
cosine_normal	TRUE	parametric	356	0.00
cosine_uniform	TRUE	parametric	349	0.00
generalized_jaccard_index	FALSE	parametric	125	
generalized_jaccard_index_normal	FALSE	parametric	104	
generalized_jaccard_index_scaled	FALSE	parametric	0	
generalized_jaccard_index_scaled_normal	FALSE	parametric	0	
generalized_jaccard_index_scaled_uniform	FALSE	parametric	0	
generalized_jaccard_index_uniform	FALSE	parametric	104	
jaccard_index	FALSE	presence-absence	64	
kendall	TRUE	non-parametric	1268	0.26
kendall_normal	TRUE	non-parametric	594	0.11
kendall_uniform	TRUE	non-parametric	607	0.11
mutual_information	FALSE	non-parametric	160	
mutual_information_normal	FALSE	non-parametric	93	
mutual_information_uniform	FALSE	non-parametric	89	
nc_score	TRUE	non-parametric	1149	0.24
nc_score_normal	TRUE	non-parametric	558	0.10
nc_score_uniform	TRUE	non-parametric	551	0.10
pearson	TRUE	parametric	319	0.00
pearson_normal	TRUE	parametric	316	0.00
pearson_uniform	TRUE	parametric	312	0.00
spearman	TRUE	non-parametric	1236	0.25
spearman_normal	TRUE	non-parametric	649	0.14
spearman_uniform	TRUE	non-parametric	637	0.13
squared_euclidean	FALSE	parametric	0	
squared_euclidean_normal	FALSE	parametric	0	
squared_euclidean_scaled	FALSE	parametric	0	
squared_euclidean_scaled_normal	FALSE	parametric	0	
squared_euclidean_scaled_uniform	FALSE	parametric	0	
squared_euclidean_uniform	FALSE	parametric	0	

Table A.7: Performance of the different similarity measures using absolute data from the K_L subset of the selection-switch experiment

Name of similarity measure	Signed	Type of similarity measure	Number of significant interactions	Proportion of negative interactions
bray_curtis	FALSE	parametric	193	
bray_curtis_normal	FALSE	parametric	139	
bray_curtis_scaled	FALSE	parametric	0	
bray_curtis_scaled_normal	FALSE	parametric	0	
bray_curtis_scaled_uniform	FALSE	parametric	0	
bray_curtis_uniform	FALSE	parametric	157	
cosine	TRUE	parametric	434	0.00
cosine_normal	TRUE	parametric	382	0.00
cosine_uniform	TRUE	parametric	380	0.00
generalized_jaccard_index	FALSE	parametric	135	
generalized_jaccard_index_normal	FALSE	parametric	112	
generalized_jaccard_index_scaled	FALSE	parametric	0	
generalized_jaccard_index_scaled_normal	FALSE	parametric	0	
generalized_jaccard_index_scaled_uniform	FALSE	parametric	0	
generalized_jaccard_index_uniform	FALSE	parametric	113	
jaccard_index	FALSE	presence-absence	75	
kendall	TRUE	non-parametric	1139	0.35
kendall_normal	TRUE	non-parametric	251	0.27
kendall_uniform	TRUE	non-parametric	269	0.28
mutual_information	FALSE	non-parametric	156	
mutual_information_normal	FALSE	non-parametric	26	
mutual_information_uniform	FALSE	non-parametric	29	
nc_score	TRUE	non-parametric	1006	0.33
nc_score_normal	TRUE	non-parametric	221	0.24
nc_score_uniform	TRUE	non-parametric	210	0.25
pearson	TRUE	parametric	403	0.05
pearson_normal	TRUE	parametric	330	0.06
pearson_uniform	TRUE	parametric	315	0.05
spearman	TRUE	non-parametric	1069	0.35
spearman_normal	TRUE	non-parametric	302	0.28
spearman_uniform	TRUE	non-parametric	295	0.30
squared_euclidean	FALSE	parametric	4	
squared_euclidean_normal	FALSE	parametric	4	
squared_euclidean_scaled	FALSE	parametric	0	
squared_euclidean_scaled_normal	FALSE	parametric	0	
squared_euclidean_scaled_uniform	FALSE	parametric	0	
squared_euclidean_uniform	FALSE	parametric	4	

Table A.8: Performance of the different similarity measures using absolute data from the r_H subset of the selection-switch experiment

Name of similarity measure	Signed	Type of similarity measure	Number of significant interactions	Proportion of negative interactions
bray_curtis	FALSE	parametric	34	
bray_curtis_normal	FALSE	parametric	22	
bray_curtis_scaled	FALSE	parametric	0	
bray_curtis_scaled_normal	FALSE	parametric	0	
bray_curtis_scaled_uniform	FALSE	parametric	0	
bray_curtis_uniform	FALSE	parametric	25	
cosine	TRUE	parametric	144	0.00
cosine_normal	TRUE	parametric	152	0.00
cosine_uniform	TRUE	parametric	148	0.00
generalized_jaccard_index	FALSE	parametric	18	
generalized_jaccard_index_normal	FALSE	parametric	18	
generalized_jaccard_index_scaled	FALSE	parametric	0	
generalized_jaccard_index_scaled_normal	FALSE	parametric	0	
generalized_jaccard_index_scaled_uniform	FALSE	parametric	0	
generalized_jaccard_index_uniform	FALSE	parametric	18	
jaccard_index	FALSE	presence-absence	146	
kendall	TRUE	non-parametric	637	0.18
kendall_normal	TRUE	non-parametric	109	0.04
kendall_uniform	TRUE	non-parametric	104	0.04
mutual_information	FALSE	non-parametric	60	
mutual_information_normal	FALSE	non-parametric	8	
mutual_information_uniform	FALSE	non-parametric	10	
nc_score	TRUE	non-parametric	545	0.17
nc_score_normal	TRUE	non-parametric	86	0.03
nc_score_uniform	TRUE	non-parametric	89	0.03
pearson	TRUE	parametric	132	0.00
pearson_normal	TRUE	parametric	114	0.00
pearson_uniform	TRUE	parametric	109	0.00
spearman	TRUE	non-parametric	631	0.17
spearman_normal	TRUE	non-parametric	109	0.04
spearman_uniform	TRUE	non-parametric	113	0.04
squared_euclidean	FALSE	parametric	0	
squared_euclidean_normal	FALSE	parametric	0	
squared_euclidean_scaled	FALSE	parametric	0	
squared_euclidean_scaled_normal	FALSE	parametric	0	
squared_euclidean_scaled_uniform	FALSE	parametric	0	
squared_euclidean_uniform	FALSE	parametric	0	

Table A.9: Performance of the different similarity measures using absolute data from the r_L subset of the selection-switch experiment

Name of similarity measure	Signed	Type of similarity measure	Number of significant interactions	Proportion of negative interactions
bray_curtis	FALSE	parametric	59	
bray_curtis_normal	FALSE	parametric	35	
bray_curtis_scaled	FALSE	parametric	0	
bray_curtis_scaled_normal	FALSE	parametric	0	
bray_curtis_scaled_uniform	FALSE	parametric	0	
bray_curtis_uniform	FALSE	parametric	37	
cosine	TRUE	parametric	243	0.00
cosine_normal	TRUE	parametric	227	0.00
cosine_uniform	TRUE	parametric	236	0.00
generalized_jaccard_index	FALSE	parametric	45	
generalized_jaccard_index_normal	FALSE	parametric	29	
generalized_jaccard_index_scaled	FALSE	parametric	0	
generalized_jaccard_index_scaled_normal	FALSE	parametric	0	
generalized_jaccard_index_scaled_uniform	FALSE	parametric	0	
generalized_jaccard_index_uniform	FALSE	parametric	27	
jaccard_index	FALSE	presence-absence	48	
kendall	TRUE	non-parametric	620	0.30
kendall_normal	TRUE	non-parametric	154	0.16
kendall_uniform	TRUE	non-parametric	158	0.16
mutual_information	FALSE	non-parametric	61	
mutual_information_normal	FALSE	non-parametric	20	
mutual_information_uniform	FALSE	non-parametric	19	
nc_score	TRUE	non-parametric	504	0.28
nc_score_normal	TRUE	non-parametric	136	0.14
nc_score_uniform	TRUE	non-parametric	141	0.15
pearson	TRUE	parametric	214	0.02
pearson_normal	TRUE	parametric	165	0.01
pearson_uniform	TRUE	parametric	165	0.00
spearman	TRUE	non-parametric	621	0.30
spearman_normal	TRUE	non-parametric	158	0.16
spearman_uniform	TRUE	non-parametric	164	0.17
squared_euclidean	FALSE	parametric	0	
squared_euclidean_normal	FALSE	parametric	0	
squared_euclidean_scaled	FALSE	parametric	0	
squared_euclidean_scaled_normal	FALSE	parametric	0	
squared_euclidean_scaled_uniform	FALSE	parametric	0	
squared_euclidean_uniform	FALSE	parametric	0	

Table A.10: Performance of the different similarity measures using data from the seawater experiment

Name of similarity measure	Signed	Type of similarity measure	Number of significant interactions	Proportion of negative interactions
bray_curtis	FALSE	parametric	109	
bray_curtis_normal	FALSE	parametric	76	
bray_curtis_scaled	FALSE	parametric	0	
bray_curtis_scaled_normal	FALSE	parametric	0	
bray_curtis_scaled_uniform	FALSE	parametric	0	
bray_curtis_uniform	FALSE	parametric	69	
cosine	TRUE	parametric	592	0.00
cosine_normal	TRUE	parametric	515	0.00
cosine_uniform	TRUE	parametric	478	0.00
generalized_jaccard_index	FALSE	parametric	96	
generalized_jaccard_index_normal	FALSE	parametric	78	
generalized_jaccard_index_scaled	FALSE	parametric	0	
generalized_jaccard_index_scaled_normal	FALSE	parametric	0	
generalized_jaccard_index_scaled_uniform	FALSE	parametric	0	
generalized_jaccard_index_uniform	FALSE	parametric	87	
jaccard_index	FALSE	presence-absence	888	
kendall	TRUE	non-parametric	4722	0.36
kendall_normal	TRUE	non-parametric	10	0.00
kendall_uniform	TRUE	non-parametric	10	0.10
mutual_information	FALSE	non-parametric	49	
mutual_information_normal	FALSE	non-parametric	0	
mutual_information_uniform	FALSE	non-parametric	0	
nc_score	TRUE	non-parametric	3727	0.35
nc_score_normal	TRUE	non-parametric	2	0.00
nc_score_uniform	TRUE	non-parametric	2	0.00
pearson	TRUE	parametric	520	0.01
pearson_normal	TRUE	parametric	277	0.00
pearson_uniform	TRUE	parametric	260	0.00
spearman	TRUE	non-parametric	5410	0.35
spearman_normal	TRUE	non-parametric	3	0.00
spearman_uniform	TRUE	non-parametric	4	0.00
squared_euclidean	FALSE	parametric	0	
squared_euclidean_normal	FALSE	parametric	0	
squared_euclidean_scaled	FALSE	parametric	0	
squared_euclidean_scaled_normal	FALSE	parametric	0	
squared_euclidean_scaled_uniform	FALSE	parametric	0	
squared_euclidean_uniform	FALSE	parametric	0	

Table A.11: Performance of the different similarity measures using data from the C subset of the biofilm experiment

Name of similarity measure	Signed	Type of similarity measure	Number of significant interactions	Proportion of negative interactions
bray_curtis	FALSE	parametric	72	
bray_curtis_normal	FALSE	parametric	63	
bray_curtis_scaled	FALSE	parametric	0	
bray_curtis_scaled_normal	FALSE	parametric	0	
bray_curtis_scaled_uniform	FALSE	parametric	0	
bray_curtis_uniform	FALSE	parametric	65	
cosine	TRUE	parametric	251	0.00
cosine_normal	TRUE	parametric	242	0.00
cosine_uniform	TRUE	parametric	245	0.00
generalized_jaccard_index	FALSE	parametric	49	
generalized_jaccard_index_normal	FALSE	parametric	45	
generalized_jaccard_index_scaled	FALSE	parametric	0	
generalized_jaccard_index_scaled_normal	FALSE	parametric	0	
generalized_jaccard_index_scaled_uniform	FALSE	parametric	0	
generalized_jaccard_index_uniform	FALSE	parametric	46	
jaccard_index	FALSE	presence-absence	127	
kendall	TRUE	non-parametric	1358	0.34
kendall_normal	TRUE	non-parametric	86	0.23
kendall_uniform	TRUE	non-parametric	97	0.23
mutual_information	FALSE	non-parametric	171	
mutual_information_normal	FALSE	non-parametric	10	
mutual_information_uniform	FALSE	non-parametric	11	
nc_score	TRUE	non-parametric	1229	0.33
nc_score_normal	TRUE	non-parametric	83	0.25
nc_score_uniform	TRUE	non-parametric	84	0.25
pearson	TRUE	parametric	238	0.11
pearson_normal	TRUE	parametric	169	0.07
pearson_uniform	TRUE	parametric	174	0.08
spearman	TRUE	non-parametric	1268	0.33
spearman_normal	TRUE	non-parametric	126	0.25
spearman_uniform	TRUE	non-parametric	127	0.25
squared_euclidean	FALSE	parametric	0	
squared_euclidean_normal	FALSE	parametric	0	
squared_euclidean_scaled	FALSE	parametric	0	
squared_euclidean_scaled_normal	FALSE	parametric	0	
squared_euclidean_scaled_uniform	FALSE	parametric	0	
squared_euclidean_uniform	FALSE	parametric	0	

Table A.12: Performance of the different similarity measures using data from the TR1 subset of the biofilm experiment

Name of similarity measure	Signed	Type of similarity measure	Number of significant interactions	Proportion of negative interactions
bray_curtis	FALSE	parametric	50	
bray_curtis_normal	FALSE	parametric	48	
bray_curtis_scaled	FALSE	parametric	0	
bray_curtis_scaled_normal	FALSE	parametric	0	
bray_curtis_scaled_uniform	FALSE	parametric	0	
bray_curtis_uniform	FALSE	parametric	47	
cosine	TRUE	parametric	197	0.00
cosine_normal	TRUE	parametric	190	0.00
cosine_uniform	TRUE	parametric	188	0.00
generalized_jaccard_index	FALSE	parametric	43	
generalized_jaccard_index_normal	FALSE	parametric	39	
generalized_jaccard_index_scaled	FALSE	parametric	0	
generalized_jaccard_index_scaled_normal	FALSE	parametric	0	
generalized_jaccard_index_scaled_uniform	FALSE	parametric	0	
generalized_jaccard_index_uniform	FALSE	parametric	39	
jaccard_index	FALSE	presence-absence	238	
kendall	TRUE	non-parametric	1216	0.30
kendall_normal	TRUE	non-parametric	168	0.13
kendall_uniform	TRUE	non-parametric	171	0.14
mutual_information	FALSE	non-parametric	272	
mutual_information_normal	FALSE	non-parametric	26	
mutual_information_uniform	FALSE	non-parametric	26	
nc_score	TRUE	non-parametric	1091	0.29
nc_score_normal	TRUE	non-parametric	134	0.13
nc_score_uniform	TRUE	non-parametric	136	0.13
pearson	TRUE	parametric	155	0.02
pearson_normal	TRUE	parametric	126	0.02
pearson_uniform	TRUE	parametric	129	0.02
spearman	TRUE	non-parametric	1257	0.30
spearman_normal	TRUE	non-parametric	201	0.15
spearman_uniform	TRUE	non-parametric	203	0.15
squared_euclidean	FALSE	parametric	0	
squared_euclidean_normal	FALSE	parametric	0	
squared_euclidean_scaled	FALSE	parametric	0	
squared_euclidean_scaled_normal	FALSE	parametric	0	
squared_euclidean_scaled_uniform	FALSE	parametric	0	
squared_euclidean_uniform	FALSE	parametric	0	

Table A.13: Performance of the different similarity measures using data from the TR2 subset of the biofilm experiment

Name of similarity measure	Signed	Type of similarity measure	Number of significant interactions	Proportion of negative interactions
bray_curtis	FALSE	parametric	42	
bray_curtis_normal	FALSE	parametric	54	
bray_curtis_scaled	FALSE	parametric	0	
bray_curtis_scaled_normal	FALSE	parametric	1	
bray_curtis_scaled_uniform	FALSE	parametric	0	
bray_curtis_uniform	FALSE	parametric	50	
cosine	TRUE	parametric	155	0.00
cosine_normal	TRUE	parametric	161	0.00
cosine_uniform	TRUE	parametric	162	0.00
generalized_jaccard_index	FALSE	parametric	27	
generalized_jaccard_index_normal	FALSE	parametric	41	
generalized_jaccard_index_scaled	FALSE	parametric	0	
generalized_jaccard_index_scaled_normal	FALSE	parametric	0	
generalized_jaccard_index_scaled_uniform	FALSE	parametric	0	
generalized_jaccard_index_uniform	FALSE	parametric	38	
jaccard_index	FALSE	presence-absence	180	
kendall	TRUE	non-parametric	841	0.25
kendall_normal	TRUE	non-parametric	84	0.06
kendall_uniform	TRUE	non-parametric	88	0.06
mutual_information	FALSE	non-parametric	161	
mutual_information_normal	FALSE	non-parametric	10	
mutual_information_uniform	FALSE	non-parametric	7	
nc_score	TRUE	non-parametric	730	0.24
nc_score_normal	TRUE	non-parametric	79	0.05
nc_score_uniform	TRUE	non-parametric	76	0.05
pearson	TRUE	parametric	127	0.00
pearson_normal	TRUE	parametric	113	0.00
pearson_uniform	TRUE	parametric	111	0.00
spearman	TRUE	non-parametric	816	0.24
spearman_normal	TRUE	non-parametric	95	0.07
spearman_uniform	TRUE	non-parametric	100	0.07
squared_euclidean	FALSE	parametric	0	
squared_euclidean_normal	FALSE	parametric	0	
squared_euclidean_scaled	FALSE	parametric	0	
squared_euclidean_scaled_normal	FALSE	parametric	0	
squared_euclidean_scaled_uniform	FALSE	parametric	0	
squared_euclidean_uniform	FALSE	parametric	0	

Table A.14: Performance of the different similarity measures using data from the TR3 subset of the biofilm experiment

Name of similarity measure	Signed	Type of similarity measure	Number of significant interactions	Proportion of negative interactions
bray_curtis	FALSE	parametric	21	
bray_curtis_normal	FALSE	parametric	15	
bray_curtis_scaled	FALSE	parametric	0	
bray_curtis_scaled_normal	FALSE	parametric	0	
bray_curtis_scaled_uniform	FALSE	parametric	0	
bray_curtis_uniform	FALSE	parametric	12	
cosine	TRUE	parametric	170	0.00
cosine_normal	TRUE	parametric	153	0.00
cosine_uniform	TRUE	parametric	141	0.00
generalized_jaccard_index	FALSE	parametric	23	
generalized_jaccard_index_normal	FALSE	parametric	22	
generalized_jaccard_index_scaled	FALSE	parametric	0	
generalized_jaccard_index_scaled_normal	FALSE	parametric	0	
generalized_jaccard_index_scaled_uniform	FALSE	parametric	0	
generalized_jaccard_index_uniform	FALSE	parametric	21	
jaccard_index	FALSE	presence-absence	80	
kendall	TRUE	non-parametric	305	0.05
kendall_normal	TRUE	non-parametric	40	0.07
kendall_uniform	TRUE	non-parametric	43	0.07
mutual_information	FALSE	non-parametric	0	
mutual_information_normal	FALSE	non-parametric	0	
mutual_information_uniform	FALSE	non-parametric	0	
nc_score	TRUE	non-parametric	193	0.02
nc_score_normal	TRUE	non-parametric	7	0.00
nc_score_uniform	TRUE	non-parametric	7	0.00
pearson	TRUE	parametric	128	0.06
pearson_normal	TRUE	parametric	83	0.08
pearson_uniform	TRUE	parametric	90	0.07
spearman	TRUE	non-parametric	308	0.05
spearman_normal	TRUE	non-parametric	51	0.06
spearman_uniform	TRUE	non-parametric	49	0.06
squared_euclidean	FALSE	parametric	0	
squared_euclidean_normal	FALSE	parametric	0	
squared_euclidean_scaled	FALSE	parametric	0	
squared_euclidean_scaled_normal	FALSE	parametric	0	
squared_euclidean_scaled_uniform	FALSE	parametric	0	
squared_euclidean_uniform	FALSE	parametric	0	

Table A.15: Performance of the different similarity measures using data from the W subset of the biofilm experiment

Name of similarity measure	Signed	Type of similarity measure	Number of significant interactions	Proportion of negative interactions
bray_curtis	FALSE	parametric	58	
bray_curtis_normal	FALSE	parametric	34	
bray_curtis_scaled	FALSE	parametric	0	
bray_curtis_scaled_normal	FALSE	parametric	0	
bray_curtis_scaled_uniform	FALSE	parametric	0	
bray_curtis_uniform	FALSE	parametric	35	
cosine	TRUE	parametric	149	0.00
cosine_normal	TRUE	parametric	143	0.00
cosine_uniform	TRUE	parametric	149	0.00
generalized_jaccard_index	FALSE	parametric	35	
generalized_jaccard_index_normal	FALSE	parametric	28	
generalized_jaccard_index_scaled	FALSE	parametric	0	
generalized_jaccard_index_scaled_normal	FALSE	parametric	0	
generalized_jaccard_index_scaled_uniform	FALSE	parametric	0	
generalized_jaccard_index_uniform	FALSE	parametric	27	
jaccard_index	FALSE	presence-absence	304	
kendall	TRUE	non-parametric	1129	0.14
kendall_normal	TRUE	non-parametric	289	0.14
kendall_uniform	TRUE	non-parametric	313	0.14
mutual_information	FALSE	non-parametric	235	
mutual_information_normal	FALSE	non-parametric	47	
mutual_information_uniform	FALSE	non-parametric	60	
nc_score	TRUE	non-parametric	1021	0.13
nc_score_normal	TRUE	non-parametric	280	0.14
nc_score_uniform	TRUE	non-parametric	273	0.14
pearson	TRUE	parametric	128	0.07
pearson_normal	TRUE	parametric	100	0.05
pearson_uniform	TRUE	parametric	107	0.06
spearman	TRUE	non-parametric	1151	0.14
spearman_normal	TRUE	non-parametric	333	0.14
spearman_uniform	TRUE	non-parametric	327	0.15
squared_euclidean	FALSE	parametric	0	
squared_euclidean_normal	FALSE	parametric	0	
squared_euclidean_scaled	FALSE	parametric	0	
squared_euclidean_scaled_normal	FALSE	parametric	0	
squared_euclidean_scaled_uniform	FALSE	parametric	0	
squared_euclidean_uniform	FALSE	parametric	0	

Table A.16: Performance of the different similarity measures on the overall data from the biofilm experiment

Name of similarity measure	Signed	Type of similarity measure	Number of significant interactions	Proportion of negative interactions
bray_curtis	FALSE	parametric	139	
bray_curtis_normal	FALSE	parametric	120	
bray_curtis_scaled	FALSE	parametric	0	
bray_curtis_scaled_normal	FALSE	parametric	0	
bray_curtis_scaled_uniform	FALSE	parametric	0	
bray_curtis_uniform	FALSE	parametric	114	
cosine	TRUE	parametric	256	0.00
cosine_normal	TRUE	parametric	263	0.00
cosine_uniform	TRUE	parametric	255	0.00
generalized_jaccard_index	FALSE	parametric	100	
generalized_jaccard_index_normal	FALSE	parametric	93	
generalized_jaccard_index_scaled	FALSE	parametric	0	
generalized_jaccard_index_scaled_normal	FALSE	parametric	0	
generalized_jaccard_index_scaled_uniform	FALSE	parametric	0	
generalized_jaccard_index_uniform	FALSE	parametric	92	
jaccard_index	FALSE	presence-absence	609	
kendall	TRUE	non-parametric	2217	0.25
kendall_normal	TRUE	non-parametric	638	0.20
kendall_uniform	TRUE	non-parametric	668	0.20
mutual_information	FALSE	non-parametric	1269	
mutual_information_normal	FALSE	non-parametric	157	
mutual_information_uniform	FALSE	non-parametric	158	
nc_score	TRUE	non-parametric	2138	0.24
nc_score_normal	TRUE	non-parametric	619	0.20
nc_score_uniform	TRUE	non-parametric	597	0.19
pearson	TRUE	parametric	256	0.11
pearson_normal	TRUE	parametric	227	0.10
pearson_uniform	TRUE	parametric	239	0.10
spearman	TRUE	non-parametric	2222	0.25
spearman_normal	TRUE	non-parametric	681	0.20
spearman_uniform	TRUE	non-parametric	688	0.20
squared_euclidean	FALSE	parametric	1	
squared_euclidean_normal	FALSE	parametric	0	
squared_euclidean_scaled	FALSE	parametric	0	
squared_euclidean_scaled_normal	FALSE	parametric	0	
squared_euclidean_scaled_uniform	FALSE	parametric	0	
squared_euclidean_uniform	FALSE	parametric	1	

Appendix B

Diagnostic plots

This is a continuation of Section 4.2.2, where the remaining diagnostic plots are shown, see Table B.1 for the references to the tables.

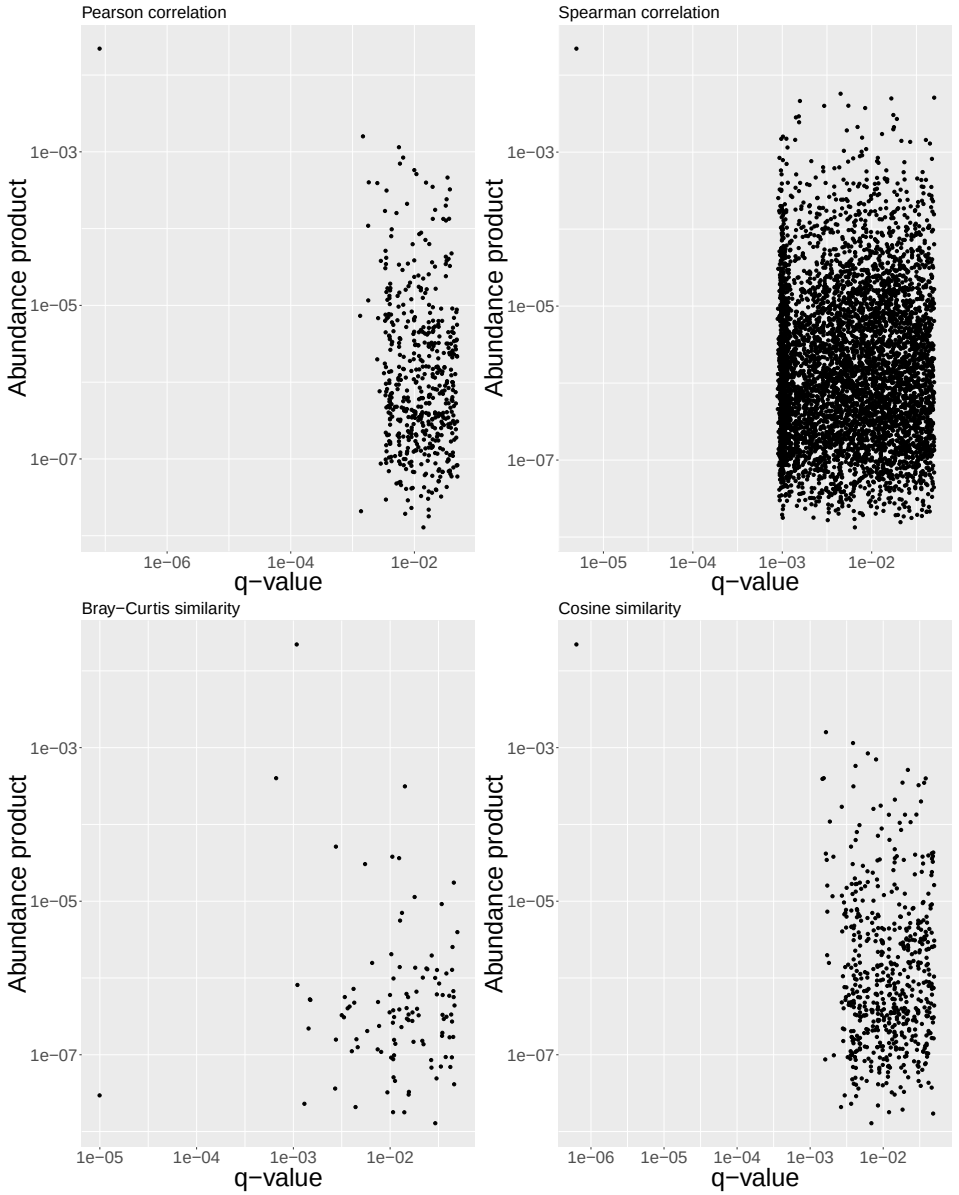


Figure B.1: Abundance product versus q-values for each significant interaction in the seawater experiment

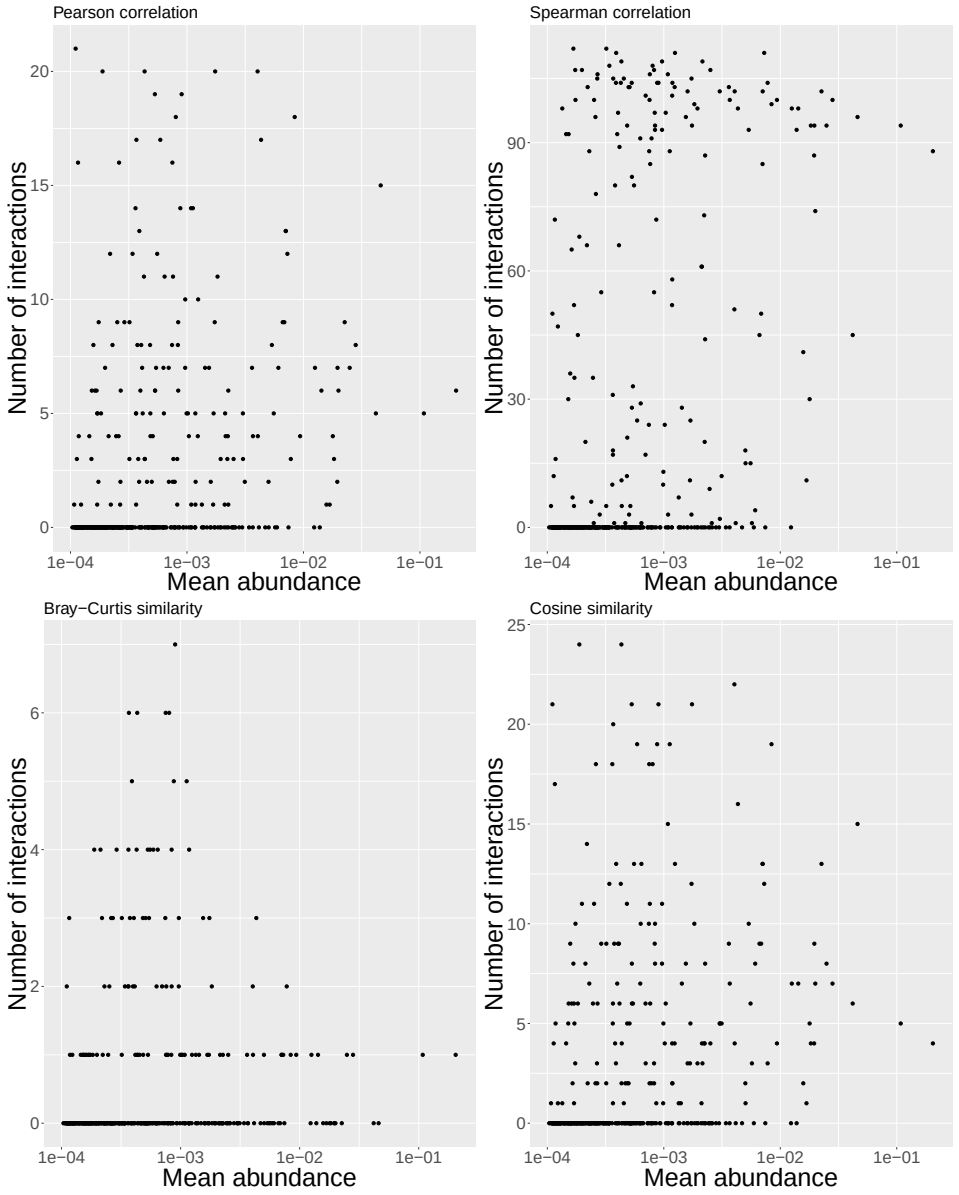


Figure B.2: Number of significant interactions versus overall mean abundance for each OTU in the seawater experiment

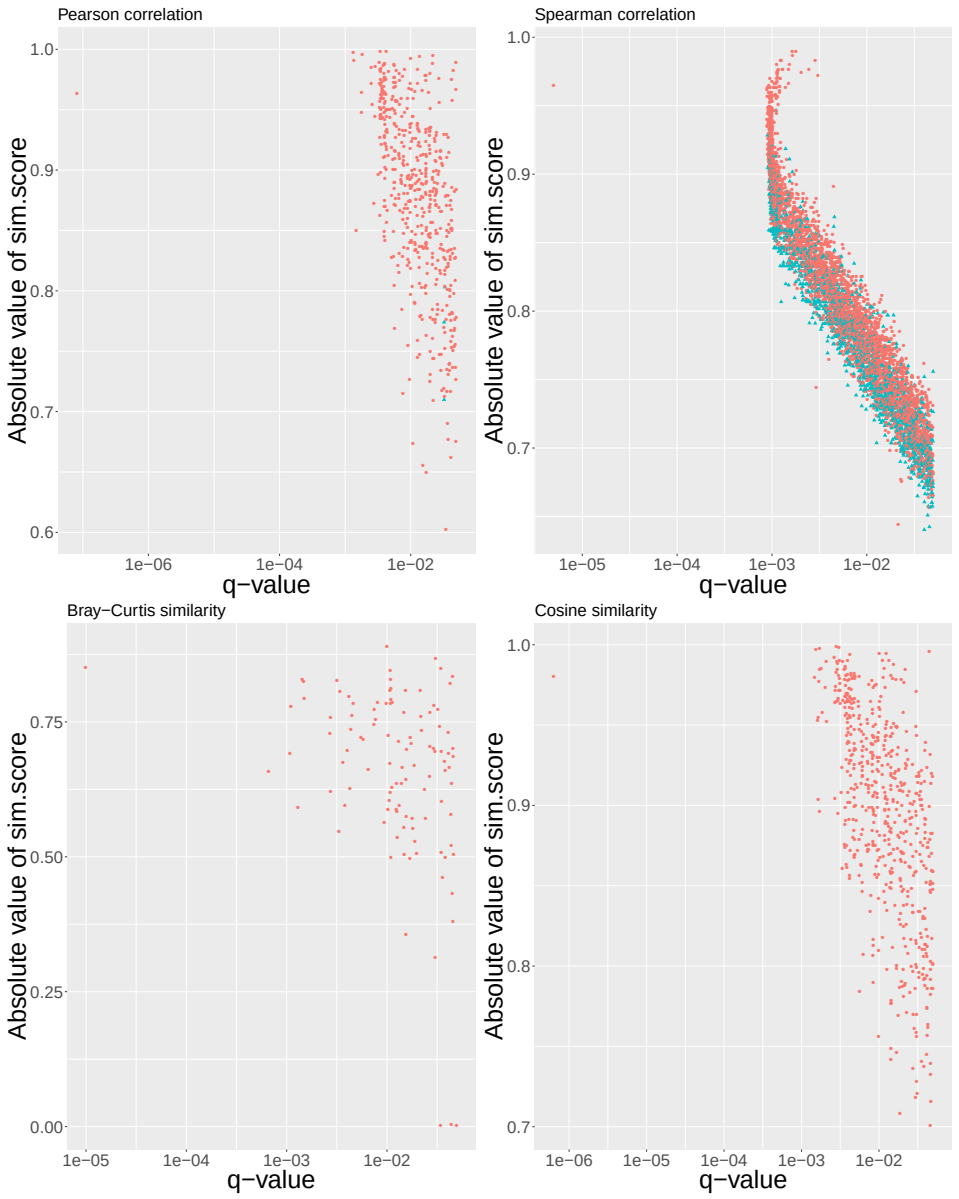


Figure B.3: Similarity scores versus q-values for each significant interaction in the seawater experiment. Red circles indicate positive interactions, whereas blue triangles indicate negative interactions.

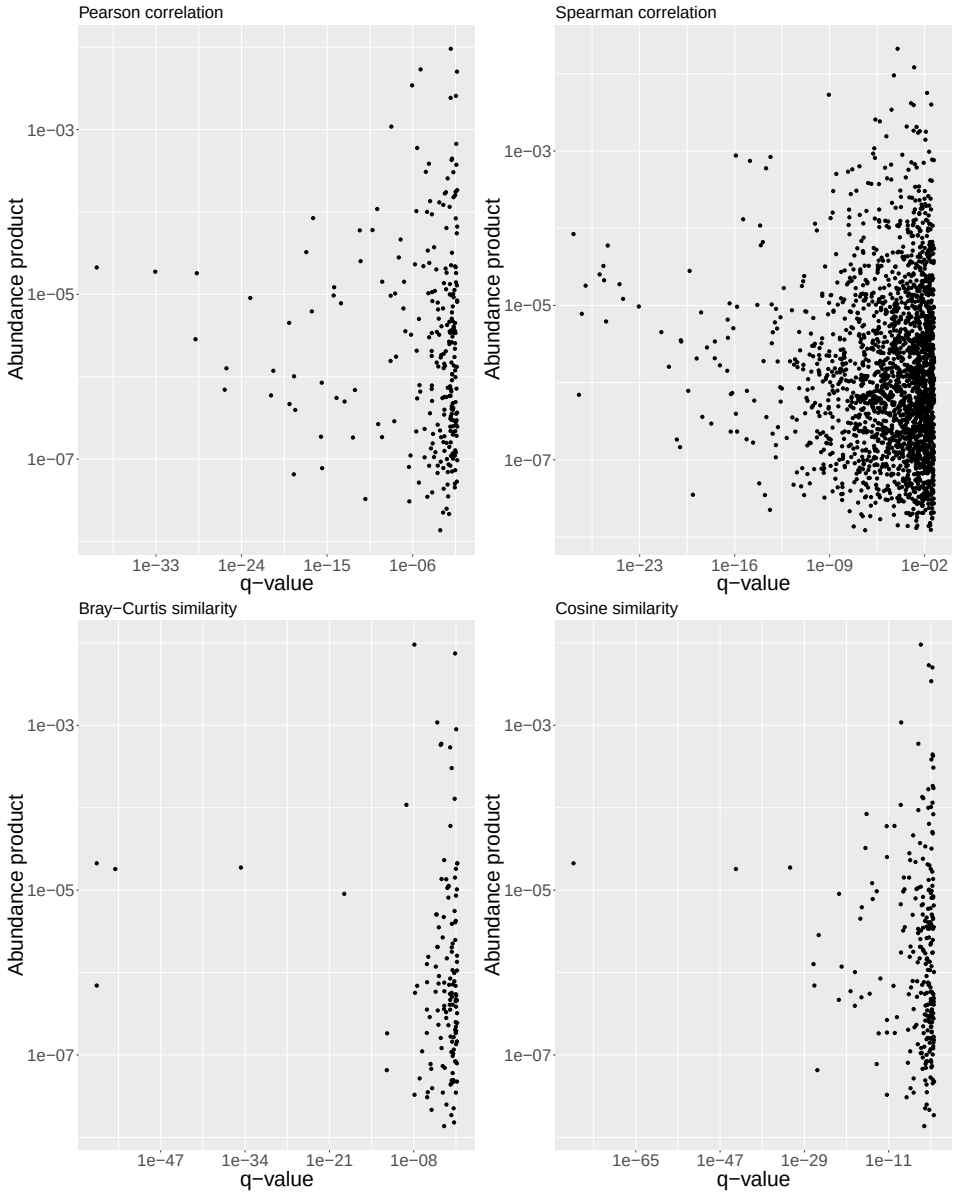


Figure B.4: Abundance product versus q-values for each significant interaction in the biofilm experiment

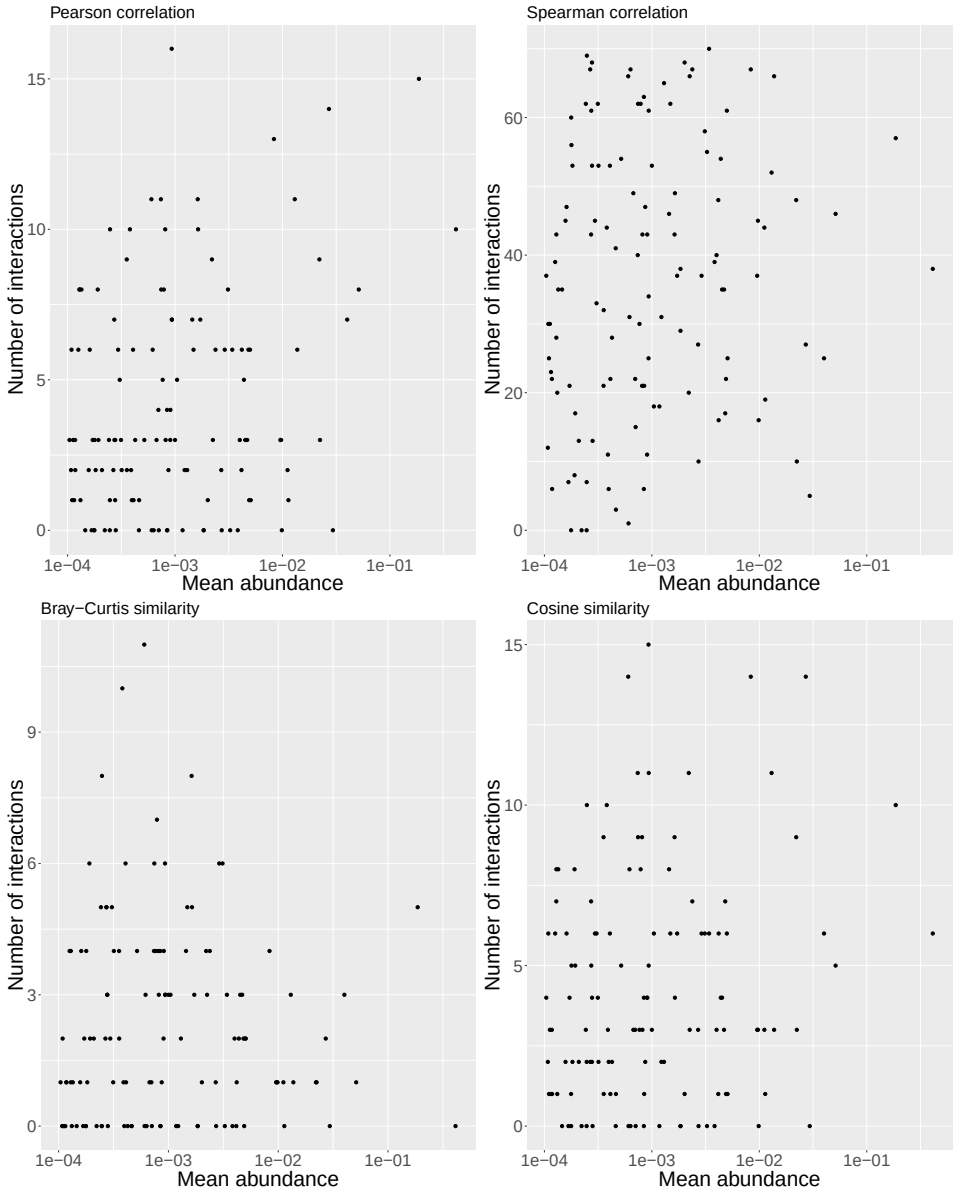


Figure B.5: Number of significant interactions versus overall mean abundance for each OTU in the biofilm experiment

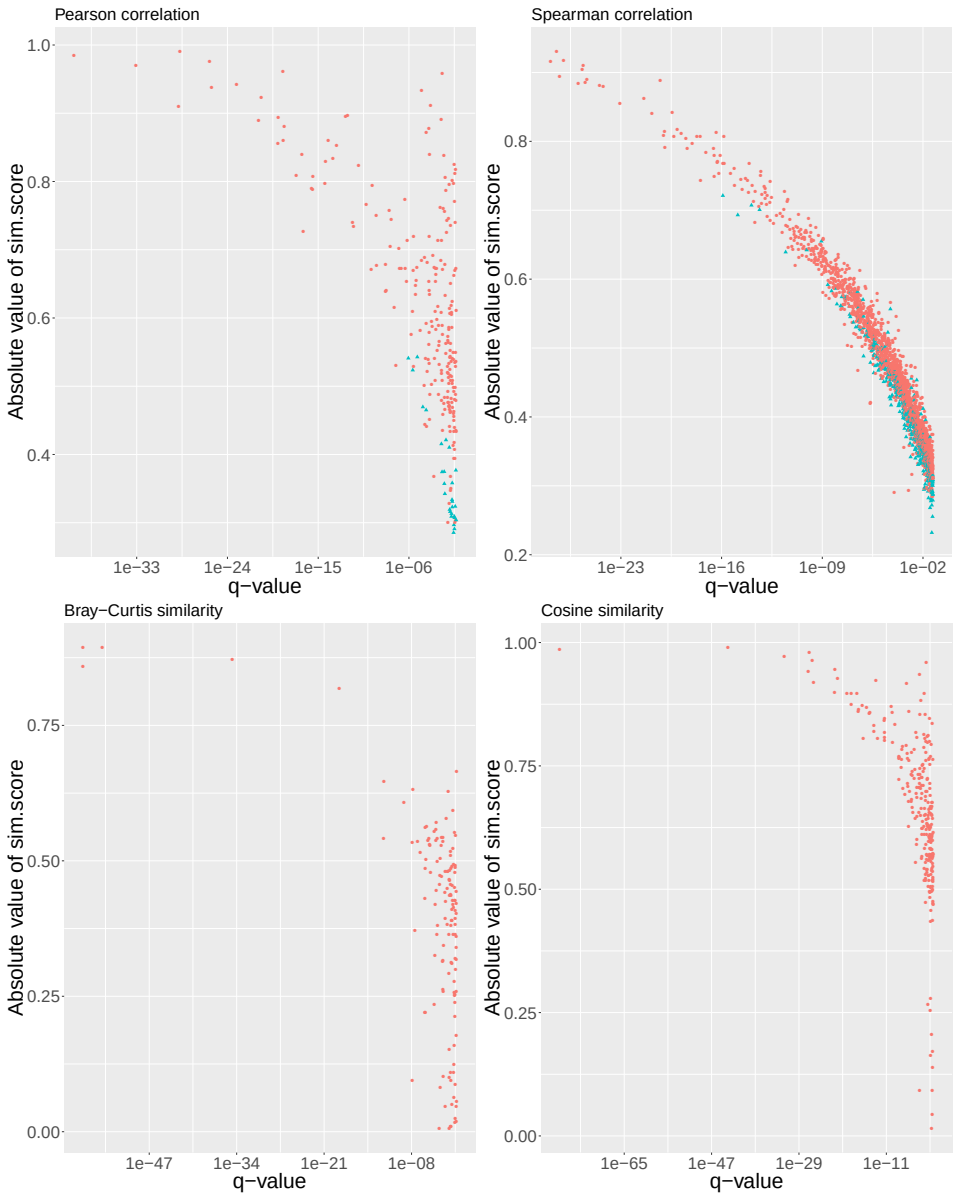


Figure B.6: Similarity scores versus q-values for each significant interaction in the biofilm experiment. Red circles indicate positive interactions, whereas blue triangles indicate negative interactions.

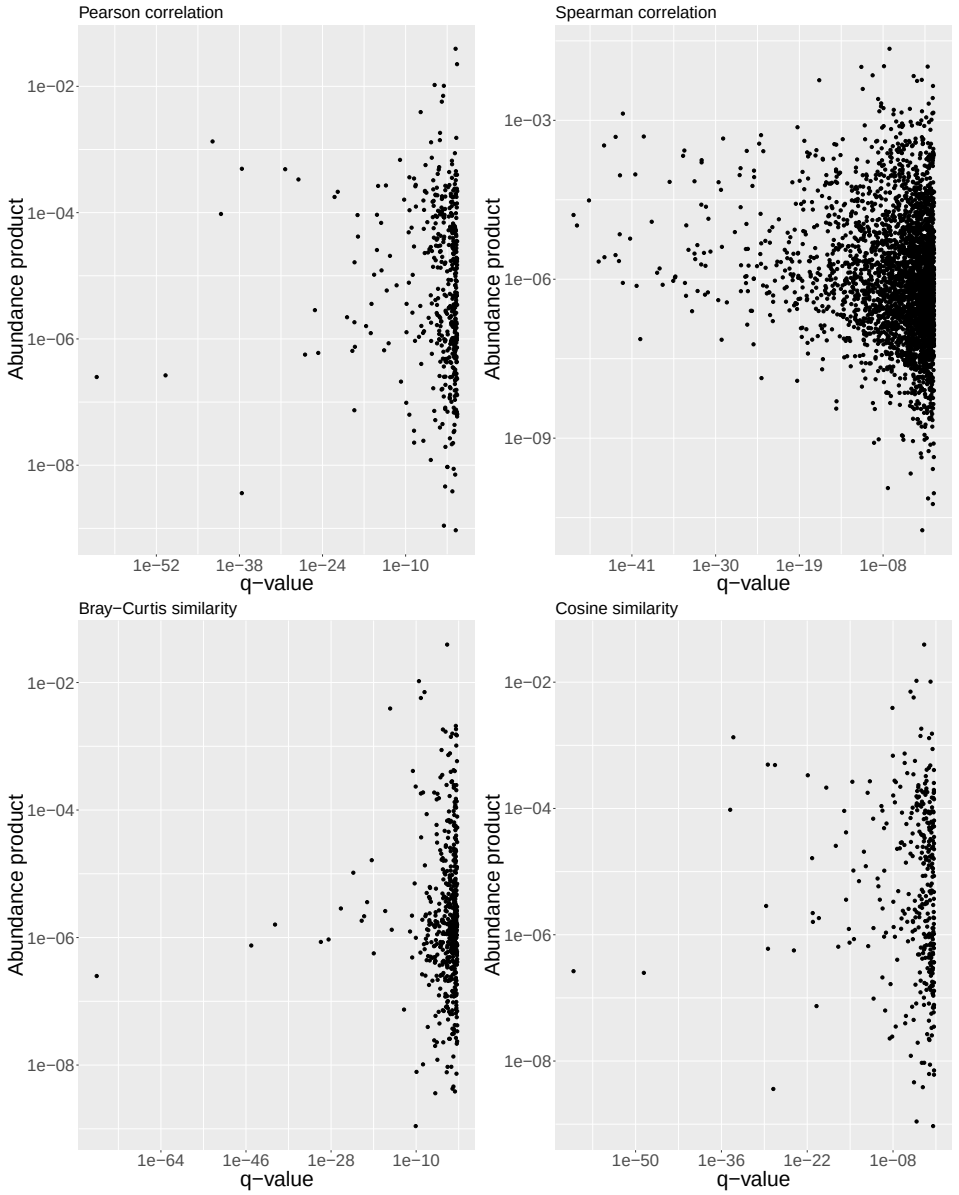


Figure B.7: Abundance product versus q-values for each significant interaction in the selection-switch experiment with relative data

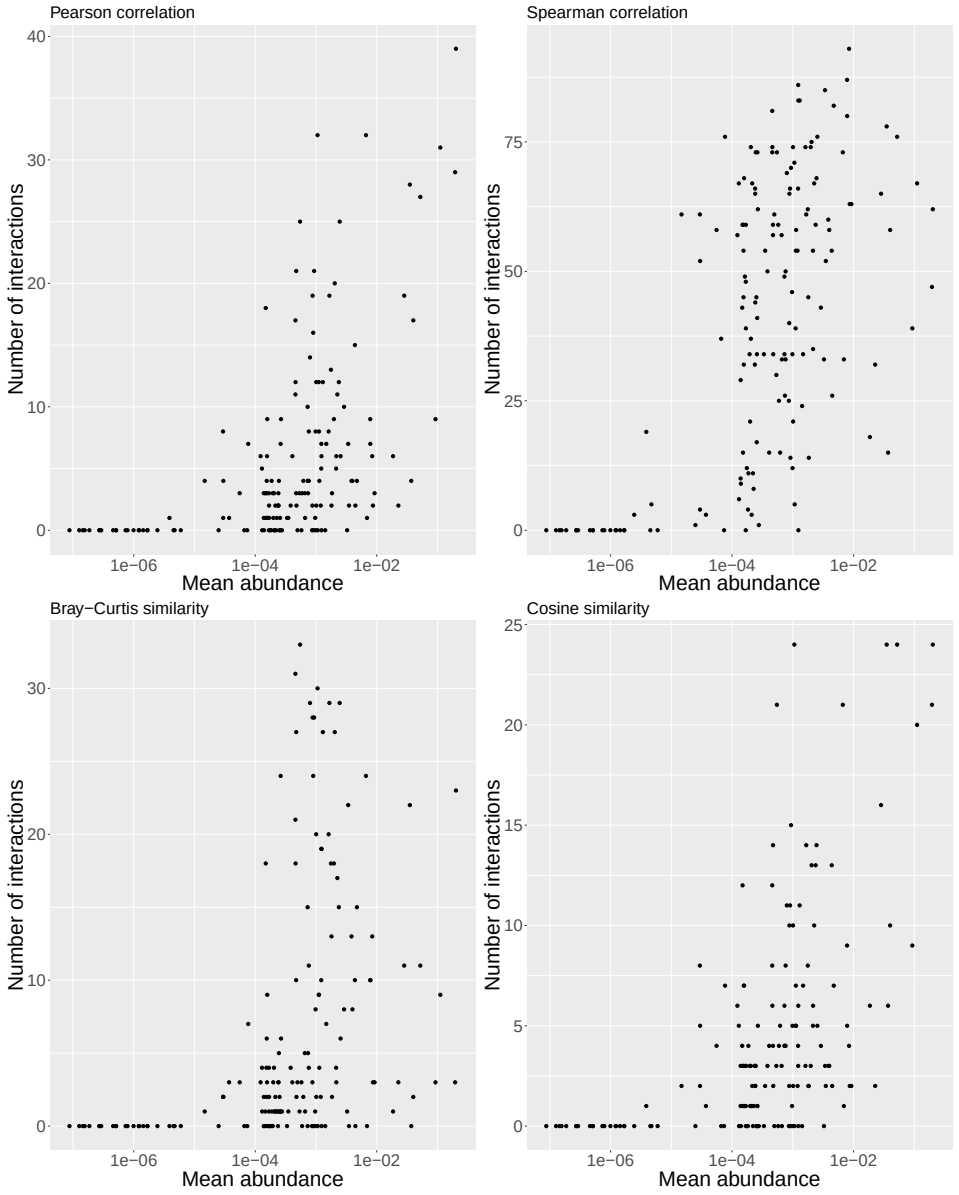


Figure B.8: Number of significant interactions versus overall mean abundance for each OTU in the selection-switch experiment with relative data

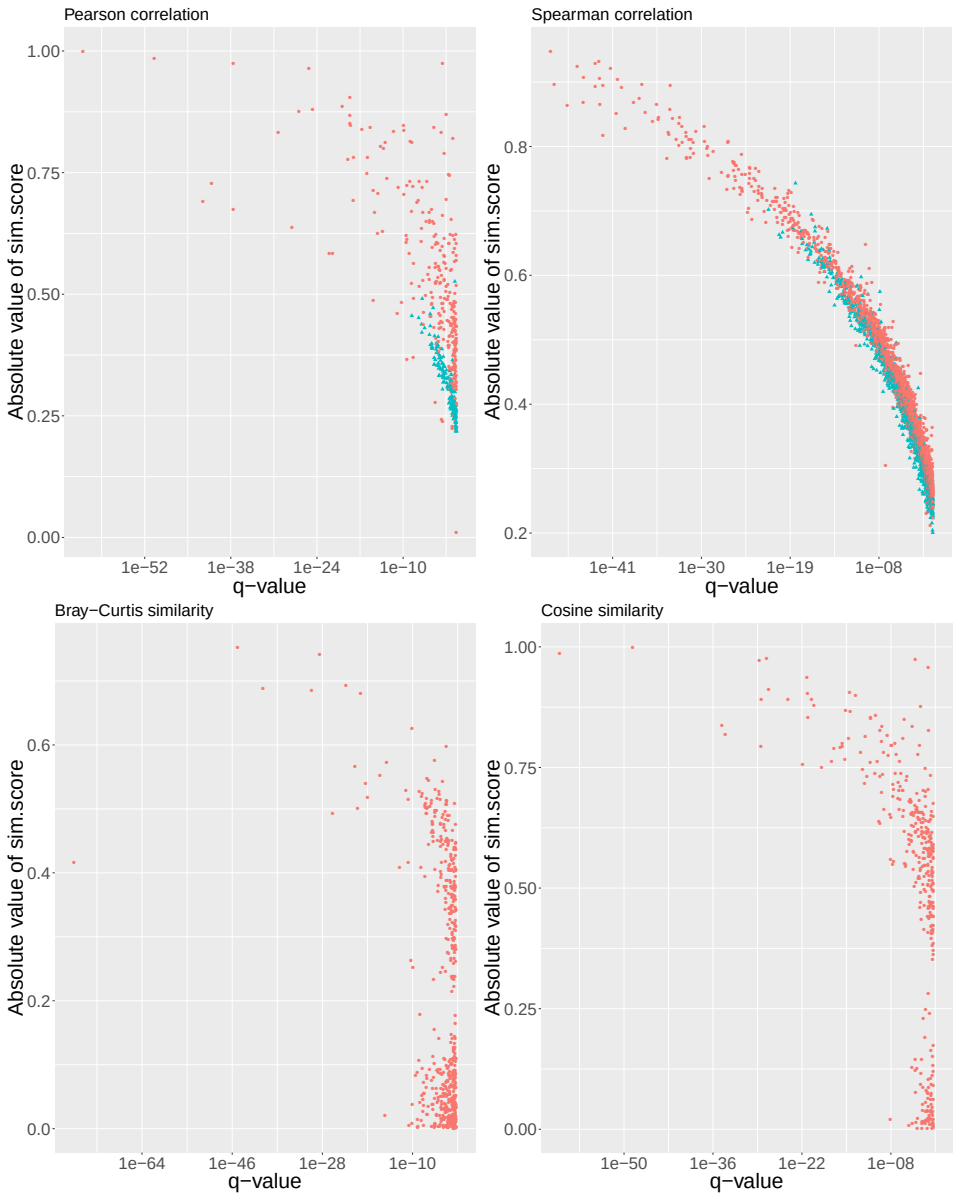


Figure B.9: Similarity scores versus q-values for each significant interaction in the selection-switch experiment with relative data. Red circles indicate positive interactions, whereas blue triangles indicate negative interactions.

Table B.1: Figure references for diagnostic plots

Dataset	Abundance product	Number of interactions VS abundance	`sim.score` VS q-values
Seawater	B.1	B.2	B.3
Biofilm	B.4	B.5	B.6
Selection-switch with relative abundances	B.7	B.8	B.9
Selection-switch with absolute abundances	4.2	4.3	4.4

Appendix C

Interaction networks

This is an extension of Section 4.2.3, where networks of significant interactions are plotted. For the seawater experiment, the networks are shown in Figure C.1, while Figure C.2 shows the networks for the biofilm experiment. Networks from the selection-switch experiment for relative data are shown in Figure C.3.

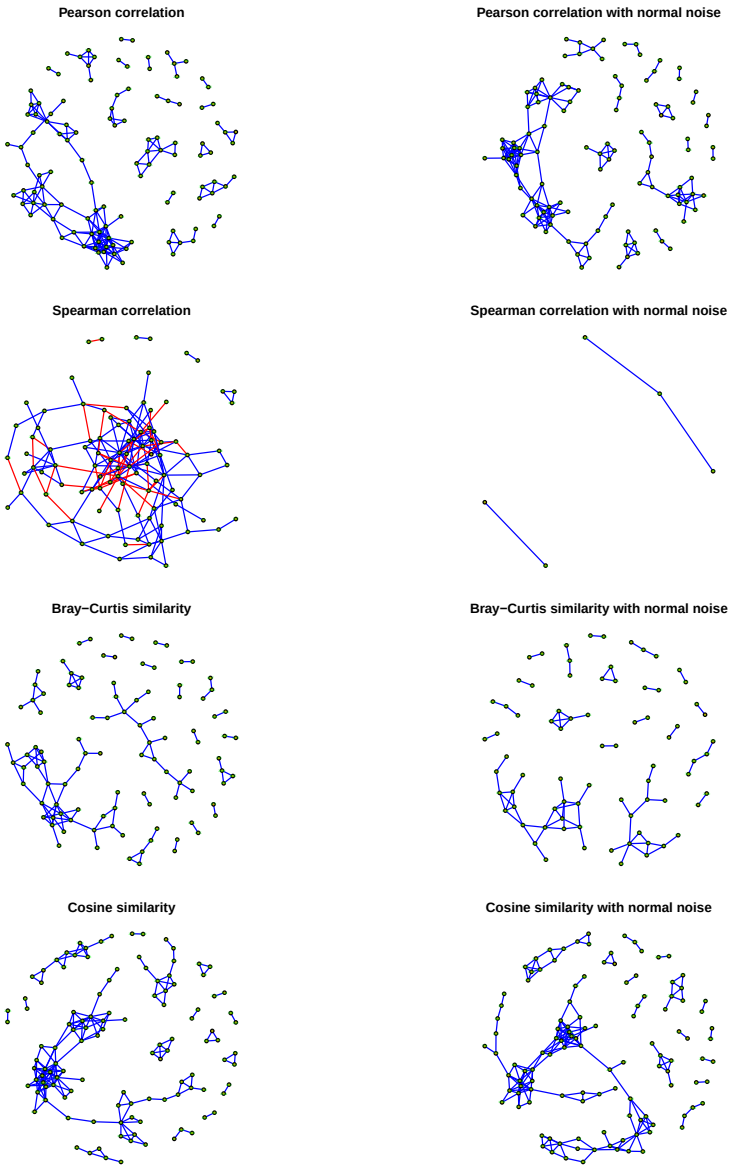


Figure C.1: Network of significant interactions for the seawater experiment. Blue edges correspond to positive interactions, whereas red edges correspond to negative interactions

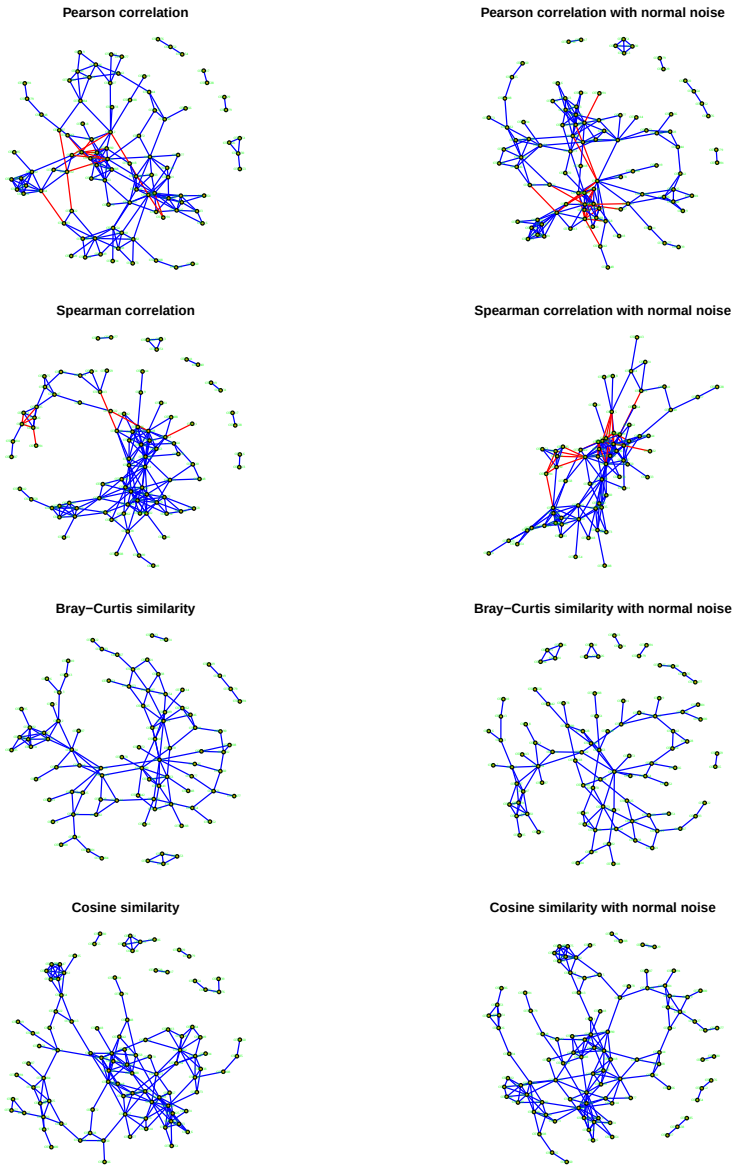


Figure C.2: Network of significant interactions for the biofilm experiment. Blue edges correspond to positive interactions, whereas red edges correspond to negative interactions

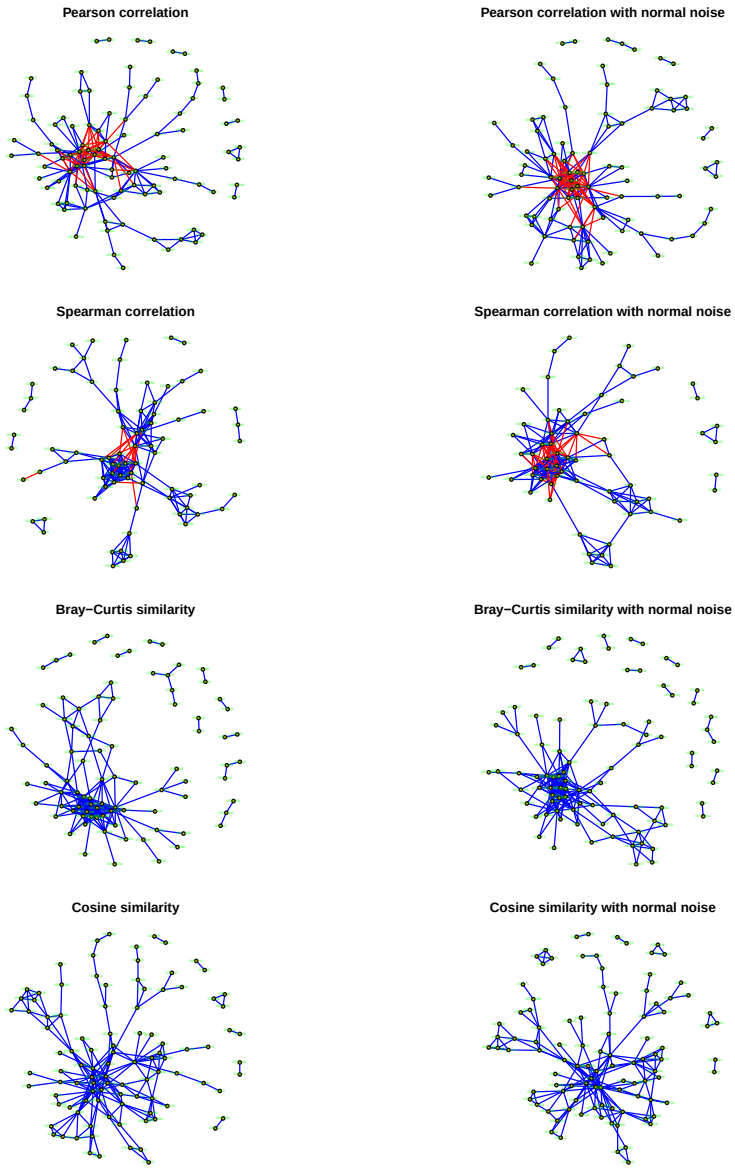


Figure C.3: Network of significant interactions for the selection-switch experiment with relative data. Blue edges correspond to positive interactions, whereas red edges correspond to negative interactions

Appendix D

PCoA plots

This is an extension of Section 4.2.4, where PCoA ordinations of the OTUs are presented. Table D.1 contains the references to the figures.

Table D.1: Figure references for PCoA plots

Dataset	Without ReBoot	With ReBoot
Seawater	D.1	D.4
Biofilm	D.2	D.5
Selection-switch with relative abundances	D.3	D.6
Selection-switch with absolute abundances	4.9	4.10

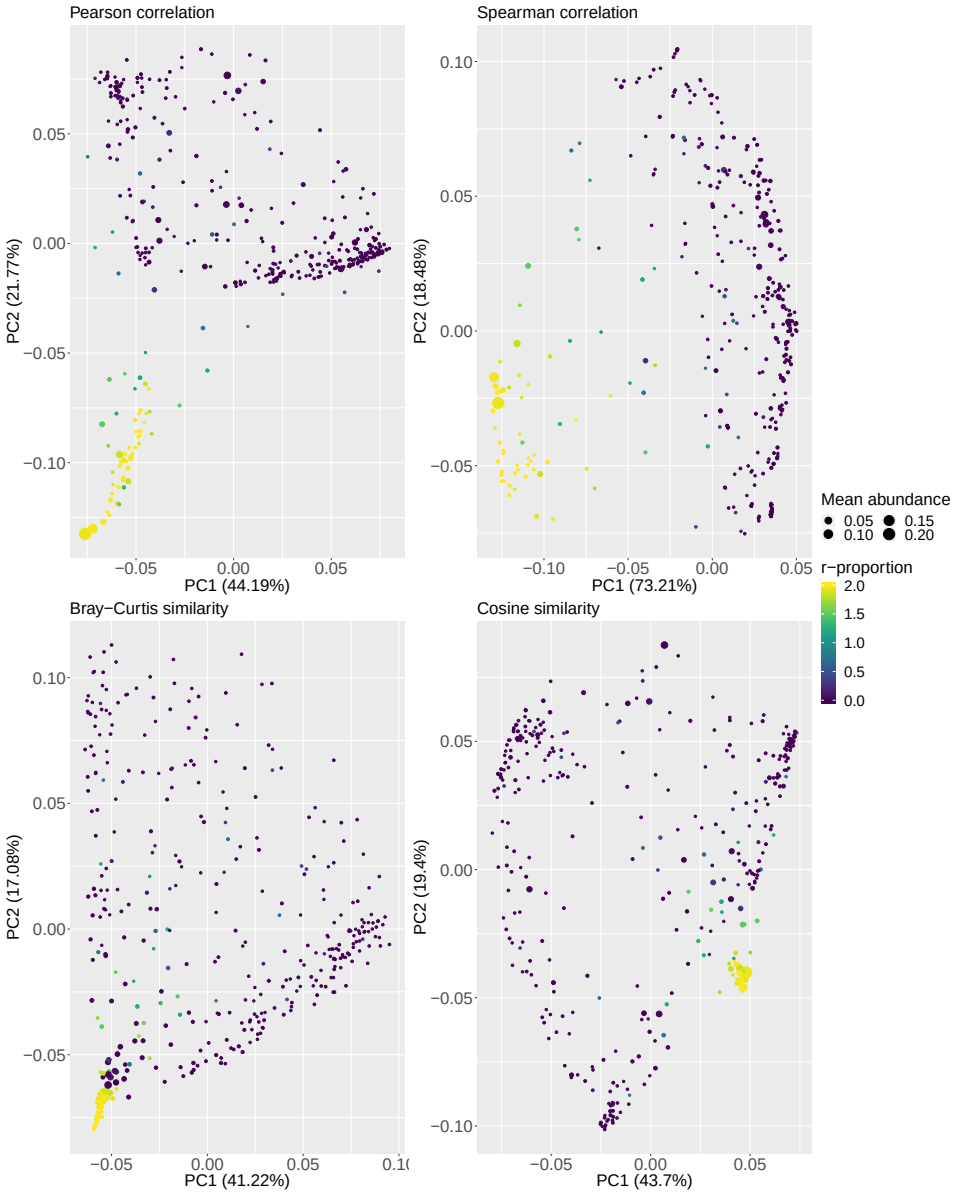


Figure D.1: PCoA ordination plots for the seawater experiment without using the ReBoot pipeline. The r-proportion is explained in Section 3.4.1. The markers sizes correspond of the overall mean relative abundance.

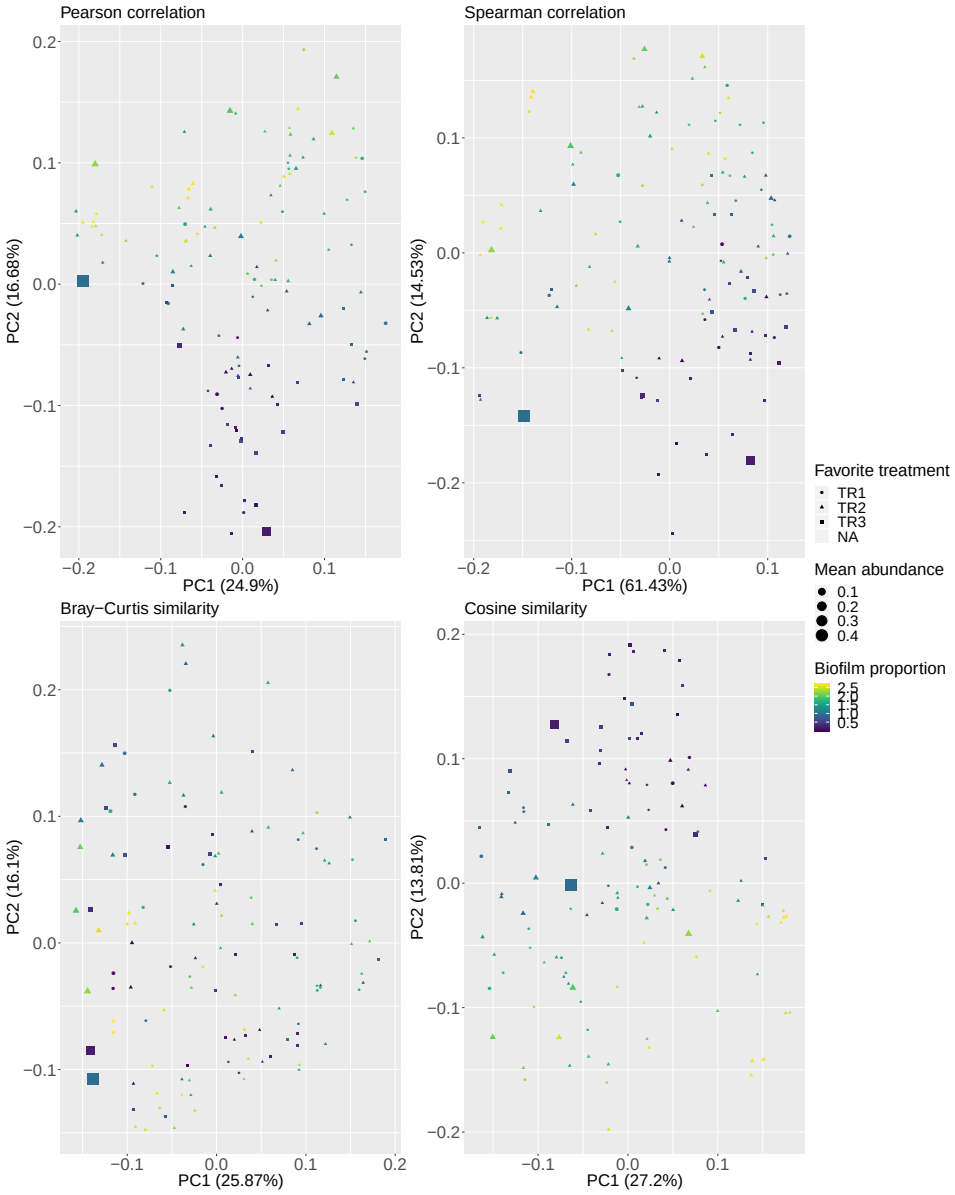


Figure D.2: PCoA ordination plots for the biofilm experiment without using the ReBoot pipeline. The favorite treatment and biofilm proportion are explained in Section 3.4.1. The marker size is determined by the overall mean abundance.

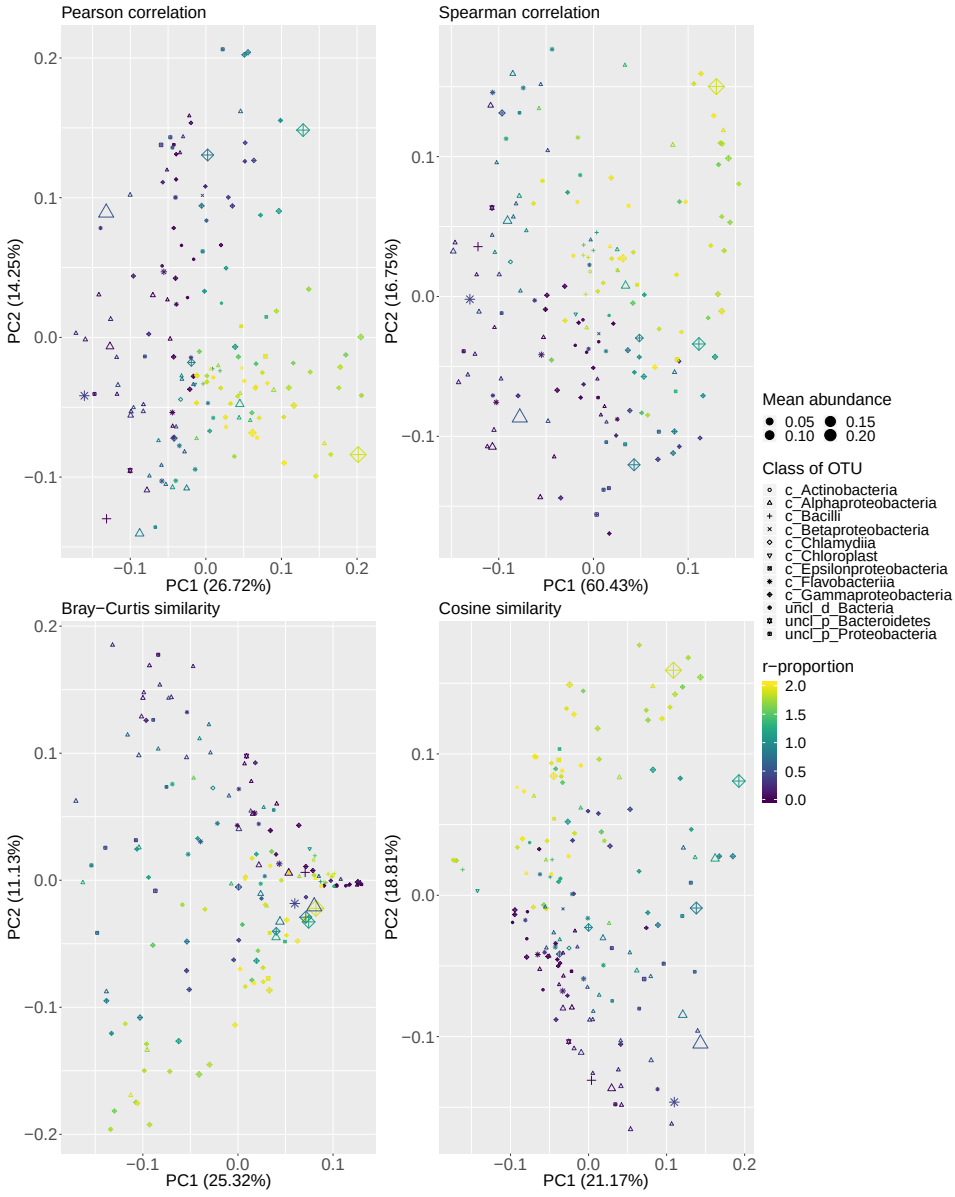


Figure D.3: PCoA ordination plots for the relative data from the selection-switch experiment without using the ReBoot pipeline. The r-proportion is explained in Section 3.4.1. In this plot, the taxonomies of the OTUs on class level are included.

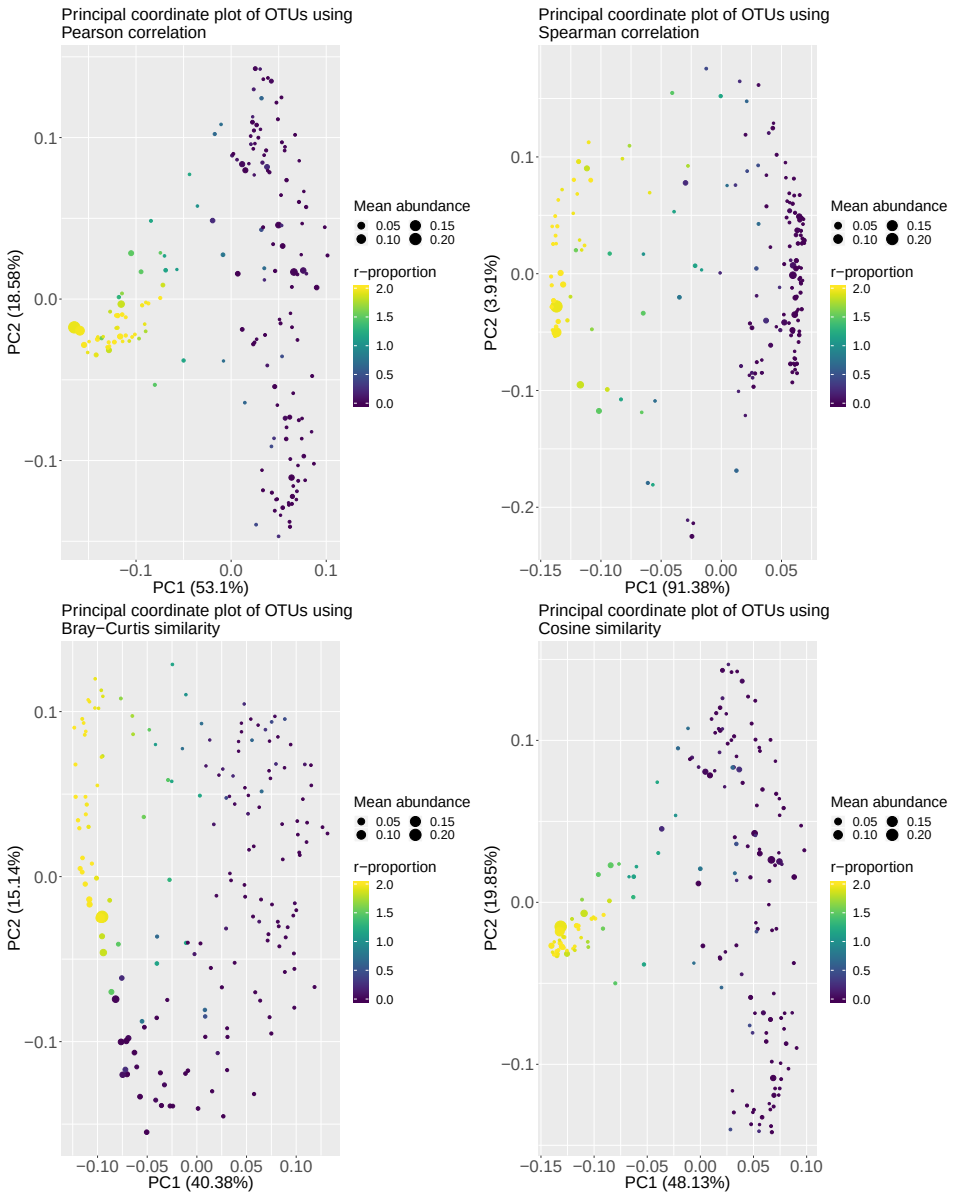


Figure D.4: PCoA ordination plots for the seawater experiment using the ReBoot pipeline. The r-proportion is explained in Section 3.4.1. The markers sizes correspond of the overall mean relative abundance.

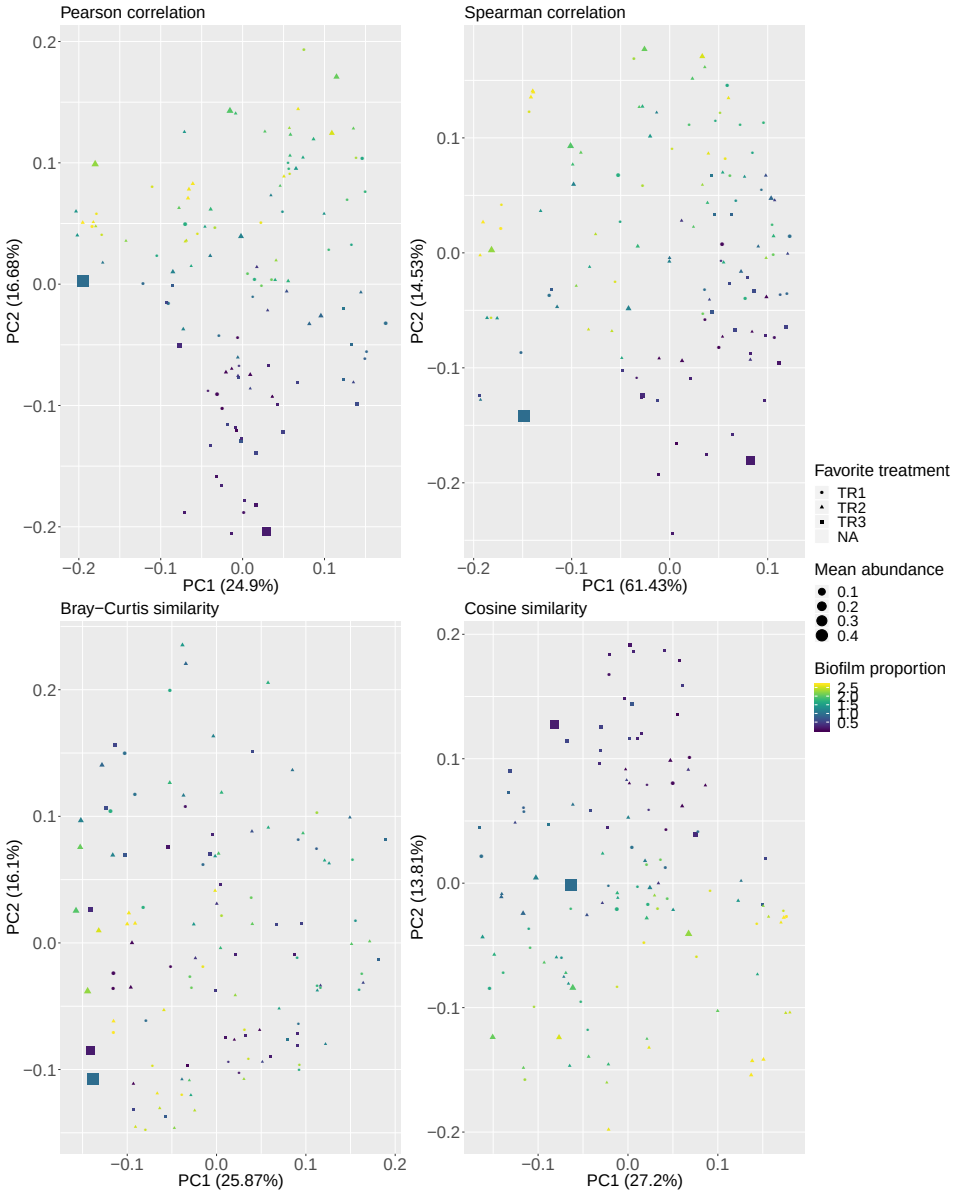


Figure D.5: PCoA ordination plots for the biofilm experiment using the ReBoot pipeline. The favorite treatment and biofilm proportion are explained in Section 3.4.1. The marker size is determined by the overall mean abundance.

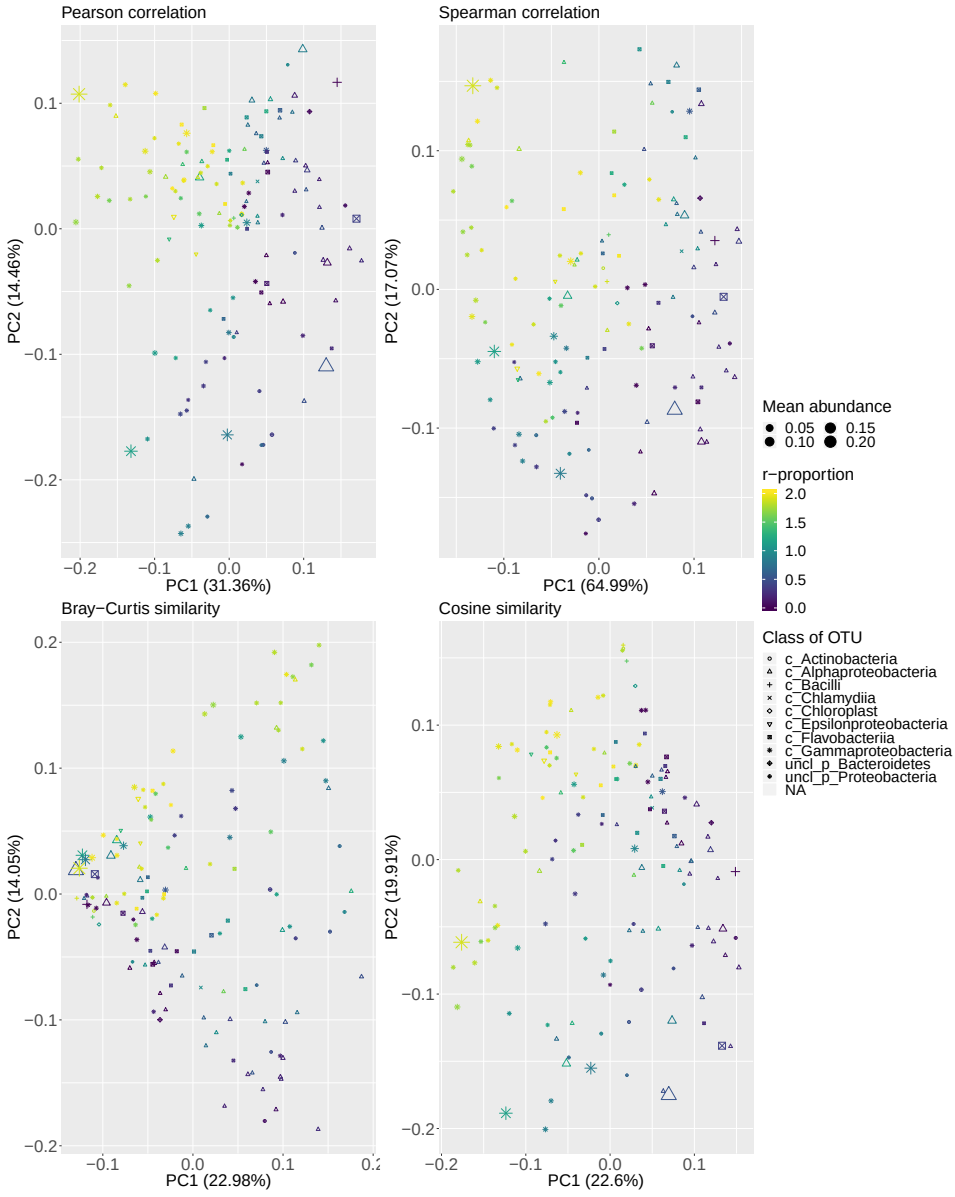
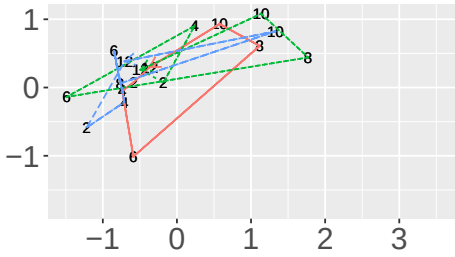


Figure D.6: PCoA ordination plots for the relative data from the selection-switch experiment using the ReBoot pipeline. The r-proportion is explained in Section 3.4.1. In this plot, the taxonomies of the OTUs on class level are included.

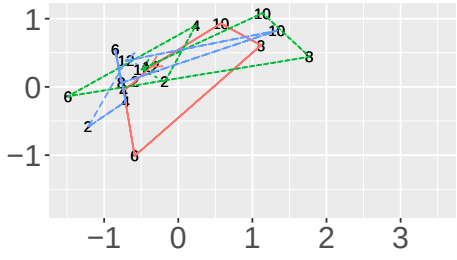
Appendix E

Trajectory plots

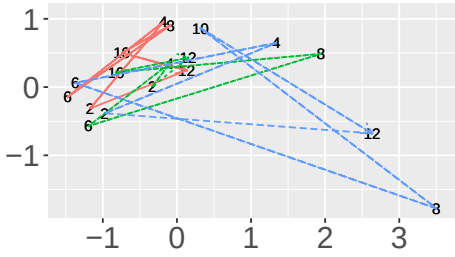
This is an extension of Section 4.3.1, where time trajectory plots are presented. For the biofilm experiment, the plots are shown in Figure E.1, while Figure E.2 shows the trajectory plots from the selection-switch experiment with relative abundances.



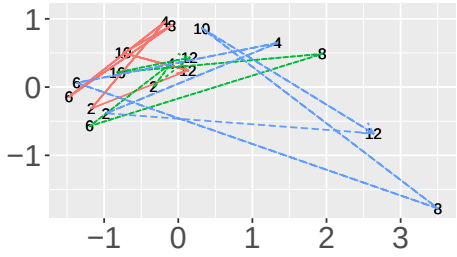
(a) Treatment 1 with water samples. The time series plotted are: Water1 (red, solid), Water2 (green, short dashes) and Water3 (blue, long dashes).



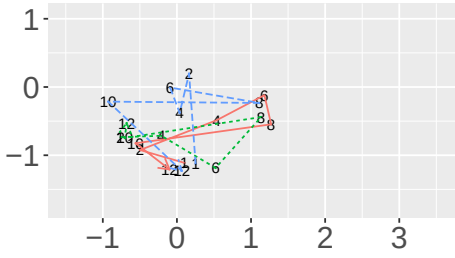
(b) Treatment 1 with samples from biofilm carriers.The time series plotted are.The time series plotted are: Carrier1 (red, solid), Carrier2 (green, short dashes) and Carrier3 (blue, long dashes).



(c) Treatment 2 with water samples. The time series plotted are: Water4 (red, solid), Water5 (green, short dashes) and Water6 (blue, long dashes).

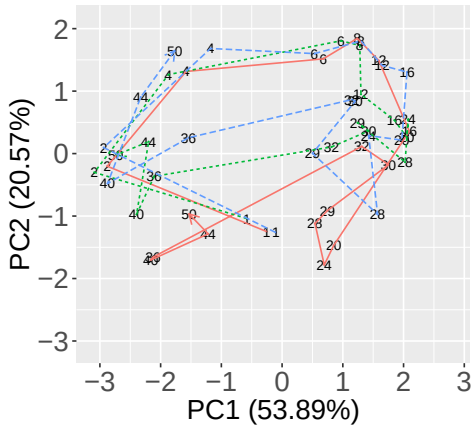


(d) Treatment 2 with samples from biofilm carriers. The time series plotted are: Carrier4 (red, solid), Carrier5 (green, short dashes) and Carrier6 (blue, long dashes).

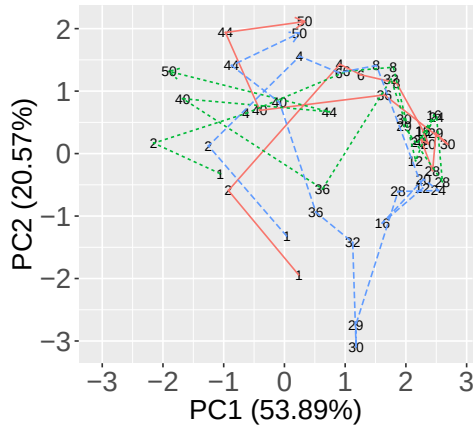


(e) Treatment 3 with water samples. The time series plotted are: Water7 (red, solid), Water8 (green, short dashes) and Water9 (blue, long dashes).

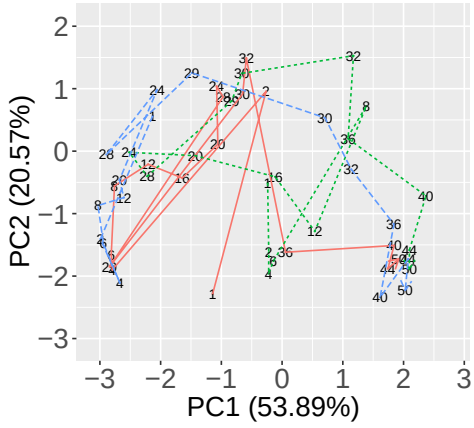
Figure E.1: Trajectory plots for the biofilm experiment. The data are ordinated by PCoA using Bray-Curtis similarity. The text labels on the points correspond to the week of sampling. All time series in the overall figure were used to make the ordination. Later, the time series stemming from identical selection regimes were superimposed on each individual subfigure.



(a) K-selected community before day 29, r-selected community after day 29, with high carrying capacity. The time series plotted are: KRH1 (red, solid), KRH2 (green, short dashes) and KRH3 (blue, long dashes).

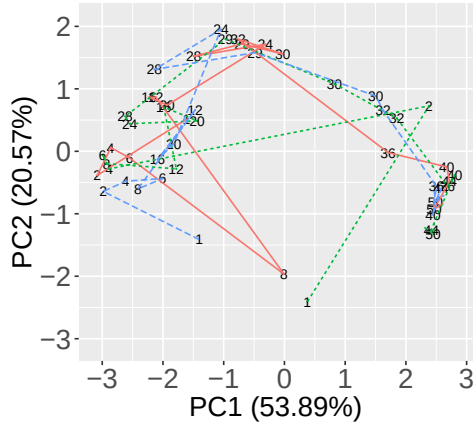


(b) K-selected community before day 29, r-selected community after day 29, with low carrying capacity. The time series plotted are: KRL1 (red, solid), KRL2 (green, short dashes) and KRL3 (blue, long dashes).



(c) r-selected community before day 29, K-selected community after day 29, with high carrying capacity. The time series plotted are: RKH1 (red, solid), RKH2 (green, short dashes) and RKH3 (blue, long dashes).

.



(d) r-selected community before day 29, K-selected community after day 29, with low carrying capacity. The time series plotted are: RKL1 (red, solid), RKL2 (green, short dashes) and RKL3 (blue, long dashes).

Figure E.2: Trajectory plots for the selection-switch experiment with relative abundances. The data are ordinated by PCoA using Bray-Curtis similarity. The text labels on the points correspond to the day of sampling. All time series in the overall figure were used to make the ordination. Later, the time series stemming from identical selection regimes were superimposed on each individual subfigure.

Appendix F

Cross-validation colorplots

This is a continuation of Section 4.3.3, presenting cross-validation colorplots for the Lotka-Volterra pipeline. For the biofilm dataset, the results are shown in Figure F.1. The cross-validation colorplot based on stringent filtering at mean relative OTU abundance 10^{-3} of the selection-switch experiment is presented in Figure F.2.

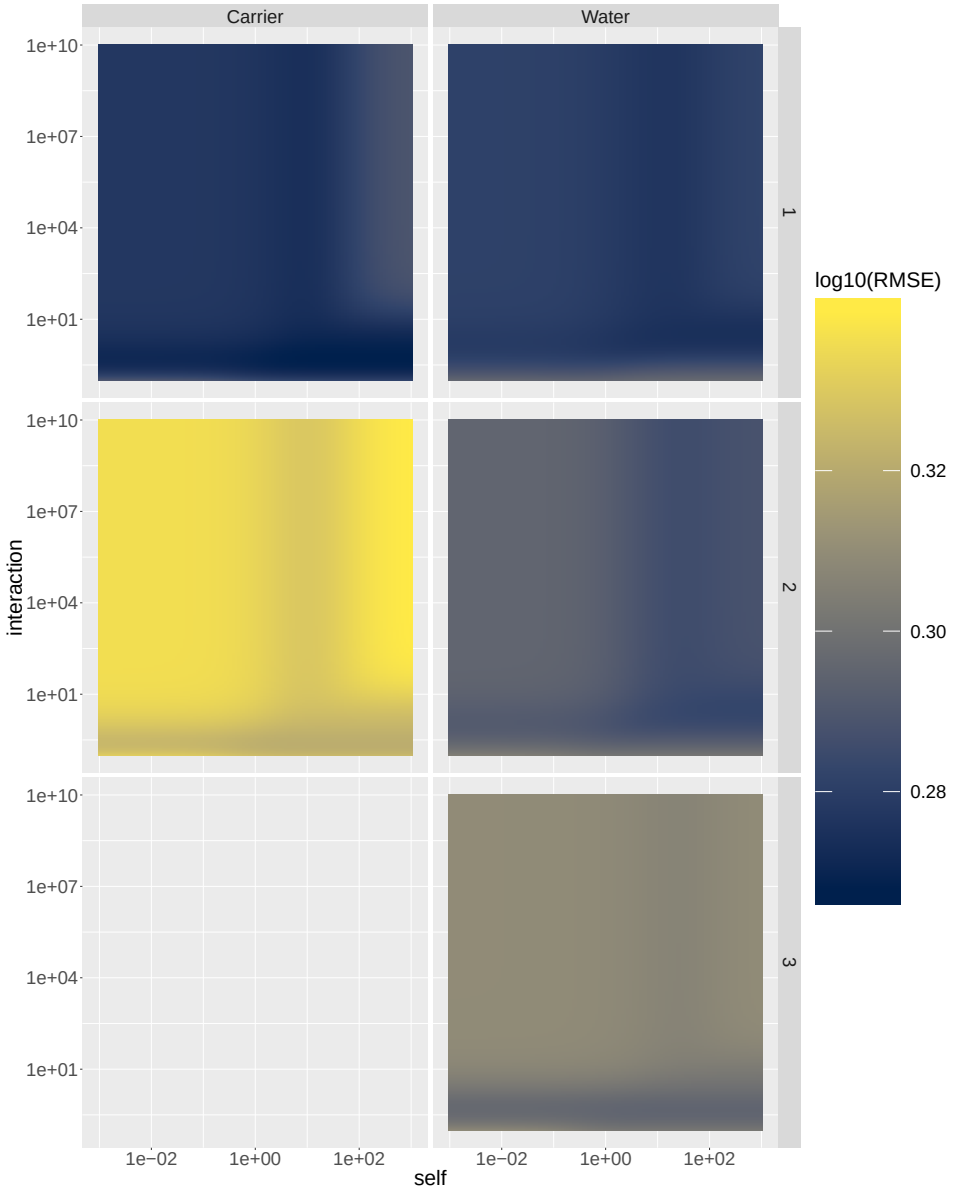


Figure F.1: Cross-validation result for the biofilm experiment with ordinary filtering at mean relative abundance 10^{-4}. The color reported corresponds to base-10 logarithm of the Root Mean Squared Error. The plots are shown for each possible combination of source of sample (in columns) and treatment (in rows: TR1,TR2, TR3).

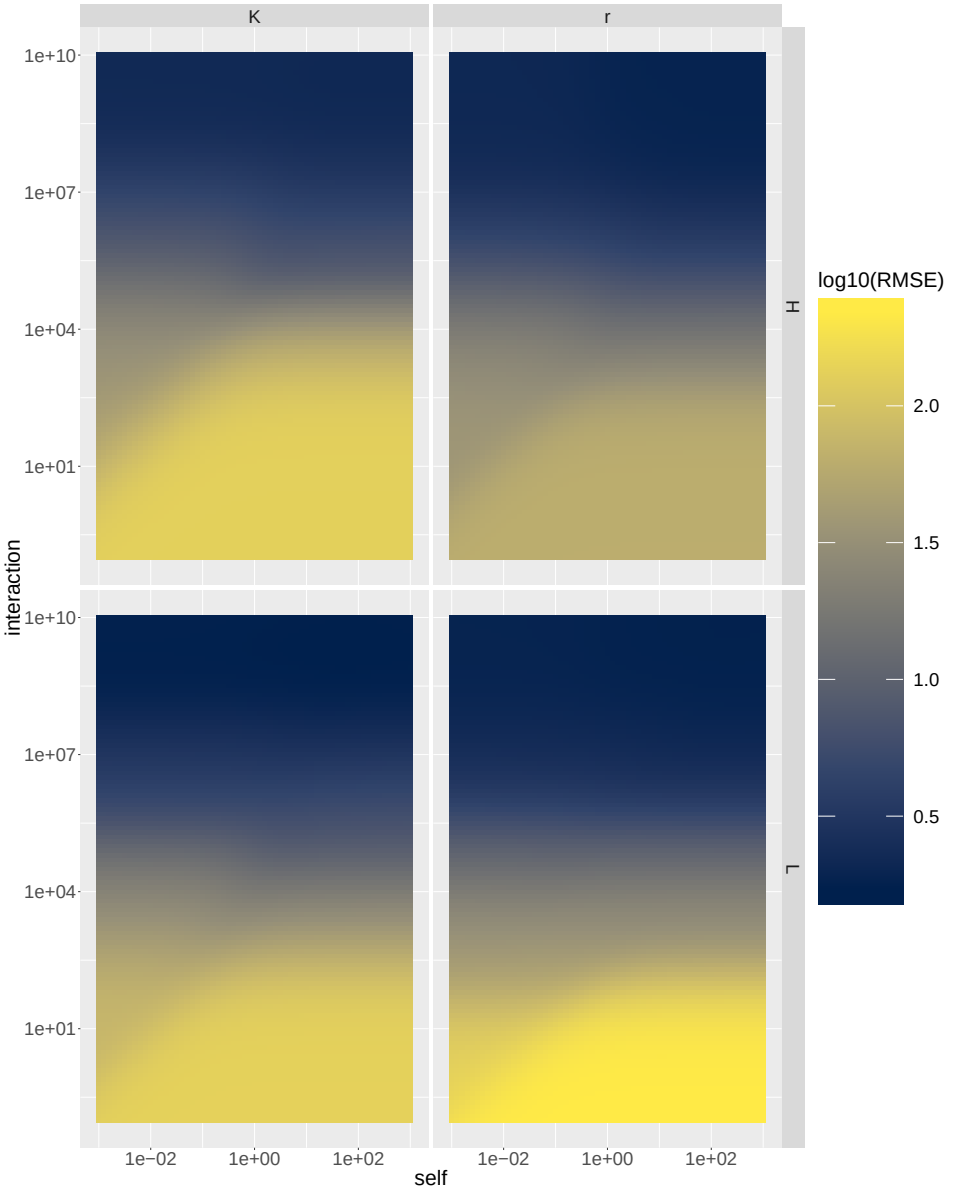


Figure F.2: Cross-validation result for the selection-switch experiment with stringent filtering at mean relative abundance 10^{-3}. The color reported corresponds to base-10 logarithm of the Root Mean Squared Error. The plots are shown for each combination of present selection regime (in columns) and nutrient supply (in rows).

Appendix G

Time trajectory plot of predicted time series

This is a continuation of Section 4.3.4, presenting time trajectory plots of predicted time series. For the selection-switch experiment with stringent filtering, this is shown in Figure G.1

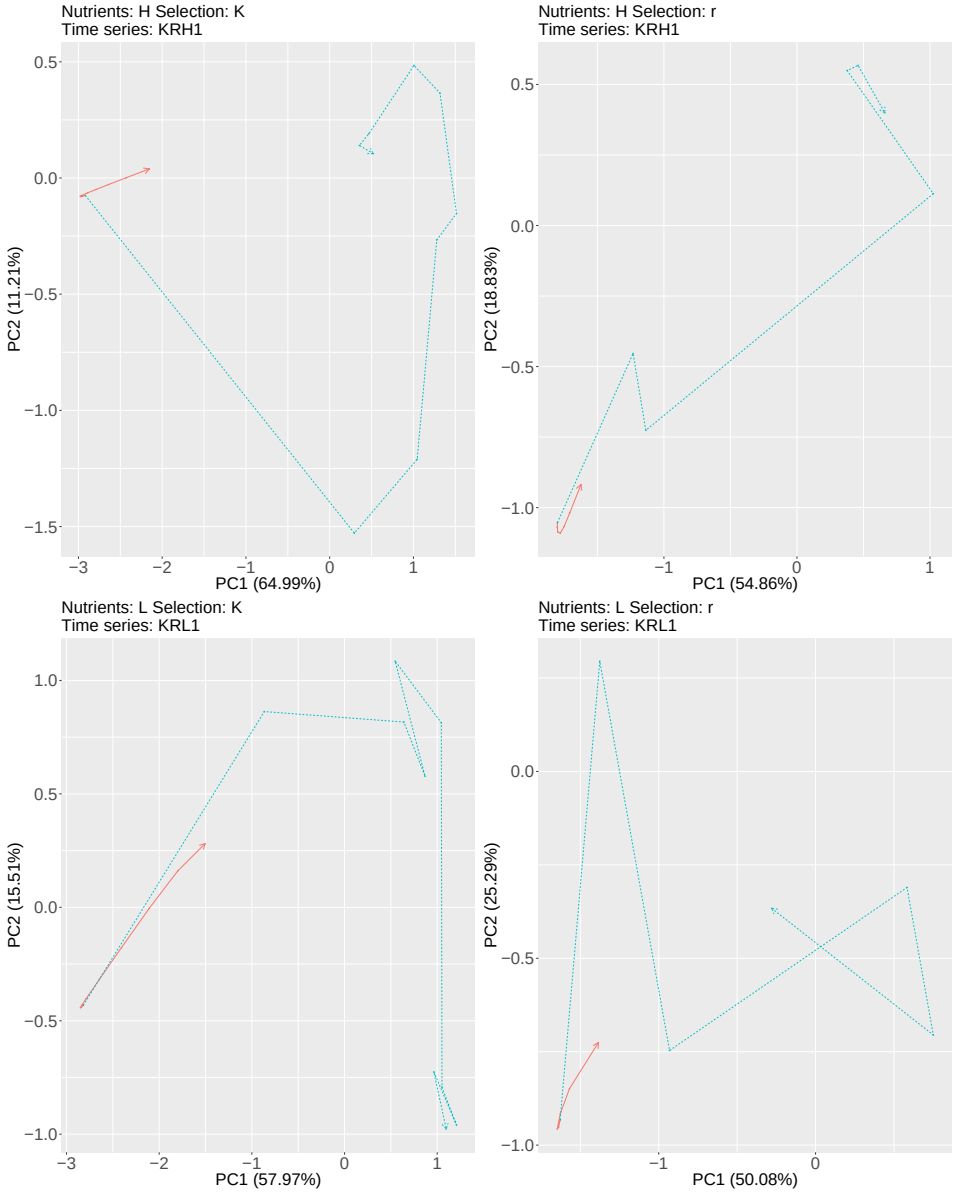


Figure G.1: Predicted (determined through inferred gLV coefficients, shown as red solid line) and reference (actual, shown as blue dashed line) time series from the selection-switch experiment with stringent filtering at mean relative abundance 10^{-3}. The time series are shown in PCoA ordinations using Bray-Curtis similarity. The ordinations are based on all time series of each subdivision.