Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN FISICA

Ciclo XXIX

Settore Concorsuale di afferenza: 02/D1

Settore Scientifico disciplinare: FIS/07

STOCHASTIC MODELING AND STATISTICAL PROPERTIES OF BIOLOGICAL SYSTEMS INFERRED FROM OMICS DATA

Presentata da: Claudia Sala

Coordinatore Dottorato:

Prof. Gastone Castellani

Relatore:

Prof. Gastone Castellani

Esame finale anno 2017

Contents

Introduction

It is well known, nowadays, that a correct description of biological systems should take into account stochasticity rather than exclusively rely on deterministic models. Biological processes are essentially random and randomness characterizes, for instance, both cellular behavior and the cellular environment [1]. Stochasticity arises, for instance, because biochemical reactions are generated by random intermolecular collisions and is typically due to two sources of noise [2]: intrinsic noise, that is generated by the dynamics of the system from the random timing of individual reactions, and extrinsic noise, that is generated by the system interacting with other stochastic systems in the cell or with the environment. The cell has several ways to deal with noise and to favor the reactions it needs. For example, it produces specific enzymes, that increase the rate of the desired reactions by lowering their activation energy. However, noise still remains a characteristic of both collisions and bonding processes and, eventually, the cell has taken advantage from such stochasticity, that is what allows, for instance, variability, adaptation and evolution.
The effect of noise can be disregarded in systems composed by large molecule populations, typically in the order of 10^{23}. In this case, in fact, what emerges is an average behavior that is nearly deterministic and can be described by a set of ordinary differential equations. In biology, however, the number of molecules is small. Typical molecule numbers of the same protein specie in a cell are usually no more than a few thousands and, indeed, fluctuations are not negligible. For instance, only one gene is involved in most activities of gene expressions and less than 20 transcriptions of mRNAs from a single gene are present in an individual bacteria [3].
The biological systems that we are going to consider in this thesis are treated from an ecological point of view. In particular, in Part I we will propose a model to describe the population dynamics of the Gut Microbiota, while in Part III we will focus on the ensemble of protein domains in bacterial genomes and we will consider it as an ecosystem, so that to obtain an insight in the genome evolution process. As in the single cell, also when we consider the dynamic of a whole population, noise is an important feature. Specifically, the dynamics of population has both deterministic and stochastic components that operate simultaneously [4] and usually include three basic forms of noise: demographic stochasticity, environmental stochasticity and sampling error [4].

- **Demographic stochasticity** refers to chance events of individual mortality and reproduction. It is usually conceived as being independent among individuals and is consequently modeled as an additive term in the dynamic equation, that has the same amount on all species, independently from their abundance. Consequently, its effect will be bigger for small populations than for the larger ones and, for this reason, it is referred to as 'density dependence stochasticity'.

- **Environmental stochasticity** refers to temporal fluctuations in the probability of mortality and the reproductive rate of all individuals in a population in the same or similar fashion. The impact of environmental stochasticity is roughly the same for small and large populations and it therefore constitutes an important risk of population decline in all populations regardless of their abundance at a given location.

From a mathematical point of view, environment noise is represented by a multiplicative term, that, being multiplied by the population abundance, has a density independent effect.

- **Measurement error**, finally, is due to uncertainties in the estimate of the population size or density, that is usually based on a sampling procedure. For populations in which a complete and accurate census is available, measurement error can be ignored.

It is clear, then, that in order to model biological systems, and specifically ecosystems, the correct approach is to use a stochastic description. In the first introductory chapter we will report some basic concepts of stochastic process theory that are required to understand the models used in the following. We will then give an overview on ecological theory. We will introduce the concept of biodiversity and explain how it can be described through the Relative Species Abundance distribution (RSA). Then, we will detail some of the most common theories that have been proposed in the past. In particular, we will focus on the class of theories that are called neutral. These were formalized in 2001 by Hubbell [5], that effectively introduced in community ecology the Ockham's razor concept, according to which science should aim at finding the minimal set of processes that can satisfactorily explain observed phenomena. The main hypothesis of neutral theories are then that species can be considered equivalent, meaning that they have the same dynamics rate, and that species interaction can be neglected.
In Part I, we will show that a purely neutral model in which species are subject to demographic noise is not appropriate to describe the Gut Microbiota ecosystem and that the assumption species equivalence needs to be relaxed. Moreover, we will also show that the biodiversity measure obtained with our modeling is able to discriminate elderly subjects that have a good or a worse health state and that it is able to predict healthy aging with much better accuracy than using biodiversity indices that are simply based on the relative abundance of species and do not take into account stochasticity. The Gut Microbiota modeling involves the computation of the RSA distribution, and as we will detail in Part I, the correct way to construct it is by clustering together 16S rRNA sequences so that to obtain a definition of species that does not rely on human made taxonomic classifications but reflects the phylogenetic relationships between bacteria. In Part II we will summarize the most common procedures that exist to cluster the 16S rRNA sequences and we will propose to use a recently developed method that has the main advantage of being totally parameter-free. Finally, in Part III we will consider the ecosystem composed by the set of protein domains in the bacterial genome. We will show that, in this case, environmental noise should also be taken into account and that differences in the RSA distribution of protein domains reflect the phylogenetic distances between bacteria.

Stochastic processes

Here we summarize some key concepts of stochastic processes that needed in the following, taking the cue mainly from the two books: 'A second course in stochastic processes' by S. Karlin and H.M. Taylor [6] and 'Stochastic processes in physics and chemistry' by N.G. Van Kampen [7].

0.1 Stochastic processes

A stochastic process is a phenomenon that evolves with time (process) and whose evolution depends on random factors (stochastic). We can define a stochastic process $\{X_t; t \in T\}$, or equivalently $\{X(t); t \in T\}$, as a family of random variables $x_1, x_2, ..., x_t, ...$ indexed by a parameter t that runs over a suitable set T. The indices t may correspond to continuous or discrete units of time, and in the last case, $\{X_t\}$ could represent the outcomes at successive trials like the result of tossing a coin or the successive observations of some characteristic of a population.

0.1.1 Markov process

A Markov process is a stochastic process with the property that, given the value of X_t, the values of X_s, $s > t$, do not depend on the values of X_u, $u < t$. This means that the probability of any particular feature behavior of the process, when its present state is known exactly, is not altered by additional knowledge concerning its past. Formally, a stochastic process $\{X_t\}$ is said to be Markovian if

$$P\{a < X_t \leq b \mid X_{t_1} = x_1, X_{t_2} = x_2, ..., X_{t_n} = x_n\} = P\{a < X_t \leq b \mid X_{t_n} = x_n\} \qquad (1)$$

whenever $t_1 < t_2 < \cdots < t_n < t$.

0.1.2 Stationary Markov process

A Markov process $\{X(t)\}$ is said to be (strongly) stationary if the joint distribution of $\{X(t)\}$ at times $t_1 + h$, $t_2 + h$, ..., $t_k + h$ is the same for every k, $h = 0, 1, 2, ...$ and for all $t_1, t_2, ..., t_k$. This is true if the statistical properties of $X(t)$ do not change over time

$$P(X_{t_1} \leq x_1, X_{t_2} \leq x_2, ..., X_{t_n} \leq x_n) = P(X_{t_1+h} \leq x_1, X_{t_2+h} \leq x_2, ..., X_{t_n+h} \leq x_n) \qquad (2)$$

for every n and for every h. This relation implies that all the existing moments do not change over time and that the family of random variables that define the process itself are independent and identically distributed (i.i.d.). In fact, if we take $n = 1$, we have $P(X_{t_1} \leq x) = P(X_{t_2} \leq x) = ... = P(X_{t_k} \leq x) = P(X_{t_1+h} \leq x)$, meaning that the x_t have the same cumulative distributions. Note that the Poisson process that we are going to describe in the next section is not a stationary process, in fact, as we will see, its mean is $\mu = \mu(t) = \lambda t$ and depends on the time t.

0.1.3 Poisson process

A Poisson process is a Markov process with continuous time and discrete state space $\mathbb{N}$. It describes the number of times a specific event occurs (success) during a time or space interval, when the success probability is small but the number of trials is big. Formally, it is defined in the following way.

Definition 1. A Poisson process $\{X_t\}$ with intensity $\lambda > 0$ is a Markov process defined on the non-negative integers which has the following properties

i) $X(0) = 0$.

ii) $\{X_t\}$ has independent increments, meaning that the number of events happening in two disjoint intervals of time are independent.

iii) $\{X_t\}$ has stationary (or homogeneous) increments, meaning that the number of events happening in an interval $[t, t + \Delta t]$ depends only on the interval length Δt and not on the position of the interval on the time axis t.

iv) $P\{X(t + h) - X(t) = 1 \mid X(t) = x\} = \lambda h + o(h)$, as $h \downarrow 0$, is the probability that one event occurs in an infinitesimal interval.

v) $P\{X(t + h) - X(t) = 0 \mid X(t) = x\} = 1 - \lambda h + o(h)$, as $h \downarrow 0$, is the probability that no event occurs in an infinitesimal interval.

vi) $P\{X(t + h) - X(t) = 2 \mid X(t) = x\} = o(h)$, as $h \downarrow 0$, is the probability of two events to occur in an infinitesimal interval.

Properties iv to vi imply that in a time interval $h \downarrow 0$ either one event occurs or no events at all and the probability that the event happens is proportional to the time interval h being equal to the the frequency multiplied by the interval width. Moreover, the probability that more than one event occurs is negligible. It follows that for a Poisson process the probability of having n successes in a finite time interval Δt is given by the random variable $X(\Delta t)$ that has a Poisson distribution

$$P\{X(\Delta t) = n\} = \frac{(\lambda \Delta t)^n}{n!} e^{-\lambda \Delta t}; n = 0, 1, ... \tag{3}$$

To prove this, we can consider the approximation of the Binomial distribution for a number of trials $N \to \infty$ and a success probability $p \to 0$. The Binomial distribution is given by

$$Bin(n; N, p) = \frac{N!}{n!(N - n)!} p^n (1 - p)^{N-n} \tag{4}$$

and the expected value of n following a Binomial distribution is $E[n] = Np$. Imagine to divide the interval Δt in N infinitesimal subsets δt, so that $\frac{\Delta t}{\delta t} = N \gg 1$ we know that in each δt at most one event occurs. So, counting the number of events in Δt is equivalent to making N trials in which p is given by the probability of a success in δt, that for a Poisson process is $p = \lambda \delta t$, as indicated by property iv. Then, the average number of successes is

$$NP = N\lambda \delta t = \frac{\Delta t}{\delta t} \lambda \delta t = \lambda \Delta t. \tag{5}$$

Moreover, δt is an infinitesimal interval, then $p = \lambda \delta t \ll 1$, namely the success probability is small and we expect to have few successes in δt. Consequently $n \ll N$ and $\frac{N!}{(N-n)!} = N(N - 1)(N - 2) \dots (N - n + 1) \approx N^n$. Moreover, we can approximate $log((1 - p)^{N-n}) =$

$(N-n)log(1-p) \xrightarrow[p \ll 1]{} -p(N-n) \xrightarrow[n \ll N]{} -Np$, so that $(1-p)^{N-n} \to e^{-Np}$. Finally, $Bin(n;N,p) \xrightarrow[N\to\infty;p\to 0]{} \frac{(Np)^n e^{-Np}}{n!}$ and remembering that $Np = \lambda\Delta t$ we obtain

$$P\{X(\Delta t) = n\} = Bin(n; N \to \infty, p \to 0) = \frac{(\lambda\Delta t)^n e^{-\lambda\Delta t}}{n!} \tag{6}$$

that is the notorious Poisson distribution.

0.1.4 Inhomogeneous Poisson process

If the probability of having an event varies with time, i.e. $\lambda = \lambda(t)$, then property iii does not hold anymore and the number of events happening in an interval $[t, t+\Delta t]$ actually depends on the time point t. In this case, $\{X(t)\}$ is called inhomogeneous Poisson process and the probability of having n events in the time interval $[s,t]$ is given by

$$P\{X(t) - X(s) = n\} = \frac{1}{n!}\left[\int_s^t \lambda(t')dt'\right]^n e^{-\int_s^t \lambda(t')dt'}. \tag{7}$$

The integral $m(s=0,t) = \int_0^t \lambda(t')dt'$ is called expected value function and the random variable $X(t) - X(s)$ follows now a Poisson distribution with expected value $m(s,t)$, $0 \le s \le t$.

0.1.5 Birth death process

A natural generalization of the Poisson process is to permit the chance of an event occurring at a given instant of time to depend upon the number of events which have already occurred. This is the case of the reproduction of living organisms, in which one could suppose that the probability of a birth at a given instant is directly proportional to the population size (abundance) at that time. Analogously, we may assume that the population abundance $X(t)$ also decrease proportionally to the current abundances by the death of members. To describe this process, we assume that $X(t)$ is a Markov process on the states 0, 1, 2, ... and that its transition probabilities $P_{ij}(t)$ are stationary, i.e. $P_{ij}(t) = P\{X(t+s) = j | X(s) = i\}$. In addition we assume that $P_{ij}(t)$ satisfy

1. $P_{i,i+1}(h) = b_i h + o(h)$ as $h \downarrow 0$, $i \ge 0$;
2. $P_{i,i-1}(h) = d_i h + o(h)$ as $h \downarrow 0$, $i \ge 1$;
3. $P_{i,i}(h) = 1 - (b_i + d_i)h + o(h)$ as $h \downarrow 0$, $i \ge 0$;
4. $P_{ij}(0) = \delta_{ij}$;
5. $d_0 = 0$, $b_0 > 0$, $d_i, b_i > 0$, $i = 1, 2, ...$

The parameters b_i and d_i are called, respectively, the infinitesimal birth and death rates. In the first two postulates we are assuming that if the process starts in state i, then in a small interval of time the probabilities of the population increasing or decreasing by 1 are essentially proportional to the length of the interval.
Since the $P_{ij}(t)$ are probabilities, we have $P_{ij}(t) \ge 0$ and $\sum_{j=0}^{\infty} P_{ij}(t) = 1$. Moreover, using the Markovian property of the process we may also derive the Chapman-Kolmogorov equation

$$P_{ij}(t+s) = \sum_{k=0}^{\infty} P_{ik}(t) P_{kj}(s). \tag{8}$$

This equation states that in order to move from state i to state j in time $t+s$, $X(t)$ moves to some state k in time t and then from k to j in the remaining time s. As a consequence, the probability of moving from i to j in $t+s$ is given by the sum of the probabilities of all the possible paths from i to j through all possible k.

0.2 Master Equation

The Chapman-Kolmogorov equation is not easy to handle in actual applications. The master equation is a more convenient version of the same equation. It is a differential equation obtained by going to the limit of vanishing time difference s. For this purpose, we first show how the transition probability $P_{ij}(s)$ behaves when s tends to zero. The first order Taylor expansion of $P_{ij}(s)$ for $s \to 0$ is

$$\begin{aligned} P_{ij}(s) &= P_{ij}(0) + P'_{ij}(s) \cdot (s-0) + o((s-0)^2) \\ &= \left(1 - s \cdot \sum_{k \neq i} W_{ik}\right) \delta_{ij} + s \cdot W_{ij} + o(s^2), \end{aligned} \tag{9}$$

where W_{ij} is the transition probability per unit time from i to j and $(1 - s \cdot \sum_{k \neq i} W_{ik})$ is the probability that no transition takes place during s.
Substituting this expression for $P_{ij}(s)$, the Chapman-Kolmogorov equation becomes

$$\begin{aligned} P_{ij}(t+s) &= \sum_{k=0}^{\infty} P_{ik}(t) \left[\left(1 - s \cdot \sum_{j'} W_{kj'}\right) \delta_{kj} + s \cdot W_{kj} \right] \\ &= \left(1 - s \cdot \sum_{j'} W_{jj'}\right) P_{ij}(t) + s \sum_{k=0}^{\infty} P_{ik}(t) W_{kj} \end{aligned} \tag{10}$$

Dividing by s and going to the limit $s \to 0$, we obtain

$$\frac{dP_{ij}(t)}{dt} = \sum_{k=0}^{\infty} \left[W_{jk} P_{ij}(t) - W_{kj} P_{ik}(t) \right]. \tag{11}$$

This is the differential form of the Chapman-Kolmogorov equation and is called master equation. The equation describes the variation of the probability of the system to be in a particular state, due to incoming and outgoing fluxes and it is usually written in the simplified form

$$\frac{dP_n(t)}{dt} = \sum_{n} \left[W_{nn'} P_{n'}(t) - W_{n'n} P_n(t) \right] \tag{12}$$

or, for a process $\{X(t)\}$ defined in a continuous state space,

$$\frac{\partial P(x,t)}{\partial t} = \int \left[W_{x'x} P(x',t) - W_{xx'} P(x,t) \right] dx'. \tag{13}$$

0.2.1 Detailed Balance and stationary solution

The Master equation 12 (or 13) describes the variation of the probability of the system with time. It is clear then that the stationary solution, that is the probability distribution of the process when it is stationary (see Sec. 0.1.2), may be found simply setting $dP/dt = 0$. However, in the general case, the master equation is impossible to solve, being a huge (or infinite) set of degenerate differential equations. Analytical or numerical solution can be

found for some special cases, such as, for instance, for the one-step birth death process (see Sec. 0.1.5). Here, the stationary solution can be easily obtained with a linear expansion of the coefficient b_n and d_n. The Master equation for the birth death process is given by

$$\frac{dP_n(t)}{dt} = b_{n-1}P_{n-1}(t) + d_{n+1}P_{n+1}(t) - (b_n + d_n)P_n(t), \tag{14}$$

and, in the stationary state, the flux should be 0 for each step, meaning that $P_n b_n = P_{n+1}d_{n+1}$. This condition is called detailed balance and, when it holds, the stationary solution can be found very easily writing the recursive solution

$$P_{n+1} = P_n \frac{b_n}{d_{n+1}}. \tag{15}$$

and expanding it, so that to obtain

$$P_n = P_0 \prod_{i=0}^{n-1} \frac{b_i}{d_{i+1}} \tag{16}$$

where P_0 can be found from the normalization condition.

0.3 Diffusion process

Most Markov processes, including the Poisson process, birth-death processes, etc., satisfy the property

$$\lim_{h\downarrow 0} \frac{1}{h} P\{|X(h) - x| > \epsilon | X(0) = x\} = \lambda(x, \epsilon) \tag{17}$$

with $\lambda(x, \epsilon)$ nonnegative and possibly positive for ϵ small. For a Poisson process, in particular, the properties iv to vi reported in Sec. 0.1.3 imply in fact that

$$\lim_{h\downarrow 0} \frac{1}{h} P\{X(h) - i = 1 | X(0) = i\} = \lambda \tag{18}$$

with $i = 0, 1, ...$ and λ indicating the mean rate of the occurrence of events. The sample realization of the Poisson process are discontinuous step functions having jumps of unit increase. In contrast to condition 17, a diffusion process has to satisfy

$$\lim_{h\downarrow 0} \frac{1}{h} P\{|X(h) - x| > \epsilon | X(0) = x\} = 0 \tag{19}$$

for every $\epsilon > 0$ and for all x in the state space I. This condition describes the fact that the sample paths of a diffusion process are continuous.
Diffusion processes are usually also characterized by two further conditions, that describe the mean and variance of the infinitesimal displacements. Let $\Delta_h X(t) = X(t + h) - X(t)$ be the increment in the process accumulated over a time interval of length h. These key conditions affirm the existence of the limits

$$\lim_{h\downarrow 0} \frac{1}{h} E[\Delta_h X(t) | X(t) = x] = \mu(x, t) \tag{20}$$

and

$$\lim_{h\downarrow 0} \frac{1}{h} E[\{\Delta_h X(t)\}^2 | X(t) = x] = \sigma^2(x, t). \tag{21}$$

The functions $\mu(x,t)$ and $\sigma^2(x,t)$ are termed infinitesimal parameters of the process and respectively represent a drift and a diffusion component. The motivation for the name infinitesimal mean and variance for $\mu(x)$ and $\sigma^2(x)$ is clear, since

$$\begin{aligned} E[\Delta_h X(t)|X(t)=x] &= \mu(x)\cdot h + o(h) \\ E[(\Delta_h X(t))^2|X(t)=x] &= \sigma^2(x)\cdot h + o(h) \end{aligned} \tag{22}$$

Finally, for what concerns higher order infinitesimal parameters, the following relations are usually satisfied:

$$\lim_{h\downarrow 0} \frac{E[|\Delta_h X(t)|^r|X(t)=x]}{h} = 0,\ r=3,4,... \tag{23}$$

0.3.1 Elementary transformation of Processes

We may want to transform an arbitrary stochastic process $\{X(t)\}$ into a new process defined by $Y(t)=g(X(t))$, where g is a continuous strictly increasing function. If $\{X(t)\}$ is a continuous path Markov process, i.e. a diffusion, then, since g is continuous and monotone, so is $\{Y(t)\}$. Thus, if $\{X(t)\}$ has infinitesimal parameters $\mu(x)$ and $\sigma^2(x)$ and g has also two continuous derivatives g' and g'', then $\{Y(t)\}$ will also have infinitesimal parameters, that are determined by

$$\begin{aligned} \mu_Y(y) &= \frac{1}{2}\sigma^2(x)g''(x) + \mu(x)g'(x) \\ \sigma_Y^2(y) &= \sigma^2(x)[g'(x)]^2, \end{aligned} \tag{24}$$

where $y=g(x)$.

Proof. We consider only the case where g is strictly increasing. The strictly decreasing case is similar. For g twice continuously differentiable, the Taylor expansion with Lagrange remainder furnishes the representation

$$g(x+\Delta x) = g(x) + \Delta x g'(x) + \frac{1}{2}(\Delta x)^2 g''(x) + \frac{1}{2}(\Delta x)^2[g''(\xi) - g''(x)] \tag{25}$$

with $x \le \xi \le x+\Delta x$. If we substitute $X(t)=x$ and $\Delta X = X(t+h)-X(t)$, we have

$$g(X(t+h)) = g(X(t)) + \Delta X g'(X(t)) + \frac{1}{2}(\Delta X)^2 g''(X(t)) + \frac{1}{2}(\Delta X)^2[g''(\xi(w)) - g''(X(t))] \tag{26}$$

where $\xi(w)$ lies between $X(t)$ and $X(t+h)$. We can rewrite this equation substituting $Y(t)=g(X(t))$, and it becomes

$$Y(t+h) - Y(t) = \Delta X g'(X(t)) + \frac{1}{2}(\Delta X)^2 g''(X(t)) + \frac{1}{2}(\Delta X)^2[g''(\xi(w)) - g''(X(t))]. \tag{27}$$

Then, dividing by h, remembering that $\Delta X = X(t+h)-X(t)$, and introducing the definition of $\mu(x)$ and $\sigma^2(x)$, we obtain

$$\begin{aligned} \mu_Y(y) = &\lim h\downarrow 0 \frac{1}{h} E[Y(t+h) - T(t)|y(T)=Y] = \\ &\mu(x)g'(x) + \frac{1}{2}\sigma^2(x)g''(x) + \frac{1}{2}\lim_{h\downarrow 0}\frac{1}{h}E[(\Delta X)^2(g''(\xi(w)) - g''(X(t)))]. \end{aligned} \tag{28}$$

Since we assumed that g'' is twice continuous, $g''(\xi(w))$ converges to $g''(X(t))$ and this leads the last limit to tend to zero, proving the transformation for the infinitesimal mean $\mu_Y(y)$.

The infinitesimal variance of $Y(t)$ is found with a similar procedure. We square Eq. 27 and obtain

$$[Y(t+h) - Y(t)]^2 = (\Delta X)^2 [g'(X(t))]^2 + R_h \tag{29}$$

with R_h that contains only terms of order $(\Delta X)^3$ or higher. We know from Eq. 23 that for $r \geq 3$ $\lim_{h \downarrow 0} \frac{1}{h} E[|\Delta X|^r | X(t) = x] = 0$, and it follows that

$$\sigma_Y^2(y) = \lim_{h \downarrow 0} \frac{1}{h} E[(Y(t+h) - Y(t))^2 | Y(t) = y] = \sigma^2(x)[g'(x)]^2 \tag{30}$$

and also the transformation for the infinitesimal variance $\sigma_Y^2(y)$ is proven.

0.4 Fokker Planck equation

Planck derived the Fokker-Planck equation as an approximation to the continuous master equation (Eq. 13) in the following way. First express the transition probability W as a function of the jump size r and of the starting point x'

$$W_{x'x} = W(x'; r) \tag{31}$$

with $r = x - x'$. The general master equation then becomes

$$\frac{\partial P(x,t)}{\partial t} = \int \{W(x-r; r)P(x-r, t) - W(x; -r)P(x,t)\} dr \tag{32}$$

We assume that only small jumps occur and consequently that there exists a $\delta > 0$ such that

$$\begin{aligned} &W(x', r) \approx 0 && \text{for } |r| > \delta \\ &W(x' + \Delta x; r) \approx W(x'; r) && \text{for } |\Delta x| < \delta. \end{aligned} \tag{33}$$

Secondly, we assume that the solution $P(x,t)$ varies slowly with x, meaning that there exists a $\delta > 0$ such that

$$P(x' + \Delta x, t) \approx P(x', t) \text{ for } |\Delta x| < \delta. \tag{34}$$

Then, it is possible to handle the shift from x to $x - r$ in the first integral in Eq. 32 by means of a Taylor expansion up to the second order, so that

$$\begin{aligned} \frac{\partial P(x,t)}{\partial t} = &\int W(x; r)P(x,t)dr - \int r \frac{\partial}{\partial x}\{W(x; r)P(x,t)\}dr + \\ &\frac{1}{2}\int r^2 \frac{\partial^2}{\partial x^2}\{W(x; r)P(x,t)\}dr - \int W(x; -r)P(x,t)dr \end{aligned} \tag{35}$$

The first and fourth terms cancel and the other two terms can be rewritten introducing the jump moments

$$a_\nu(x) = \int_{-\infty}^{+\infty} r^\nu W(x; r)dr = \int_{-\infty}^{+\infty} (x - x')^\nu W(x; r) = E[(x - x')^\nu] \tag{36}$$

so that the master equation approximation becomes

$$\frac{\partial P(x,t)}{\partial t} = -\frac{\partial}{\partial x}\{a_1(x)P(x,t)\} + \frac{1}{2}\frac{\partial^2}{\partial x^2}\{a_2(x)P(x,t)\}. \tag{37}$$

This equation is called Fokker-Planck equation. It is an approximation of the master equation. The terms $a_1(x)$ and $a_2(x)$ correspond to the mean and variance of the jump

$r = x - x'$. Note that for a diffusion process these correspond to $a_1(x) = \mu(x)h$ and $a_2(x) = \sigma^2(x)h$ and go to zero for infinitesimal jumps h. The space state of x here has to be continuous and the equation can be interpreted as a continuity equation

$$\frac{\partial P(x,t)}{\partial t} = -\frac{\partial J(x,t)}{\partial x}, \tag{38}$$

where the probability flux $J(x,t)$ is given by a drift and a diffusion term

$$J(x,t) = a_1(x)P(x,t) - \frac{1}{2}\frac{\partial}{\partial x}a_2(x)P(x,t). \tag{39}$$

0.4.1 The Langevin approach

Suppose we have a system whose macroscopic behavior is known and we also know that there must be fluctuations. This may be the case, for instance, of a Brownian particle, that is a heavy particle immersed in a fluid of light molecules that randomly collide with it, making its velocity varying by a large number of (uncorrelated) jumps. The Langevin approach to describe such system is the following.

1. Write the deterministic macroscopic equations of motion of the system. For the velocity of the Brownian particle, this would be $\frac{dV}{dt} = -\gamma V$, where γ is a constant coefficient.

2. Add a noise term in the form of an external force $L(t)$, that represents the effect of the molecules of the surrounding fluid. $L(t)$ is called white noise and has the following properties:

   (a) it has null average: $\langle L(t)\rangle = 0$.

   (b) its autocorrelation function is $\langle L(t)L(t')\rangle = \Gamma\delta(t-t')$, meaning that each collision with the molecules of the surrounding fluid is practically instantaneous and that successive collisions are uncorrelated.

   (c) $L(t)$ is Gaussian.

3. Γ describes the mean square fluctuations, as we prove in the following. So, adjust Γ so that the stationary solution reproduces the correct mean square fluctuations.

Let us consider a generic physical system with an equation of motion $\dot{x} = A(x)$. Following the Langevin approach, we add a white noise term to describe fluctuations and obtain

$$\dot{x} = A(x) + L(t). \tag{40}$$

Its solution is equivalent to the Fokker-Planck equation

$$\frac{\partial P(x,t)}{\partial t} = -\frac{\partial}{\partial x}A(x)P(x,t) + \frac{\Gamma}{2}\frac{\partial^2 P(x,t)}{\partial x^2}. \tag{41}$$

In fact, for each function L, Eq. 40 uniquely determines $x(t)$ when $x(0)$ is given. Since the values of L at different time are stochastically independent, it follows that x is Markovian. Hence, it obeys the master equation. We can compute the jump moments for x following Eq. 40.

$$\Delta x = \int_t^{t+\Delta t} A(x(t'))dt' + \int_t^{t+\Delta t} L(t')dt' \tag{42}$$

and the average will be given by

$$a_1(x) = \langle \Delta x\rangle = A(x(t))\Delta t + O(\Delta t)^2. \tag{43}$$

Analogously,

$$\begin{aligned}\langle(\Delta x)^2\rangle =&\langle\left(\int_t^{t+\Delta t} A(x(t'))^2\right)^2\rangle \\ &+2\int_t^{t+\Delta t}dt'\int_t^{t+\Delta t}dt''\langle A(x(t'))L(t'')\rangle \\ &+\int_t^{t+\Delta t}dt'\int_t^{t+\Delta t}dt''\langle L(t')L(t'')\rangle. \end{aligned} \tag{44}$$

The first term is of order $(\Delta t)^2$ and therefore does not contribute to $a_2(x)$. The last term equals $\Gamma\Delta t$ according to the property 2b. Finally, the second term can be rewritten expanding $A(x(t'))$ as

$$2A(x(t))\Delta t\int_t^{t+\Delta t}dt''\langle L(t'')\rangle+2A'(x(t))\int_t^{t+\Delta t}dt'\int_t^{t+\Delta t}dt''\langle(x(t')-x(t))L(t'')\rangle+\cdots\approx o(\Delta t). \tag{45}$$

So, we proved that $a_2(x)=\Gamma\Delta t$.
Finally, if we consider a nonlinear Langevin equation

$$\dot{x}=A(x)+C(x)L(t), \tag{46}$$

we can reduce it to the preceding case dividing by $C(x)$ and transforming $x\to\tilde{x}$, with $\tilde{x}=\int\frac{dx}{C(x)}$:

$$\begin{aligned} d\tilde{x} &= \frac{dx}{C(x)} \\ d^2\tilde{x} &= \frac{d^2(x)C(x)-C'(x)dx}{C^2(x)} \\ \tilde{A}(\tilde{x}) &= \frac{A(x)}{C(x)} \\ \tilde{P}(\tilde{x}) &= P(x)C(x). \end{aligned} \tag{47}$$

The equivalent Fokker-Planck equation is then

$$\frac{\partial\tilde{P}(\tilde{x},t)}{\partial t}=-\frac{\partial}{\partial\tilde{x}}\tilde{A}(\tilde{x})\tilde{P}(\tilde{x},t)+\frac{\Gamma}{2}\frac{\partial^2\tilde{P}(\tilde{x},t)}{\partial\tilde{x}^2}. \tag{48}$$

and transforming back to the original x

$$\frac{\partial P(x,t)}{\partial t}=-\frac{\partial}{\partial x}[A(x)+\frac{1}{2}\Gamma C(x)C'(x)]P(x,t)+\frac{\Gamma}{2}\frac{\partial^2}{\partial x^2}[C(x)]^2P(x,t) \tag{49}$$

and $a_1=[A(x)+\frac{1}{2}\Gamma C(x)C'(x)]\Delta t$ and $a_2=\Gamma[C(x)]^2\Delta t$.
Analogously, we could choose to directly estimate a_1 and a_2 using the same approach of Eq. 43-44.
Note that there is a problem concerning Eq. 46. In fact, $L(t)$ can be visualized as a sequence of delta peaks arriving at random times. If we consider one of these times $\check{t}$, according to Eq. 46, when $t=\check{t}$, the delta function in $L(\check{t})=\infty$ causes a jump in $x(\check{t})$ that also goes to infinity. Hence, the value of x at the time that the delta function arrives is undetermined, and therefore also the value of $C(x)$. The equation does not specify whether one should insert in $C(x)$ the value of x before the jump, after the jump or perhaps the mean of both. Stratonovich opted for the mean value and rewrote Eq. 46 as

$$x(t+\Delta t)-x(t)=A(x(t))\Delta t+C\left(\frac{x(t)+x(t+\Delta t)}{2}\right)\int_t^{t+\Delta t}L(t')dt'. \tag{50}$$

This choice leads to Eq. 49. First let us rewrite C as

$$C\left(\frac{x(t)+x(t+\Delta t)}{2}\right)=C\left(\frac{x(t)+x(t+\Delta t)+x(t)-x(t)}{2}\right)=C\left(x(t)+\frac{\Delta x}{2}\right). \quad (51)$$

Then, assuming that C is differentiable, we expand it as

$$C\left(x(t)+\frac{\Delta x}{2}\right)\approx C\left(x(t)\right)+\frac{\Delta x}{2}\frac{dC(x(t))}{dx}+... \quad (52)$$

It follows that the second term of Eq. 50 is

$$\int_t^{t+\Delta t} C\left(x(t)+\frac{\Delta x}{2}\right)L(t')dt'=C\left(x(t)\right)\int_t^{t+\Delta t}L(t')dt'+\frac{dC(x(t))}{dx}\int_t^{t+\Delta t}\frac{\Delta x}{2}L(t')dt' \quad (53)$$

The first integral is zero according to the property 2a of the Langevin approach. The second one can be rewritten substituting $\Delta X=\int_t^{t+\Delta t}A(x(t'))dt'+\int_t^{t+\Delta t}(C\cdot L)dt'$, so that it becomes

$$\begin{aligned}&\frac{dC(x(t))}{dx}\int_t^{t+\Delta t}\frac{\Delta x}{2}L(t')dt'=\\&=\frac{1}{2}\frac{dC(x(t))}{dx}\int_t^{t+\Delta t}dt'\left[\int_{t'}^{t'+\Delta t'}A(x(t''))L(t'')L(t'')dt''+\int_{t'}^{t'+\Delta t'}C(x(t))L(t'')L(t'')dt''\right].\end{aligned} \quad (54)$$

The first integral addend is $o(\Delta t)$ as in Eq. 45. Hence,

$$\begin{aligned}\int_t^{t+\Delta t}C\left(x(t)+\frac{\Delta x}{2}\right)L(t')dt'=&\frac{1}{2}\frac{dC(x(t))}{dx}C(x(t))\int_t^{t+\Delta t}dt'\int_{t'}^{t'+\Delta t'}L(t'')L(t'')dt''=\\&\frac{1}{2}C(x(t))\frac{dC(x(t))}{dx}\Gamma\end{aligned} \quad (55)$$

and we find

$$\langle\Delta x\rangle\approx A(x(t))\Delta t+\frac{1}{2}C(x(t))\frac{dC(x(t))}{dx}\Gamma\Delta t \quad (56)$$

and then follows Eq. 49, as anticipated.
Ito, instead, opted for the value of x before the arrival of the delta peak, so that

$$\int_t^{t+\Delta t}C(x(t'))L(t')dt'=C(x(t))\int_t^{t+\Delta t}L(t')dt'=0 \quad (57)$$

where the first equivalence follows from the Ito assumption according to which the value of C in the interval $[t,t+\Delta t]$ is equal to the value at the beginning of the interval $C(x(t))$, and the second equivalence comes from the property 2a for the white noise L. Consequently,

$$\langle\Delta x\rangle\approx A(x)\Delta t \quad (58)$$

and Eq. 49 is now equivalent to the Fokker-Planck equation

$$\frac{\partial P(x,t)}{\partial t}=-\frac{\partial}{\partial x}A(x)P(x,t)+\frac{\Gamma}{2}\frac{\partial^2}{\partial x^2}[C(x)]^2P(x,t). \quad (59)$$

Apparently, this interpretation is not compatible with the familiar way of transforming variables and in fact new transformation laws have been formulated to deal with the Ito interpretation.

Ecological Theory

0.5 Ecosystems and biodiversity

Ecology aims to study the interactions between organisms and their environment as an integrated system [8]. One of its main interests is biodiversity, which refers to the varieties of species, genes, and, on a wider scale, ecosystems. Biodiversity increases the ecosystem performances and productivity, so as its stability over time, through the supplement of functional traits variety [9]. The total biodiversity of an ecological landscape is called gamma diversity and is determined by two independent components, the mean species diversity in sites or habitats at local scale (α diversity) and the differentiation among those habitats (β diversity) [10]. When studying a single ecosystem, we are interested in particular in the α diversity, that implies species richness and allows an healthy degree of competition, preventing pathogenic or invasive species to prevail.

0.5.1 Diversity indices

Several definitions of α diversity exist according to what diversity itself is assumed to be. Diversity indices, for instance, characterize how many different species compose the ecosystem, taking simultaneously into account how evenly individuals are distributed among those species. Shannon [11] and Pielou's [12] indices are entropy measures that are maximized when all species have the same proportion of individuals, corresponding to maximum diversity. If the total number of species in the ecosystem is S and p_i is the fraction of individuals that belong to species i, Shannon's index is defined as

$$H = -\sum_{i=1}^{S} p_i log(p_i). \tag{60}$$

Pielou's index is the normalized version of Shannon's index, given by

$$J = \frac{H}{H_{max}}, \tag{61}$$

where H_{max} is the maximum value that H can have, that is when all species have the same proportion of individuals: $H_{max} = -\sum_{i=1}^{S} \frac{1}{S} log(\frac{1}{S})$. Simpson's index measures the degree of concentration when individuals are classified into species. It is the probability that two individuals taken at random from the dataset of interest belong to the same species [13] and is given by

$$\gamma = \sum_{i=1}^{S} p_i^2. \tag{62}$$

These indices, however, have been strongly criticized because of their sensitivity to the few commonest species [14]. For this reason, fitting a model to the data seems a better option, enabling the diversity statistics to be estimated much more efficiently than by direct evaluation of the relative frequencies [14].

0.5.2 Relative Species Abundance distribution (RSA)

Diversity evaluation can be obtained specifically by fitting the Relative Species Abundance distribution (RSA), that refers to how common or rare a species is relative to other species. As we will point out in the following, the usual way to represent the RSA is in the form of Preston plot, that is plotting how many species ($y-$axes) have a certain number of individuals ($x-$axis, usually in logarithm to base 2). RSA is closely related to Shannon entropy. Maximum diversity, for instance, is obtained when individuals are equally distributed into species, that is when entropy is maximal. In this situation, all species have the same proportion of individuals and RSA is peaked in one bin, that is the '1 individual' bin if the number of individuals is equal to the number of species. In this case, in fact, maximum diversity is obtain when each species has one individual and consequently the number of species with one individual (RSA's bin 1) is the total number of species.
RSA distribution, has widely fascinated ecologists, since, in an extensive number of ecological communities, ranging from open-ocean planktonic copepod, to tropical bat, but also in rain forest trees and British breeding birds, RSA has been shown to follow very similar patterns [5]. Moreover, RSA distributions appear to be drawn from a single family of distributions, extending from the Log Series to a highly skewed and unveiled Log Normal. This led to several efforts in modeling the RSA of ecosystems arising from statistical or dynamical clues, that may be classified in two major approaches: inductive and deductive. In the early years, when the study of relative species abundance was in its infancy, the inductive approach dominated. Observed distributions of the numbers of individuals per species in collections were fitted to statistical distributions with little or no attempt to give a theoretical explanation or to define the sampling universes from which the collections were made. More recently, instead, different attempt have been made in order to derive a theory of relative species abundance from first principles, i.e. based on hypotheses about how ecological communities were organized.

0.6 Inductive approaches

0.6.1 Fisher's Log-Series distribution

A milestone of the inductive approach was the work proposed by Fisher, Cobert and Williams in 1943 [15]. Fisher considered the number of individuals observed when repeatedly sampling a species as a counting process or, more precisely, a Poisson process (see Sec. 0.1.3). Then, under the assumption that the population was homogeneous, the distribution of the number of species with a certain number of individuals would be

$$P(n, \lambda) = \frac{e^{-\lambda}\lambda^n}{n!}. \tag{63}$$

This is the Poisson distribution, where n is the number of observed individuals in each sample and λ is the expected number of individuals $E[n] = \lambda$. Obviously, λ will depend on the sample size and on the density of individuals in the sampled species. Moreover, if all the species were equally frequent in the community, we would have obtained the same distribution if, rather than considering many samples of the same species, we would have considered a single sample from the whole community and counted the number of individuals in each species.
On the other hand, if the community was heterogeneous, meaning that different species are present with different frequencies, each species j would be characterized by a different λ_j and the distribution of the number of individuals in each species would be a mixture of

Poisson distribution. In this case, in fact, the counting process would be an inhomogeneous Poisson process (see Sec. 0.1.4).
Moreover, since the values of λ must be positive, Fisher made the simplest assumption and proposed that they were distributed according to a Gamma distribution with parameters p and k, so that

$$dP(\lambda) = \frac{1}{(k-1)!} p^{-k} \lambda^{k-1} e^{-\lambda/p} d\lambda. \tag{64}$$

The probability of observing n individuals is now given by

$$\begin{aligned} P(n) &= \int_0^\infty dP(\lambda) P(n,\lambda) d\lambda \\ &= \int_0^\infty \frac{1}{(k-1)!} p^{-k} \lambda^{k-1} e^{-\lambda/p} \frac{e^{-\lambda}\lambda^n}{n!} d\lambda \\ &= \frac{(k+n-1)!}{(k-1)!n!} \frac{p^n}{(1+p)^{k+n}} \end{aligned} \tag{65}$$

that is a Negative Binomial distribution.
When the parameter k is very high, we obtain again a Poisson distribution, while when the population is very heterogeneous, k becomes small and tends to zero and the Gamma distribution for λ becomes very skewed. In nature, the abundance of different species generally vary greatly, in fact a large number of species are so rare that their chance of inclusion is small and as a consequence $k \to 0$. However, $k = 0$ cannot be observed in reality because the total number of species has to be finite. Moreover, such limiting case could not occur if the zero abundance class was observable, because the distribution would then wholly consist of such cases. However, Fisher considered the case $k = 0$ for the truncated distribution valid for $n \geq 1$. He set $x \equiv p/(1+p)$ and replaced the constant factor in the denominator by a new constant factor, $1/\alpha = (k-1)!$. The predicted RSA distribution turned out to be the Log-Series, that is the limit case of the Negative Binomial distribution in case of highly heterogeneity and that is valid only for $n \geq 1$. The Log-Series RSA is given by

$$P_{RSA}(n) = \alpha \frac{x^n}{n};\ n \geq 1 \tag{66}$$

and Fisher proved that it fits well data from butterflies in Malaya and moths collected over a four-year period at the Rothamsted Experimental Station in England.
According to Fisher's model, the expected number of species with 1, 2, 3, 4, ..., n individuals is given by αx, $\alpha x^2/2$, $\alpha x^3/3$, $\alpha x^4/4$, ... , $\alpha x^n/n$ for $0 < x < 1$. Adding all terms, the total number of species, S, is expected to be $\alpha[-ln(1-x)]$, and the total number of individuals in the collection, N , is $\alpha x/(1-x)$. The parameter α is known as Fisher's α and, together with the total number of individuals N, completely summarizes the RSA distribution. Fisher's α is a widely used measure of species diversity because it is theoretically independent of sample size [15], even if for some datasets it turns out to be only approximately constant, changing slowly over large ranges in sample size [5].

0.6.2 Preston Log-Normal distribution

Fisher's α parameter has been generally used to estimate diversity. However, the Log-Series model has been also criticized, mainly for not being a good fit to the data, especially when increasing sample sizes. In 1948, Preston argued that RSA distributions were actually Log-Normal, partly due to the Central Limit Theorem, and that the Log-Series resulted in fact from under-sampling [16]. Analyzing for instance bird species abundances, Preston noted, in fact, that RSA were often bell-shaped curves, such that species having intermediate abundances were more frequent than very rare species. Due to the

Log-Normal behavior of the RSA, Preston proposed to represent it in with the $x-$axis (number of individuals) in logarithm to base 2, so that to visualize the normality of the logarithmic abundances. Such plot is called Preston plot and can be built categorizing the abundances in bins (called octaves) that are limited by the powers of 2 (1, 2, 4, 8, etc.). The Log-Normal distribution is continuous, not discrete as in the case of the Log-Series. However, Preston's method of categorizing abundances provides a simple way to approximate the distribution by a discrete-valued function, as follows. Let S_0 be the number of species in the modal octave of abundance. Let S_R be the number of species in the $R-$th octave (or doubling abundance class) to the left or right of the modal octave. Then, the so called Species Curve can be written as

$$S_R = S_0 e^{-a^2 R^2}, \tag{67}$$

with $R = 0, 1, 2, ...$ and where a is a constant that depends on the parameter σ of the Log-Normal, $a = 1/\sqrt{2\sigma}$.
Over the past half century, the Log-Normal distribution has been fitted successfully to a far larger number of relative species abundance distributions than has the Log-Series distribution, particularly as bigger sample sizes have become available [5]. To explain his Log-Normal distribution, Preston argued that the shape of the RSA observed by Fisher and his colleagues was an artifact of small sample size. In the Log-Series, the expected number of species is always larger in the rarest abundance category, consisting of singleton species. However, in a small sample, one should observe only a truncated distribution of relative abundances, comprising only the most common species. This is because common species are generally collected sooner than rare species. As sample size increases, Preston predicted that more and more of the Log-Normal distribution would be revealed and the reason for which Fisher had not noted it was that he did not consider the importance of sample size, because of the theoretically expected constancy of Fisher's α in collections of different sizes. However, the proposal for a Log-Normal distribution let the apparent invariance of Fisher's α unexplained and, furthermore, in recent years, as larger sample sizes of relative species abundance have become available and the abundances of very rare species have become better known, it has become increasingly apparent that observed distributions of relative species abundance are actually seldom log-normally distributed. Empirical RSA, instead, appear to be Log-Normal to the right of the mode in the right-hand tail representing common species. But they almost always show a strong negative skewness that cannot be explained neither with Fisher or with Preston's distribution.

0.7 Deductive approaches

After the initial inductive attempts, several efforts have been made to derive a theory of relative species abundance from first principles, that was based on hypotheses about how ecological communities were organized. The motivation to this approach was the idea that the RSA patterns were so ubiquitous that there had to be an underlying general mechanism that theory could elucidate.

0.7.1 Niche models and MacArthur broken stick

The first deductive theory for ecological system, was proposed by MacArthur in 1957 [17]. MacArthur hypothesized that groups of trophically similar species in ecological communities simply randomly divided up a common pool of limiting resource and their relative abundances were proportional to the fraction of total resource each utilized. The partitions in which the resource pool is subdivided are called 'niche' and these kind of models that determine the distribution of abundances of individuals among species based on how

species break up such pool are called niche apportionment models. MacArthur idealized the resource pool as a stick of unit length. Suppose a community of S species randomly divides up the common resource. Now randomly partition the resource pool by throwing $S-1$ random points onto the unit stick. Then, break the stick at each random point, and rank the fragments from shortest to longest. The expected relative abundance of the $i-$th rarest (shortest) species, y_i, should then be given by

$$E[y_i] = \frac{1}{S}\sum_{x=i}^{S}\frac{1}{x}. \tag{68}$$

An extension to this formulation considers a broken-stick model in which the partition of the limiting resource is nonrandom and was proposed independently by Motomura [18] and Whittaker [19]. While MacArthur's model assumed the population to be homogeneous, Motomura and Whittaker allowed for heterogeneity, considering that the community may be characterized by some hierarchical structure. The expected RSA was found applying the following rule. Let the most dominant species take over a fraction k of the total resource pool, leaving the fraction $1-k$ for all other species. Then, let the second most dominant species sequester the same fraction k of the remaining resource, leaving the fraction $(1-k)^2$ for all remaining species, and so on.
Sugihara [20] further improved the broken-stick model noting that repeatedly or sequentially breaking the broken stick would eventually produce a Log-Normal distribution of fragment lengths. He thus proposed the following sequential breakage rule. Take the stick and make the first random break. Choose one of the two fragments at random and break it randomly. Then choose one of the three fragments at random and break it, and so on. What Sugihara discovered was that the resulting distribution was not only Log-Normal, but it was the distribution predicted by Preston (Eq. 67). However, this model has several shortcomings, such as, for instance, the lack of interpretation for its supposed first principles [5]: the biological analog to what is done mathematically in sequential breakage is not clear and, moreover, there is no 'stopping rule' inherent in the theory that fixes how many sequential breaks to carry out. This means that the number of species in the community is a free parameter that does not follow from the theory. Finally, broken stick models can also be faulted for having little or nothing to say about sampling issues or how they might be tested with data from real communities.

0.7.2 MacArthur and Wilson model and the neutrality assumption

All the early deductive approaches belonged to the niche perspective and, going after Lotka-Volterra's idea of coexistence as a static equilibrium, did not take into account the dynamical processes that may have generated a particular RSA. The first deductive model that was also based on the idea of a dynamical equilibrium was proposed by MacArthur and Wilson in 1967 [21]. MacArthur and Wilson were the first to introduce the concept of neutrality in ecology. Neutrality can be considered as a null hypothesis to niche theory and implies that, at a given trophic level in a food web, all species are equivalent in their birth, death and dispersal rates, when measured on a per-capita basis [22].
MacArthur and Wilson noted that with the current static idea that considered island communities fixed over ecological time-scale ($\sim 10^3$ years), it could not be explained why islands nearly always have fewer species than areas on continents of the same size. The authors proposed a new theory in which the number of species on the island could change as a result of two opposing forces: immigration from the continents of species not already present on the island, and extinction of species present on the island.
Furthermore, once island populations went extinct, it would take the same species longer to recolonize the island than it would take them to disperse among adjacent areas on the

mainland. Thus, other things being equal, species would spend a smaller fraction of total time resident on a given island than in the same-sized area of the mainland. Given these assumptions, the equation that controls the species dynamics is

$$\frac{dS}{dt} = I(S) - E(S), \tag{69}$$

where $I(S)$ is the immigration rate of new species and $E(S)$ is the extinction rate. The stationary state is found setting $dS/dt = 0$ and is achieved when the immigration and extinction rates become equal. Note that, at the steady state, the immigration and extinction rates are equal but not null. What remains steady is the number of species on the island S and not their identity. For this reason, this model is called neutral, in the sense that all species are considered as equivalent and have the same dynamic rates.

0.7.3 Caswell's abundance random walk

Following the idea of neutrality, Caswell erected his model, considering communities as collections of completely noninteracting species in which each species undergoes an independent random walk in abundance. Therefore, the total size of the community fluctuates. New species enter the community as a Poisson process (see Sec. 0.1.3) with probability λ per unit time. This immigration probability, as in the theory of island biogeography, is independent of the identity of the species and of the number and identities of the species already present, except that only species not currently present are allowed to immigrate. This is equivalent to assuming that immigration makes a negligible contribution to the population dynamics of a species already present. Caswell assumed a linear birth-death process in which the stochastic per capita birth and death rates, b and d, are equal: $b = d$. This is a pure drift process or random walk. The transition probabilities from a population of size N_i to size N_{i-1}, N_i , or N_{i+1} at time $t + dt$ are linear functions N_i of at time t, as follows:

$$\begin{aligned} P(N_{i-1}|N_i) &= d \cdot N_i \\ P(N_i|N_i) &= 1 - (b + d) \cdot N_i \\ P(N_{i+1}|N_i) &= b \cdot N_i \end{aligned} \tag{70}$$

Caswell's model is very important because it was the first one to be explicitly based on birth, death, and dispersal processes. However, it has several problems. First, its results differ substantially from observed community relative abundance patterns and look decidedly not lognormal on a Preston plot of octaves of abundance. Then, the size of the community J grows without bound over time, in fact, J turns out to be a Negative Binomial random variable with mean $E[J] = t \to \infty$ (elapsed time), and variance $Var[J] = t(t + 1) \to \infty$ as $t \to \infty$. Finally, the expected number of species in the community, $E[S]$, is linearly proportional to the colonization rate of new species per unit time, λ, and the logarithm of the elapsed time:

$$E[S] = Var[S] = \nu \cdot ln(t + 1). \tag{71}$$

Caswell's model can be improved with the addition of the assumption of a finite community size (due to limited resource availability) and minor changes in the birth, death, and dispersal processes.

0.7.4 Hubbell's Unified Neutral Theory of Biodiversity

Hubbell developed his Unified Neutral Theory of Biodiversity inspired by the ideas of neutrality and dynamic equilibrium of MacArthur and Wilson. He observed that, in general,

population densities are constant, meaning that large landscapes are always saturated with individuals, and he thus treated communities as being formed by a fixed number of individuals, usually denoted by J. More precisely, Hubbell distinguished between a dispersal-limited local community of size J and a so-called metacommunity from which species can (re)immigrate and which acts as a heat bath to the local community. The distribution of species in the metacommunity was given by a dynamic equilibrium of speciation and extinction, as in MacArthur and Wilson's model. Both community dynamics were modeled by appropriate urn processes, where each individual is represented by a ball with a color corresponding to its species. To build the metacommunity distribution, imagine to choose randomly and with a certain rate some individuals that will reproduce. For each chosen individual, this corresponds to add a further ball of its own color to the urn. Since one basic assumption is saturation, reproduction has to happen at the cost of removing another random individual from the urn. At a different rate, single individuals in the metacommunity are replaced by elements of an entirely new species.
The urn scheme for the metacommunity of J_M individuals is the following. At each time step take one of the two possible actions:

- With probability $1 - \nu$ draw an individual at random and replace another random individual from the urn with a copy of the first one.
- With probability ν draw an individual and replace it with an individual of a new species.

The urn scheme for the local community of fixed size J is very similar to the one for the metacommunity. At each time step take one of the two actions:

- With probability $(1-m)$ draw an individual at random and replace another random individual from the urn with a copy of the first one.
- With probability m replace a random individual with an immigrant drawn from the metacommunity.

The metacommunity is changing on a much larger timescale and is assumed to be fixed during the evolution of the local community. The resulting distribution of species in the local community and expected values depend on four parameters, J, J_M, θ and m, that is a dispersal parameter. When $m = 1$, that is in the limit of no dispersion, the local community is just a sample from the metacommunity. This is the only one for which Hubbell [5] presented analytical results. If $m = 0$, instead, the local community is completely isolated from the metacommunity and all species will go extinct except one. Finally, if dispersal limitation is present ($0 < m < 1$), we have an intermediate state between domination of the most common species and a sampling from the metacommunity, where singleton species are most abundant. This is the biologically more interesting situation and is characterized by a unimodal species distribution in a Preston plot, often fitted by a Log-Normal distribution. Hubbell provided only numerical results for this case. Analytical results, instead, have been found successively following two different approaches classified as forwards- and backwards-in-time [23].
The forwards-in-time perspective uses a master equation approach with a Markovian description of states and transitions. This approach is the one followed by Volkov [24] as we will detail in the following, and has in general resulted in exact analytical expressions and various approximations for the 'expected number of species with a certain abundance' in a sample of J individuals from a dispersal-limited local community. Note that the expected number of species with a certain abundance is the classical approach to study commonness and rarity in community ecology and also a very useful tool in exploring the behavior of

community models. However, it cannot be used to obtain accurate estimates of the model parameters.
The backwards-in-time perspective takes a genealogical, coalescent-type approach where community members are traced back to the ancestors that once immigrated into the community. This approach produces analytical expressions for the 'joint multivariate probability of observing S species with abundances n_1, n_2,...,n_S' in a sample of J individuals from the local community. If we denote this collection by $\vec{D}$, i.e. $\vec{D} = (n_1, n_2, ..., n_S)$, the joint multivariate probability is $P(\vec{D}|\theta, m, J)$ and can be used in maximum likelihood estimation of model parameters from species-abundance data or other methods based on the likelihood, but is less useful for studying the behavior of the model [23].
For the case $m = 1$, the species abundance distribution was found by Hubbell himself following Ewens's sampling formula [25][26] and is given by

$$P(\vec{D}|\theta, m, J) = \frac{J!}{\prod_{i=1}^{S} n_i \prod_{j=1}^{J} \Phi_j!} \frac{\theta^S}{(\theta)_J}, \tag{72}$$

where Φ_j is the observed number of species with abundance j and $(\theta)_J$ is the Pochhammer symbol defined as

$$(\theta)_J = \prod_{i=1}^{J} (\theta + i - 1) = \frac{\Gamma(\theta + J)}{\Gamma(\theta)} \tag{73}$$

The expected number of species in the metacommunity having exactly n individuals, instead, was found to be [27]

$$E[S_M(n)] = \frac{\theta}{n} \frac{\Gamma(J_M + 1)\Gamma(J_M + \theta - n)}{\Gamma(J_M + 1 - n)\Gamma(J_M + \theta)} \tag{74}$$

where $\theta = (J_M - 1)\nu/(1 - \nu) \approx J_M \nu$ is called fundamental biodiversity number. For large metacommunities and $n \ll J_M$ one recovers the Fisher Log-Series distribution

$$E[S_M(n)] \approx \frac{\theta}{n} \left(\frac{J_M}{J_M + \theta} \right)^n. \tag{75}$$

The fundamental biodiversity number θ is thus asymptotically identical to Fisher's α. In particular, when θ is small, the expected RSA is Log-Series-like, while, as θ becomes larger, it becomes more Log-Normal-like. At infinite diversity, in the limit $\theta \to \infty$, every individual sampled represents a new and different species, regardless of how large a sample is taken. At the other extreme, when $\theta = 0$, the distribution collapses to a single monodominant species throughout the metacommunity.
With dispersal limitation ($m < 1$), the joint multivariate probability distribution is derived in [28] based on the Hypergeometric distribution, and turned out to be

$$P(\vec{D}|\theta, m, J) = \frac{J!}{\prod_{i=1}^{S} n_i \prod_{j=1}^{J} \Phi_j!} \frac{\theta^S}{(I)_J} \sum_{A=S}^{J} K(\vec{D}, A) \frac{I^A}{(\theta)_A}, \tag{76}$$

where $I = (J - 1)m/(1 - m)$ and $K(\vec{D}, A)$ are coefficients fully determined by the data. The expected number of species, instead was derived in [27] and is given by

$$E[S(n)] = \frac{\theta}{(I)_J} \binom{J}{n} \int_0^1 (Ix)_n [I(1 - x)]_{J-n} \frac{(1 - x)^{\theta - 1}}{x} dx. \tag{77}$$

0.7.5 Volkov's Negative Binomial

In 2007, Volkov, Hubbell and others [24] derived a much simpler analytical stationary solution for the birth-death process with density dependence (demographic noise), following a master equation approach. In agreement with Hubbell's neutrality assumption, they considered all species as equivalent, meaning that they all have the same dynamic rates. Moreover, inter-species interaction is neglected and the community is supposed to have reached its steady state.
From a deterministic prospective, the number of individuals n of a given species evolves according to the differential equation

$$\frac{dn}{dt} = b_n \cdot n - d_n \cdot n + S, \tag{78}$$

where b_n and d_n denote the per-capita density-independent birth and death rates, with $b_{-1} = 0$ and $d_0 = 0$, while the presence of the constant influx S produces a density dependence effect. S can arise due to effective rates of immigration, emigration, speciation or extinction in a local community, so as due to intraspecific interactions, and, being independent from the species abundance n, it increases all species of the same amount, causing a rare species advantage. Because of the neutral assumption of species equivalence, the per-capita model rates are the same for all the species and Eq. 78 describes the whole ecosystem. Introducing stochasticity, the deterministic process 78 can be rewritten in terms of the probability of having n individuals in a species using the master equation (see Sec. 0.2)

$$\frac{\partial P_n(t)}{\partial t} = P_{n-1}(t)b_{n-1} + P_{n+1}(t)d_{n+1} - P_n(b_n + d_n), \tag{79}$$

where S has been included in the birth term, setting $b_n = b \cdot (n + \Upsilon)$, and $\Upsilon = S/b$, while the death rate is simply $d_n = d \cdot n$.
The stationary solution is obtained with the linear expansion method (see Sec. 0.2.1), exploiting the birth-death process detailed balance condition:

$$P_n = P_0 \prod_{i=0}^{n-1} \frac{b_i}{d_{i+1}} = P_0 \prod_{i=0}^{n-1} \frac{b \cdot (\frac{S}{b} + i)}{d \cdot (i+1)}. \tag{80}$$

Considering $n > 0$ and deducing P_0 from the normalization condition $\sum_{n\geq 0} P_n = 1$, Volkov et al. obtained the stationary solution, that turns out to be a Negative Binomial distribution

$$P_n = P_{RSA} = \frac{(1 - \frac{b}{d})^{S/b}}{\Gamma(S/b)} \frac{(\frac{b}{d})^n}{n!} \Gamma(n + \frac{S}{b}). \tag{81}$$

The average number of species with n individuals, given the total number of species N, is

$$\phi_n = N \cdot P_n. \tag{82}$$

Since the zero abundance class can not be observed, the mean number of observed species actually is

$$N_{obs} = N - \phi_0 = N - N \cdot (1 - b/d)^{S/b}, \tag{83}$$

and the average number of species in the community, that is the RSA normalization factor, is give by

$$N = \frac{N_{obs}}{1 - (1 - b/d)^{S/b}}. \tag{84}$$

Under these assumptions, the RSA distribution turns out to be the (truncated) Negative Binomial

$$P_{RSA}(n) = \frac{N_{obs}}{1 - (1 - b/d)^{S/b}} \frac{(1 - b/d)^{S/b}}{\Gamma(S/b)} \frac{(b/d)^n}{n!} \Gamma(n + S/b). \tag{85}$$

Moreover, when the influx parameter $\Upsilon = S/b$ is very small, we have $P_{RSA} \sim \theta \frac{(b/d)^n}{n}$ and the RSA resembles Fisher's Log-Series distribution with biodiversity number

$$\theta = \frac{N_{obs}(1-b/d)^{S/b}}{[1-(1-b/d)^{S/b}]\Gamma(S/b)} = \frac{N_{obs}}{[(1-b/d)^{-S/b}-1]\Gamma(S/b)}. \tag{86}$$

On the other hand, for larger Υ, the RSA distribution exhibits a clear interior mode at abundance $n > 1$ and becomes Log-Normal-like, with the rare species constituting a smaller fraction of all the species. In this way, the Negative Binomial distribution is able to satisfactorily describe many of the situations commonly observed and discussed in ecology, resembling both a Log-Series and a Log-Normal RSA [5]. We will see in Part I that the birth death model with a density dependent influx is appropriate to describe the Gut Microbiota ecosystem, even if, in this case, we will have to relax the neutrality assumption and to consider two niches, in which the equivalence of species holds, that are characterized by different dynamic rates. However, for other ecosystems, the RSA still turns out to be better described by a Log-Normal distribution as originally observed by Preston. We will see in Part II that this is the case of protein domains. A dynamic model that results in a Log-Normal was proposed in 1996 by Engen and Lande [29]. The main differences with Volkov's models are the assumption of Gompertzian death and the introduction of a constant environmental variances that acts independently on each species and adds up to the demographic noise. In the next section we detail the derivation of the Log-Normal distribution based on these hypothesis.

0.7.6 Engen and Lande's Poisson Log-Normal

Engen and Lande [29] supposed that species were generated by an inhomogeneous Poisson process, with rate $w(t)$ and that they evolve independently.
Let $f(x;t)$ be the distribution of the abundance $X(t)$ of a certain species that entered the community before the present time $t_0 = 0$. Moreover, let $p(t)$ be the probability that such species has not gone extinct before the present time. Then, as reported in the following, Engen and Lande proved that the abundances of species that are in the community at the present time are generated by an inhomogeneous Poisson process with rate

$$\lambda(x) = \int_0^\infty w(-t)p(t)f(x;t)dt. \tag{87}$$

Proof. Let Ω_1, $\Omega_2 \in \mathbb{R}^+$ be non-overlapping intervals. Let $Y_1(t)$ and $Y_2(t)$ be the contributions to the species number with abundances in Ω_1 and Ω_2 respectively, from species entering the community in the time interval $(t, t+\delta t)$, $t < t_0$. Since we supposed that species enter the community following a Poisson process, the properties iv to vi reported in Sec. 0.1.3 suggest that:

- the probability that no species enters the community in the time interval δt, so that $Y_1(t) = Y_2(t) = 0$, is $1 - w(t)\delta t + o(\delta t)$;
- the probability that one species enters the community, i.e. $Y_1(t) = 1$ and $Y_2(t) = 0$ or $Y_1(t) = 0$ and $Y_2(t) = 1$, is $w(t)\delta t + o(\delta t)$;
- the probability that two species enter the community, i.e. $Y_1(t) = Y_2(t) = 1$, is negligible.

Moreover, the probability of having a species in $Y_1(t)$ will be given by the probability of having a new species entering the community through the Poisson process, that such

species does not go extinct and that its abundance distribution is in Ω_1. Thus,

$$P(Y_1(t)=1,Y_2(t)=0) = w(t)p(t_0-t)\delta t \int_{\Omega_1} f(x;t_0-t)dx + o(\delta t)$$
$$P(Y_1(t)=0,Y_2(t)=1) = w(t)p(t_0-t)\delta t \int_{\Omega_2} f(x;t_0-t)dx + o(\delta t)$$
$$P(Y_1(t)=0,Y_2(t)=0) = 1 - w(t)p(t_0-t)\delta t \int_{\Omega_1\cup\Omega_2} f(x;t_0-t)dx + o(\delta t) \quad (88)$$
$$(89)$$

Here, the (t_0-t) argument in p and f indicates that we are considering species that enter the community at a time preceding the current time t_0. Let us set $I_j = \int_{\Omega_j} f(x;t_0-t)dx$. Then, the joint moment generating function $G_y(k) = \int e^{iky}P_y(y)dy$ for $Y_1(t)$ and $Y_2(t)$ takes the form

$$E[e^{uY_1(t)+vY_2(t)}] = 1+(e^u-1)w(t)p(t_0-t)I_1(t)\delta t+(e^v-1)w(t)p(t_0-t)I_2(t)\delta t+o(\delta t) \quad (90)$$

Taking the logarithm and considering the approximation $log(1+x)\approx x$ for small x, we find the corresponding cumulant generating function

$$K_t(u,v) = [(e^u-1)I_1(t)+(e^v-1)I_2(t)]p(t_0-t)w(t)\delta t + o(\delta t) \quad (91)$$

The total cumulant generating function for the species number in Ω_1 and Ω_2 can be found by splitting the time interval $(-\infty,t_0)$ up to intervals of length δt and adding all the contributions of $K_t(u,v)$ for these intervals, that are given by Eq. 91.

$$K(u,v) = \sum_{-\infty}^{+\infty}[(e^u-1)I_1(t)+(e^v-1)I_2(t)]p(t_0-t)w(t)\delta t. \quad (92)$$

In the limit $\delta t\to 0$, the sum converges to the integral

$$K(u,v) = (e^u-1)\phi_1 + (e^v-1)\phi_2 \quad (93)$$

where

$$\phi_j = \int_{-\infty}^{t_0} I_j(t)w(t)p(t_0-t)dt \quad (94)$$

for $j=1,2$. This is the cumulant generating function of a Poisson distribution, for which in general $K(h)=\sum_j \mu_j(e^h-1)=\sum_j\lambda(x_j)x_j(e^h-1)$. If we choose $\Omega_1=[x_0,x]$, we find that

$$\phi_1(x) = \int_{-\infty}^{t_0}\int_{x_0}^{x} f(y;t_0-t)w(t)p(t_0-t)dydt \quad (95)$$

If $f(x;t)$ is continuous, changing the order of integration and taking the derivative with respect to x, we find that the abundances follow an inhomogeneous Poisson process with rate

$$\lambda(x) = \frac{d\phi_1}{dx} = \int_{-\infty}^{t_0} w(t)f(x;t_0-t)p(t_0-t)dt \quad (96)$$

that we can rewrite as

$$\lambda(x) = \int_{t_0}^{\infty} w(t_0-t)p(t)f(x;t)dt. \quad (97)$$

Diffusion approximation In order to determine the abundance model $\lambda(x)$, the species abundances are described as a stochastic process $\{X(t)\}$ that follows the dynamic equation

$$\frac{dx}{dt} = rx - xg(x) \tag{98}$$

where $r = r + \sigma_r(x)dB(t)/dt$ is the growth rate, with $dB(t)/dt$ being a white noise with mean 0 and variance 1, and $g(x)$ is the death rate that will be defined in the following together with $\sigma_r(x)$.
Let us introduce the transformation $y = ln(x)$, that also implies $dy = dx/x$. The corresponding stochastic differential equation for y is

$$\frac{dy}{dt} = [r - g(e^y)] + \sigma_r(e^y)\frac{dB(t)}{dt}. \tag{99}$$

As pointed out before, we suppose that the variance $\sigma_r^2(e^y)$ has two components, that are the environmental and the demographic noise (see also the Introduction). Environmental stochasticity is due to changing environments that act simultaneously on all individuals in the population and it is assumed to be a constant σ_e^2. The demographic component, on the other hand, reflects the differences in fertility and survival among individuals. This effect acts on each individual independently. Consequently it is inversely proportional to the species number, and it is assumed to be $\sigma_d^2 e^{-y}$. Hence, the total variance is given by $\sigma_r^2(y) = \sigma_e^2 + \sigma_d^2 e^{-y}$.
Applying the Ito approach (see Sec. 0.4.1, Eq. 59) to the process $\{Y(t)\}$ we obtain the corresponding Fokker-Planck equation

$$\frac{\partial P(y)}{\partial t} = -\frac{\partial}{\partial y}[r - g(e^y)]P(y) + \frac{1}{2}\frac{\partial^2}{\partial y^2}\sigma_r^2(y)P(y). \tag{100}$$

By introducing the Fokker-Planck equation, we are approximating the process for each species to a diffusion process. More precisely, $\{Y(t)\}$ is a diffusion process with infinitesimal mean $\mu(y) = r - g(e^y)$ and infinitesimal variance $\sigma_r^2(y) = \sigma_e^2 + \sigma_d^2 e^{-y}$ and is indeed defined over a continuous state space. We can now transform back to x through the transformation $x = e^y$. Following the rules for the transformation of the infinitesimal parameters of a diffusion process given by Eq. 24 in Sec. 0.3.1, we obtain

$$\begin{aligned} \mu(x) &= \left[r + \frac{1}{2}\frac{\sigma_d^2}{x} + \frac{1}{2}\sigma_e^2\right]x - xg(x) \\ \sigma^2(x) &= \sigma_d^2 x + \sigma_e^2 x^2 \end{aligned} \tag{101}$$

The Fokker-Planck equation for the diffusion process $\{X(t)\}$ is then

$$\frac{\partial P(x)}{\partial t} = -\frac{\partial}{\partial x}\mu(x)P(x) + \frac{1}{2}\frac{\partial^2}{\partial x^2}\sigma^2(x)P(x). \tag{102}$$

Note that, without the log-transformation passage, we would have obtained a different Fokker-Planck equation. In particular, we would not have the noise terms in $\mu(x)$.
In order to find the stationary solution, we set $\partial P(x)/\partial t = 0$ and expand the second order derivative so that to obtain

$$\frac{1}{2}\sigma^2(x)\frac{\partial^2 P(x)}{\partial x^2} - \mu(x)\frac{\partial P(x)}{\partial x} - \frac{\partial \mu(x)}{\partial x}P(x) = 0. \tag{103}$$

Now, let us suppose $\{X(t)\}$ to be a diffusion process with absorbing barriers a and b, $a < b$. In our contest, the left-hand barrier a represents extinction, while the right-hand

barrier b would be an upper limit to the species abundance that we will move towards infinity.
Eq. 103 is a second order differential equation of the form

$$Ly \equiv p(x)\frac{d^2y(x)}{dx^2} + q(x)\frac{dy(x)}{dx} + r(x)y(x) = f(x) \tag{104}$$

with $f(x) = 0$ and boundary conditions at $x = a$ and $x = b$

$$y(a) = y(b) = 0. \tag{105}$$

If the homogeneous equation $Ly = 0$ admits no nonzero solution fulfilling the boundary conditions, then Eq. 104 can be inverted in the form of an integral operator

$$y(x) = \int_a^b G(x,\xi)f(\xi)d\xi \tag{106}$$

where $G(x,\xi)$ is commonly referred to as Green function of the boundary value problem. The Green function has the property that, for each $\xi \in (a,b)$, the function $G(x,\xi)$ solves the equation $Ly = 0$ and also satisfies the boundary conditions, so that the solution of our problem (Eq. 103) will be given by $G(x,\xi)$. In order to obtain the Green function, let us consider two solutions of $Ly = 0$, $y_1(x)$ and $y_2(x)$, such that y_1 satisfies the left boundary condition $y_1(a) = 0$ and y_2 instead satisfies the right-hand one, $y_2(b) = 0$. Since we supposed that there exists no nonzero solution that satisfies both boundary conditions, we know that y_1 and y_2 are linearly independent. Consequently, the Wronskian determinant $W(y_1, y_2) = y_1y_2' - y_1'y_2$ will be nonzero. We know from the differential equation theory, that the Green function for such problem is

$$G(x,\xi) = \begin{cases} \frac{y_1(\xi)y_2(x)}{p(\xi)W(\xi)}, & \text{for } a \le \xi \le x \le b \\ \frac{y_1(x)y_2(\xi)}{p(\xi)W(\xi)}, & \text{for } a \le x \le \xi \le b \end{cases} \tag{107}$$

and is a solution of the homogeneous problem.
We expect a function of the form

$$S(x) = \int_x s(\eta)d\eta = \int_x e^{-\int_\eta \left[2\mu(u)/\sigma^2(u)\right]du}d\eta \tag{108}$$

to satisfy the differential equation 103. We thus take the two solutions

$$\begin{aligned} y_1 &= \frac{S(x) - S(a)}{S(b) - S(a)} \\ y_2 &= \frac{S(b) - S(x)}{S(b) - S(a)} \end{aligned} \tag{109}$$

so that $y_1(a) = 0$ but $y_1(b) \neq 0$ and $y_2(a) \neq 0$ but $y_2(b) = 0$. The Wronskian determinant is given by

$$W(y_1, y_2) = \frac{s(x)}{S(b) - S(a)} \tag{110}$$

and the Green function turns out to be

$$G(x,\xi) = \begin{cases} \frac{S(\xi)-S(a)}{S(b)-S(a)}\frac{S(b)-S(x)}{p(\xi)W(\xi)}, & \text{for } a \le \xi \le x \le b \\ \frac{S(x)-S(a)}{S(b)-S(a)}\frac{S(b)-S(\xi)}{p(\xi)W(\xi)}, & \text{for } a \le x \le \xi \le b, \end{cases} \tag{111}$$

as proven in [6]. Note that $G(x,\xi) = G(\xi,x)$. Moreover, since we defined $\lambda(x)$ for $x \in \Omega_1 = [x_0, x]$ (see Eq. 95), we are in the case $a \le \xi = x_0 \le x \le b$, and the solution of our problem is given by

$$G(x_0,x) = \frac{S(x_0) - S(a)}{S(b) - S(a)}\frac{S(b) - S(x)}{s(x)p(x)}, \text{ for } a \le x_0 \le x \le b. \tag{112}$$

We let now $b \to \infty$, so that to avoid an upper limit to the species abundances. Since our model has the density regulation term $-xg(x)$, it follows that, for $x = b \to \infty$, $-\frac{\mu(x)}{\sigma^2(x)} \sim +\frac{g(x)}{x}$, and since $S(x = b)$ is the integral of an increasing exponential, also $S(x = b) \to \infty$. So, we can approximate $S(b) - S(x) \sim S(b) - S(a) \sim S(b)$ and write

$$G(x_0, x) = \frac{S(x_0) - S(a)}{s(x)p(x)}, \text{ for } a \leq x_0 \leq x \leq b. \tag{113}$$

Let us define extinction to occur at $x = a = 1$. Then, if we choose the starting point $x_0 = x(t_0) = 1$, we obtain $G(1, x) = 0$, which means that species immediately go extinct with probability 1. This is a special property of the diffusion process that is not realistic for processes with discrete states $x = 1, 2, ...$ However, we can still use the diffusion approximation for all $x \geq 1$ if we redefine speciation in the following way. Suppose, that speciation occurs when the abundance reaches the value $x = 1 + \delta x$. The Green function for $x_0 = 1 + \delta x$ and $a = 1$ is

$$G(1 + \delta x, x) = \frac{S(1 + \delta x) - S(1)}{s(x)p(x)} \tag{114}$$

Then, as $\delta x \to 0$, $S(1 + \delta x) = S(1) + \delta x S'(1) + o(\delta x)$, with $S'(1) = s(1)$ by definition. Hence, the Green function tends to

$$G(1 + \delta x, x) = \frac{\delta x \cdot s(1)}{s(x)p(x)} + o(\delta x) \tag{115}$$

and substituting $s(x) = e^{-\int_{-\infty}^{x}[2\mu(u)/\sigma^2(u)]du}$ and $p(x) = \frac{1}{2}\sigma^2(x)$, we obtain

$$G(1 + \delta x, x) = \frac{2\delta x}{\sigma^2(x)} e^{\int_1^x [2\mu(u)/\sigma^2(u)]du} \tag{116}$$

If we let the speciation rate $\mu_0 = w_0/\delta x \to \infty$ and $\delta x \to 0$ so that $\mu_0 \delta x \to w_0 > 0$, the abundance model becomes

$$\lambda(x) = 2\frac{w_0}{\sigma^2(x)} e^{\int_1^x [2\mu(u)/\sigma^2(u)]du}. \tag{117}$$

Note that the speciation parameter $w_0 = \mu_0 \delta x$ equals the speciation rate μ_0 when $\delta x = 1$. Hence, we can interpret w_0 as the rate at which species reach an abundance equal to 2. Finally, we substitute the definitions for the infinitesimal mean and variance (Eq. 101), and we assume that the density regulation is given by the Gompertzian model $g(x) = \gamma ln(x+\epsilon)$ (see Sec. 9.4 for details), where $\epsilon = \sigma_e^2/\sigma_d^2$. The solution of the integral is

$$2\int_1^x \frac{\mu(u)}{\sigma^2(u)} du = \left[-\left(ln(x + \epsilon) - \frac{r}{\gamma} \right)^2 + \left(ln(1 + \epsilon) - \frac{r}{\gamma} \right)^2 \right] \frac{\gamma}{\sigma_e^2} + ln(x) \tag{118}$$

and the abundance model takes the form

$$\lambda(x) = \frac{\alpha w_0}{x + \epsilon} e^{-\frac{1}{2}\frac{[ln(x+\epsilon) - r/\gamma]^2}{\sigma_e^2/2\gamma}} \tag{119}$$

where

$$\alpha = \frac{2}{\sigma_e^2} e^{\frac{\gamma}{\sigma_e^2}\left[ln(1+\epsilon) - \frac{r}{\gamma}\right]^2}. \tag{120}$$

The abundance model, in conclusion, has a lognormal distribution translated by $-\epsilon$, with mean r/γ and variance $\sigma_e^2/2\gamma$. Note that $x = \epsilon$ represents a very small abundance, and for $x \gg \epsilon$ the translation may be ignored. The abundance model derived by Engen and

Lande is hence mathematically equivalent to the familiar sequential broken stick model proposed by Sugihara [20] that we described in Sec. 0.7.1. Note that, if we neglect the translation term ϵ, the demographic noise σ_d^2 has no effect on the shape of the abundance curve but only affects the constant α.
The model by Engen and Lande can be seen as the generalization of the broken stick approach to the lognormal species abundance distribution with the advantage of resulting in a stationary distribution. The abundances at each time position are the points of an inhomogeneous Poisson process, where the rate $\lambda(x)$ completely characterizes all properties of the community structure and follows a Log-Normal distribution.

Heterogeneity case Engen and Lande proved that their formulation also includes the case of heterogeneity, in which we consider that different species may be generated by different stochastic processes [29]. In particular, we can suppose that the process for one species entering the community at time t belongs to some family of processes with parameter θ sampled from some distribution $\pi(\theta)$. θ is interpreted as the mean growth rate r among species and, considering a normally distributed r, we obtain an expression for $\lambda(x)$ that is equivalent to Eq.119.

Interspecific density regulation and correlated environmental noise A further extension of the model may be achieved introducing competition between species and correlated environmental noise. In particular, as detailed in [29], if we add an interspecific density regulating term, that is a death term acting equally on all individuals, and an environmental variance component also common to all species, the form of the diffusion process describing the population does not change.

Part I
Healthy aging prediction through ecological modeling of gut microbiota

Chapter 1

Introduction

The human gut is home to a diverse and complex community of trillions of microorganisms, that play a central role in human health, including metabolism, physiology, nutrition and immune function [30]. The collective genome of these symbiotic microorganisms (called microbiome) is tightly integrated with the human genome, making humans 'superorganisms' [31], whose health state is defined by the interaction between the microbiota and the host living environment [32][33]. Disruption of the gut microbiome, termed dysbiosis, has been observed in several pathological conditions and disorders [32][34], ranging from metabolic diseases (e.g. metabolic syndrome, obesity, type 2 diabetes [35][36][37][38]) to immune diseases (e.g. Crohn's Disease, Ulcerative Colitis, type 1 diabetes mellitus [39], multiple sclerosis [40], celiac disease and allergies [41]), but also in colorectal cancer [42], rheumatoid arthritis [43], Irritable Bowel Syndrom [44], recurrent *Clostridium difficile* colitis [45], and many others. Dysbiosis is frequently accompanied by significant loss of microbial diversity or key functional groups, in conjunction with overgrowth of pathogenic bacteria or fungi. This sort of changes in the gut microbiota composition are usually linked with an increased energy harvest from ingested food [35] and an inflammatory response by the host, which contributes to disease development. The result is the set up of a 'vicious circle' in which the gut microbiota is further disrupted, beneficial bacteria are reduced and opportunistic colonizers, typically pathogens, are permitted to compete, supporting a persistent inflammatory state of the gut [46][32][47]. Several authors have questioned this chicken-egg problem and some speculate that there is a causal effect of the reduction in microbiota diversity on human disease [34]. Loss of microbiota diversity appears, in fact, as the most constant finding of intestinal dysbiosis and related conditions [34]. For example, obesity is associated with altered representations of specific bacterial species, which often differ between studies, but loss of diversity is constantly retrieved with up to 20% loss of phylogenetic diversity [48][49][50]. The gut microbiota composition is influenced by a range of factors including the microbial species acquired at birth, host genetics, immunological factors and lifestyle. Loss of microbiota diversity is a feature of industrialized countries [34] and many of its candidate risk factors can be recognized in some life style typical of these societies, such as certainly eating behavior [51], lack of physical exercise [52], the disruption of biological clock [53], antibiotic consumption [54] and also aging [55], in addition to the general health state. In the work by Claesson et al. [55] is shown that the healthy food diversity index (HFD) positively correlates with microbiota diversity, indicating that a healthy, diverse diet promotes a more diverse gut microbiota. Moreover, differences in the microbiota composition and diversity were also found between older people ($>$ 65 years) and younger adults. Changing the amount of ingested fiber and fat has a profound influence on the composition of the gut microbiota and its metabolic products both over a short period and in the longer-term [54]. Phyla positively associated with fat but negatively associated with fiber are predominantly *Bacteroidetes* and *Acti-*

nobacteria, whereas *Firmicutes* and *Proteobacteria* show the opposite association [51] [56]. Dietary composition, modification and interventions can have a great impact on the gut microbiota diversity. Bacteria, in fact, are specialized in the fermentation of different substrates, so that complex diets provide a range of growth-promoting and growth-inhibiting factors for specific phylotypes [57]. In particular, food components which are indigestible for human enzymes (e.g. fiber) provide substrates for the intestinal microbial community and this is why agrarian diets high in fruit/legume fiber are associated with enhanced gut microbial diversity [58].
Given the importance of diversity in the development and maintenance of the gut microbial community and of the host health state, we focus here on modeling the population dynamics that may lead to a certain composition and diversity of the Gut Microbiota, considering an ecological description. In order to fulfill such purpose, we base our analysis on 16S rRNA sequecing data. As it turned out during the 1990s, in fact, the majority (about 80%) of microbes observed by microscopy or sequencing of fecal specimens are not recoverable by culture [59][33]. To overcome this limitation, Metagenomics collects and analyses the genetic material present in an environmental sample. When studying GM, the usually sampled gene is the 16S rRNA gene (or some of its regions), that is a component of the 30S small subunit of prokaryotic ribosomes. This is a highly conserved sequence, meaning that phylogenetically similar bacteria, that are closed from an evolutionary point of view, have very similar 16S rRNA, while bacteria belonging, for instance, to distinct phyla will have more different 16S rRNA. As a consequence, clustering 16S rRNA sequences according to their similarity through some *de novo* method, enables to identify bacterial species independently of human made taxonomic classifications, besides without relying on culturing techniques or needing to know in advance which microorganism to look for, as required for example by microarrays. Clusters of 16S rRNA sequences are called Operational Taxonomic Unites (OTUs) and constitute a new definition of species, that is much more accurate if we aim to obtain ecological information about the Gut Microbiota and its biodiversity.
We remind that one way to evaluate the internal diversity of an ecosystem is to compute the so called Relative Species Abundance distribution (RSA), which counts the number of species that have a certain number of individuals (see Sec. 0.5.2). Thus, computing OTUs and deriving their abundances we will be able to estimate the Gut Microbiota biodiversity by computing its RSA and we will also manage to test hypothesis over its the ecological dynamics, as we will see.
In a previous work we have shown that a proper model to describe the Gut Microbiota population in two cattle rumens, two swines and one chicken was the one proposed by Volkov [24] (see Sec. 0.7.5). Moreover, we proved that animals belonging to different species were characterized by different diversity parameters θ [60]. Here, we aim to model the GM population of human individuals starting from 16S rRNA sequencing. As described in the following, the analysis requires a first bioinformatic processing of the data that enables to compute the RSA distribution. Then, we will show that a redefinition of the RSA model is required to better describe the Gut Microbiota population. Finally, we will characterize the GM biodiversity of subjects with different health state, age and eating behavior using the new model biodiversity index.

Chapter 2

Material and Methods

2.1 The ELDERMET and ELDERMETpart datasets

We analyzed data from Claesson et al. [55], whose fasta files are available on MG-RAST under the Project ID 154, and the ELDERMET data [61], whose fastx files are available on the Sequence Read Archive under BioProject PRJNA283106. Here we will refer to the first dataset as ELDERMETpart because it can be considered as a subset of the ELDERMET dataset, as clarified in the following. In both datasets, DNA was extracted from faecal samples, and sequence reads from 16S rRNA gene V4 amplicons were generated with 454 Genome Sequencer FLX Titanium platform. The ELDERMETpart dataset includes 164 elderly subjects, non-antibiotic-treated, stratified by community residence setting: (1) community-dwelling, $n = 81$; (2) attending an out-patient day hospital, $n = 20$; (3) in short- term (<6 weeks) rehabilitation hospital care, $n = 12$; (4) in long-term residential care (long-stay), $n = 51$. For each subject we also have dietary information. In particular habitual dietary intake had been assessed using a validated, semiquantitative, food frequency questionnaire (FFQ), and four dietary groups (DGs) had been identified by authors: DG1 ('low fat/high fibre') and DG2 ('moderate fat/high fibre') included 98% of the community and day hospital subjects, and DG3 ('moderate fat/low fibre') and DG4 ('high fat/low fibre') included 83% of the long-stay subjects. Other clinical parameters included in the dataset are: gender, CCI, FIM, Barthel, MMSE, Weight, BMI, CC, Diastolic BP, Systolic BP, GDT, MNA, CRP, IL-6, IL-8 and TNFα.
In particualar, the FIM [62], Barthel [63] and MMSE [64] indices are related to the physical and cognitive state of the elderly person. The FIM score ranges from 18 (dependent) to 126 (fully independent) in the six sections of self-care, sphincter control, mobility, locomotion, communication and social cognition. The Barthel index, instead, ranges from 0 (sever dependence) to 20 (total independence) and is based on ten variables describing activities of daily living and mobility. Finally, the MMSE score measure cognitive impairment and takes values from 0 to 30, where a value greater than 24 usually indicates normal cognition, while a value less than 9 suggests sever cognitive impairment. We found out that in the considered datasets, these three indices were highly correlated among them and also with the residence setting and the dietary group, so that people with lower physical and cognitive abilities were also those in rehabilitation or long-term hospital care and composed most of the DG3 and DG4 dietary groups, that are those richer in fatty acids and poorer in fibers. For this reason, as shown in Sec. 3.2, we summarized the clinical information dividing the subjects in two groups that we will call healthy and unhealthy.
The mean subject age is 78 ($\pm$8 s.d.) years, with a range of 64 to 102 years, and all are of Caucasian (Irish) ethnicity. The study also includes 13 young adults with a mean age of 36 (66 s.d.) years.
From the ELDERMET dataset we selected the 13 young samples, plus those subjects for

which we had at least 3 samples. These usually refer to 3 time points, that we call $T0$, $T1$ that is around 3 months after $T0$ and $T2$ that is around 3 months after $T1$. When more than one sample was present for the same subject at the same time point, results were averaged. The 13 young subjects are the same as those in the ELDERMETpart dataset and are present only at time $T0$. In total, the ELDERMET dataset includes 97 subjects at time $T0$, 69 at time $T1$ and 77 at time $T2$. All the 97 subjects of time $T0$, that include the 13 young controls, were also part of the ELDERMETpart dataset, that instead did not include times $T1$ and $T2$. This is the reason for which we labeled the first dataset as ELDERMETpart. Clinical variables for the ELDERMET dataset include: antibiotics, Age, BMI, Gender, Stratum, MNA, FIM, MMSE, Barthel, IL6, IL8, IL10 and TNFa. Among these, the only clinical parameter available for the young group was Age in both datasets. Starting from 16S rRNA sequences we computed *de novo* OTUs, so that to assess the phylogenetic relationships between species irrespectively of taxonomic classification.

2.2 Preprocessing

Raw sequencing reads were filtered according to the following criteria. ELDERMET data were preprocessed starting from fastq files with FASTX Toolkit v.0.0.13 [65]. Sequences were trimmed to be no longer than 350 base pairs (bp) and were discarded when shorter than 150 bp. Sequences with ambiguous bases (Ns) were also discarded and quality filtering was performed so that all sequences had a Phred-33 quality score [66] greater than 25 in at least 50% of their bases. Fastq files were not available for ELDERMETpart data. For this reason, preprocessing was performed starting with fasta files. Sequences were filtered with mothur v.1.31.2 [67] according to the following requirements: (1) no ambiguous bases (Ns); (2) read-lengths not shorter than 150 bp or longer than 350 bp; (3) homopolymers not longer than 8 bp. Trimmed sequences were clustered into OTUs following the UPARSE pipeline [68]. We will discuss the importance of *de novo* clustering of 16S rRNA sequences into OTUs in Part II. As we will see, many clustering procedures have been proposed for this purpose, but here we chose to use UPARSE, that is one of the more standard and advisable ones [68].

2.3 Clustering 16S rRNA into OTUs

After pooling samples and a first dereplication step, sequences were sorted by decreasing abundance and singletons were discarded. Then, the clustering algorithm was applied. UPARSE uses a greedy algorithm to find a biologically relevant set of OTUs in which all pairs of OTU representative sequences (cluster centroids) should have pair-wise sequence similarity less than a specified threshold (e.g. 97%), chimeric sequences should be discarded and all non-chimeric input sequences should match at least one OTU representative sequence with similarity higher or equal to the threshold. Such set of OTUs is found with the following strategy. First, sequences are ordered so that to have the more reliable reads at the top of the input file, since these will be more likely to be chosen as cluster centroids, as explained in the following. By default, since high-abundance reads are more likely to be correct amplicon sequences, and hence are more likely to be true biological sequences, UPARSE considers input sequences in order of decreasing abundance. The first input sequence will be the first sequence in the OTU database D. Then, each other input sequence is compared to the sequences in D, that are the OTU representative sequences (centroids), and a maximum parsimony model of the sequence is found using UPARSE-REF. The method tries to explain a given sequence S with the fewest possible events starting from sequences in D, where 'events' are mutations arisen from PCR or

sequencing errors. This is done by constructing a model sequence M using one or more sequences from the database (refseqs). Typically, M is a single refseq representing a non-chimeric amplicon. Otherwise, M is made from m refseq segments that are concatenated to represent a chimeric amplicon. If M has one segment, i.e. is a single refseq, then the distance between M and S is defined to be the number of mismatches, which are interpreted as sequencer or PCR errors. This is obtained giving a score 0 for each match and -1 for each mismatch. In case of chimeric model, a score -3 is added for each chimeric breakpoint. There are three cases: (a) the similarity between the UPARSE-REF model and an existing OTU is more or equal to the threshold, (b) the model is chimeric, or (c) the model is less similar than the threshold to any existing OTU. In case (a), the input sequence becomes a member of the OTU. In case (b), the input sequence is discarded. In case (c), the input sequence is added to the database and becomes the representative sequence (centroid) of a new OTU.
UPARSE requires to chose a similarity threshold, that defines the phylogenetic resolution at which we observed the GM ecosystem. In UPARSE it is not recommended to use similarity thresholds lower than 97%, since this would imply an increase of false negative chimeras during the chimera detection step. However, clustering 16S rRNA sequences into OTUs at lower similarities interestingly gives an insight in the bacteria phylogeny at different scales. For this reason, we used the following procedure to obtain OTUs at 93% and 95%, besides 97%, similarity thresholds [69]. We define the final required OTUs radius r (similarity threshold $= (1 - r) \cdot 100\%$), and we first compute OTUs using UPARSE at radius $r/2$. Then, we run UCLUST [70] to re-cluster the obtained representative sequences, together with their whole cluster, at radius r.
UCLUST starts with an empty database in memory and then reads the sequences in input order, as UPARSE. The algorithm takes the first sequence as first OTU representative sequence, and process all other sequences according to the following statement: if a sequence is similar to an OTU representative sequence within the similarity threshold, then the query is assigned to its OTU; if a sequence is instead not similar to any OTU representative sequence, then it will become the representative sequence for a new cluster. In this way, running UPARSE with a radius $r/2$ and UCLUST with a radius r, we generate clusters in which all member sequences have a distance $\leq r$ to their cluster centroid.

2.4 Modeling the Gut Microbiota RSA

Once OTUs are obtained, the empirical RSA distribution is computed counting how many OTUs have a certain number of individuals and can be used to verify theoretical hypothesis. For instance, according to the neutral theory proposed by Volkov in [24], RSA should exhibit a Negative Binomial shape (see Sec.0.7.5). As detailed in Chapter 3, our results actually show that GM RSA has a heavy tail that could not been explain by a pure neutral model, especially when clustering sequences into OTUs at 97% of similarity, that is when we consider a low phylogenetic level. A similar deviation from neutrality has already been observed in animals Gut Microbiota [60] and for the coral-reef ecosystem [71]. For this reason, we propose to relax the neutral hypothesis of Volkov's model and to consider two niches within which neutrality holds.
The logistic model with interaction is given by

$$\dot{x_i} = x_i \left[\alpha_i - \sum_{j=1}^{2} \gamma_{ij}^* x_j \right], \; i = 1, 2 \tag{2.1}$$

where $\alpha_i = b_i^* - d_i^*$ represents the net growth rate of the population, whereas the symmetric positive defined matrix γ_{ij}^* introduces the effect of limit resources and competition in the

environment. We study the behavior of the system near the stable equilibrium point through linearization of the equations 2.1. First, we compute the equilibrium point by setting $\dot{x_i} = 0$, so that

$$\alpha_i x_i = \sum_{j=1}^{2} \gamma^*_{ij} x_j x_i. \tag{2.2}$$

Assuming

$$\begin{aligned} \alpha_1\gamma^*_{22} - \alpha_2\gamma^*_{12} > 0 \\ \alpha_2\gamma^*_{11} - \alpha_1\gamma^*_{21} > 0 \end{aligned} \tag{2.3}$$

the only stable equilibrium point is

$$\dot{x_i}^{eq} = \sum_{j=1}^{2} \gamma^{*-1}_{ij} \alpha_j = \sum_{j=1}^{2} \gamma^{*-1}_{ij} b^*_j - \sum_{j=1}^{2} \gamma^{*-1}_{ij} d^*_j \equiv b_i - d_i \tag{2.4}$$

and from Eq. 2.3 it follows that (x_1^{eq}, x_2^{eq}) belongs to the first quadrant and measures the abundance of the two populations. Linearizing the system 2.1 around the equilibrium point, we get

$$\dot{x_i} = x_i^{eq} \left[\alpha_i - \sum_{j=1}^{2} \gamma^*_{ij} x_j \right], \; i = 1, 2 \tag{2.5}$$

and calling the constant $x_i^{eq}\alpha_i \equiv S_i$ and setting $x_i^{eq}\gamma^*_{ij} \equiv \gamma_{ij}$, we have the deterministic equations

$$\dot{x_i} = S_i - \sum_{j=1}^{2} \gamma_{ij} x_j; \quad i = 1, 2. \tag{2.6}$$

Here, x_i is the number of individuals belonging to a specific species of niche i ($i = 1, 2$), S_i refers to density dependent effects, and

$$\gamma = \begin{bmatrix} (d_1 - b_1) & \gamma_{12} \\ \gamma_{21} & (d_2 - b_2) \end{bmatrix}$$

where b_i and d_i are the birth and death rates in niche i, while γ_{12} and γ_{12} are the interaction terms. The two populations, x_1 and x_2, belonging to the two niches, are characterized by two ranges of the birth (b_1, b_2), death (d_1, d_2) and density dependent (S_1, S_2) rates. However, because of the species equivalence assumption within each niche, equation 2.6 holds for every species in i-th niche. Moreover, in the simplest scenario in which, beyond inter-species interaction, also inter-population interaction is neglected, $\gamma_{12} = \gamma_{21} = 0$ and the time evolution of the probability that a species contains n individuals at time t, $P(n, t)$ is regulated by the master equation

$$\dot{P}(n_1, n_2, t) = \sum_{i=1}^{2} (E_i^+ - 1) \left[d_i n_i - E_i^- b_i n_i \right] P(n_i, t) \tag{2.7}$$

where E is the 'step operator' that is defined for any function $f(n)$ that depends on an integer variable n as $E^+[f(n)] = f(n + 1)$ and $E^-[f(n)] = f(n - 1)$. Since the two populations are supposed to be uncoupled, the stationary solution for the total population is given by the product of the two stationary solutions $P(n_1, n_2) = \prod_{i=1}^{2} P(n_i)$, where $P(n_i)$ satisfies the detailed balance condition

$$E_i^+ P(n_i) = \frac{b_i n_i + S_i}{E_i^+ d_i n_i} P(n_i). \tag{2.8}$$

It follows that

$$P(n_1, n_2) = \prod_{m_1=0}^{n_1-1} \frac{b_1 m_1 + S_1}{d_1(m_1+1)} \prod_{m_2=0}^{n_2-1} \frac{b_2 m_2 + S_2}{d_2(m_2+1)} P(0,0) \tag{2.9}$$

and factorizing $P(n_1 = 0, n_2 = 0) = P(n_1 = 0)P(n_2 = 0)$, we obtain the join stationary solution

$$P(n_1, n_2) = P(n_1) \cdot P(n_2) \tag{2.10}$$

$$= \frac{(1 - \frac{b_1}{d_1})^{\frac{S_1}{b_1}}}{\Gamma(S_1/b_1)} \frac{(\frac{b_1}{d_1})^{n_1}}{n_1!} \Gamma(n_1 + S_1/b_1) \cdot \frac{(1 - \frac{b_2}{d_2})^{\frac{S_2}{b_2}}}{\Gamma(S_2/b_2)} \frac{(\frac{b_2}{d_2})^{n_2}}{n_2!} \Gamma(n_2 + S_2/b_2), \tag{2.11}$$

where $P(n_1 = 0)$and $P(n_2 = 0)$ have been obtained normalizing the two probability distributions to 1. The marginal distributions are

$$P(n_i) = \frac{(1 - \frac{b_i}{d_i})^{\frac{S_i}{b_i}}}{\Gamma(S_i/b_i)} \frac{(\frac{b_i}{d_i})^{n_i}}{n_i!} \Gamma(n_i + S_i/b_i), \tag{2.12}$$

$i = 1, 2$. Equation 2.12 refers to the probability for a species belonging to population i to have n_i individuals. Following the handling of [24], the number of species observed in population i with n_i individuals is

$$\phi_{n_i} = \sum_{k=1}^{N_i} I_{n_i,k} \tag{2.13}$$

where N_i is the total number of species in niche i that may potentially be present in the community and $I_{n_i,k}$ is a random variable that takes value 1 with probability $P(n_i)$ and 0 with probability $1 - P(n_i)$. The average number of species containing n_i individuals is given by

$$\phi_{n_i} = \sum_{k=1}^{N_i} I_{n_i,k} = \sum_{k=1}^{N_i} P(n_i) = N_i P(n_i). \tag{2.14}$$

However, since species with zero individuals cannot be revealed, the average number of species observed in the population i is

$$N_{obs_i} = N_i - \phi_0 = N_i - \sum_{k=1}^{N_i} (1 - \frac{b_i}{d_i})^{S_i/b_i}, \tag{2.15}$$

from which

$$N_i = \frac{N_{obs_i}}{1 - (1 - \frac{b_i}{d_i})^{S_i/b_i}}. \tag{2.16}$$

The RSA distribution for population i is given by

$$P_{RSA}(n_i) = \frac{N_{obs_i}}{1 - (1 - b_i/d_i)^{S_i/b_i}} \frac{(1 - b_i/d_i)^{S_i/b_i}}{\Gamma(S_i/b_i)} \frac{(\frac{b_i}{d_i})^{n_i}}{n_i!} \Gamma(n_i + S_i/b_i), \tag{2.17}$$

where

$$\theta_i = \frac{N_{obs_i}(1 - \frac{b_i}{d_i})^{S_i/b_i}}{[1 - (1 - \frac{b_i}{d_i})^{S_i/b_i}]\Gamma(S_i/b_i)} = \frac{N_{obs_i}}{[(1 - \frac{b_i}{d_i})^{-S_i/b_i} - 1]\Gamma(S_i/b_i)}. \tag{2.18}$$

is the biodiversity number as previously outlined. Finally, assuming that the experimental data are a sample of the initial populations $x_{1,2}$, whose relative frequency is defined by

the ratio of the equilibrium states x_1^{eq}/x_2^{eq}, the empirical numerousness distribution of the different species reflects the distribution of the initial populations. The probability to detect a specie of numerousness n is the sum of probability to find it in the population x_1 or x_2 separately

$$P(n) = P_1 P(x_1 = n) + P_2 P(x_2 = n)$$

where P_i is the probability to detect a specie of the populations x_i $(i = 1, 2)$ and, being $P(n)$ normalized, we can assume $P_1 = \alpha$ and $P_2 = (1 - \alpha)$, $\alpha \in [0, 1]$ and, given the total number of observed species N_{obs}, we will have $N_{obs_1} = \alpha N_{obs}$ and $N_{obs_2} = (1 - \alpha)N_{obs}$. The total RSA distribution is described by a mixture of two Negative Binomials

$$P_{RSA}(n) = N_{obs}\left[\alpha \cdot \theta_1 \frac{(b_1/d_1)^n}{n!}\Gamma(n + S_1/b_1) + (1 - \alpha) \cdot \theta_2 \frac{(b_2/d_2)^n}{n!}\Gamma(n + S_2/b_2)\right], \tag{2.19}$$

and each population is associated with a biodiversity number: θ_1 and θ_2.
We have seen here how the simplest relaxation for the neutrality hypothesis is obtained considering a mixture of two Negative Binomials (2NB) rather than a single one (1NB). A further step away from the neutral model would be, for instance, to consider a mixture of three Negative Binomials (3NB), that would describe three non-interacting populations. Increasing the number of Negative Binomials in the mixture, indeed, also increases the number of parameters in the model and this may generate over-fitting issues. In order to assess whether our hypothesized 2NB model was a better description for the Gut Microbiota RSA than 1NB and 3NB, we fitted the ELDERMET and ELDERMETpart data with the Approximate Bayesian Computation (ABC) rejection algorithm detailed in the following and weighted the goodness of fit with the number of parameters through the Akaike Information Criterion (AIC), also detailed afterwards.

2.5 Fitting the Gut Microbiota RSA with ABC

When fitting the Gut Microbiota RSA data with common methods such as Maximum Likelihood Estimation or the least squares, several problems arise. In fact, the data present very heavy tails that, without any penalization, would dominate the fit. The method would then weight the tail points more than the left-hand part of the curve, because of their copiousness, and it would thus give good estimates for the abundant species and much less accurate evaluations for the rarest ones. A solution may be to fit the logarithmic transformed abundances. For this purpose, the Preston plot representation comes to aid. However, fitting directly the Preston plot bins is a rough approximation and the fitting methods usually become pretty sensitive to the initial values of parameters, besides having difficulties in discriminating between distributions that are similar in shape, such as the Negative Binomial and the Log-Normal, some common hypothesis for the RSA.
For these reasons, we chose to fit the model implementing the Approximate Bayesian Computation (ABC) rejection algorithm, that has the further advantage of returning a posterior distribution for each parameter, rather than simply its mean and error. In ABC, a set of points for the model parameters is first sampled from their predefined prior distributions. Then, a dataset G of the same dimension as the observed data O is simulated from the theoretical model with the sampled parameters. If the generated dataset is too different from the observed data, the sampled parameters values are discarded, otherwise they are accepted. The process is iterated several times, in our case 10^7 times, and a posterior distribution for the parameters is obtained from the accepted cases.
The two Negative Binomials of the RSA mixture, have been modeled through their mean $\mu = \frac{(1-b/d)(S/b)}{b/d}$ and variance $\sigma^2 = \frac{(1-b/d)(S/b)}{(b/d)^2}$. This choice was due to the fact that for rare species, the Negative Binomial tends to a Poisson distribution, that is the success

probability $p = 1 - b/d$ tends to 0 and the corresponding Negative Binomial shape parameter S/b tends to infinity. In this case, it is difficult to reach convergence in the parameters posterior distributions and it is more feasible to deal with μ and σ^2.

2.5.1 Prior distributions

In order to determine appropriate prior distributions for the model parameters, we followed a two steps approach and exploit the fact that the RSA of the different samples are supposed to have the same priors. In the first round, we defined very wide prior distributions so that to be sure that the true values of parameters would be included. Then we ran ABC on all samples and kept few selected parameters: 2 values for each subject of the ELDERMETpart database; 4 for the ELDERMET database. With all these parameters we composed the posterior distributions of round 1. We fitted the posteriors with Gamma (or Beta for the mixture parameters α and β) distributions via Markov Chain Monte Carlo (MCMC) method from pymc python module. Then, we used the fitted distributions as priors for the second round, we transformed them so that to satisfy the required conditions and we re-ran the ABC fit. Finally, for each sample we considered all the accepted parameters and obtained their posterior distributions.
The chosen priors for the first round were Inverse-Gamma distributions for μ and σ and the Uniform distribution for the mixture parameters. In particular, as detailed in the results, we compared the three models 1NB, 2NB and 3NB. For the single Negative Binomial (1NB), with parameters μ and σ, we used the prior distributions:

- $\mu \sim \textit{Inv-Gamma}(1, 1)$,
- $\sigma \sim \mu + \textit{Inv-Gamma}(1, 1)$,

The mixture of two Negative Binomials (2NB) has parameters μ_1, σ_1, μ_2, σ_2, plus the mixture parameter α, and their priors were:

- $\mu_1 \sim \textit{Inv-Gamma}(1.5, 1)$,
- $\sigma_1 \sim \mu_1 + \textit{Inv-Gamma}(1.5, 1)$,
- $\mu_2 \sim \mu_1 + \textit{Inv-Gamma}(1.5, 1)$,
- $\sigma_2 \sim \mu_2 + \textit{Inv-Gamma}(1.5, 1)$,
- $\alpha \sim Uniform(0, 1)$.

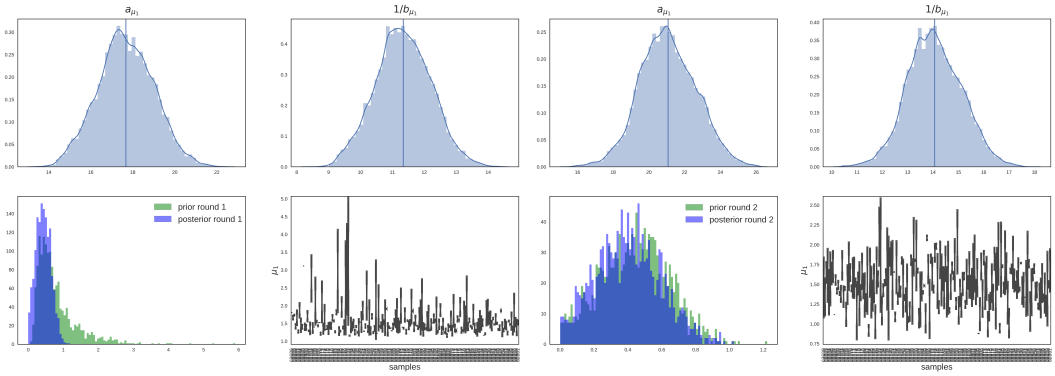
The mixture of three Negative Binomials (3NB) has parameters μ_1, σ_1, μ_2, σ_2, μ_3, σ_3 and mixture parameters α and β and its prior distributions were:

- $\mu_1 \sim \textit{Inv-Gamma}(1.5, 1)$,
- $\sigma_1 \sim \mu_1 + \textit{Inv-Gamma}(1.5, 1)$,
- $\mu_2 \sim \mu_1 + \textit{Inv-Gamma}(1.5, 1)$,
- $\sigma_2 \sim \mu_2 + \textit{Inv-Gamma}(1.5, 1)$,
- $\mu_3 \sim \mu_1 + \mu_2 + \textit{Inv-Gamma}(1.5, 1)$,
- $\sigma_3 \sim \mu_3 + \textit{Inv-Gamma}(1.5, 1)$,
- $\alpha \sim Uniform(0, 1)$.
- $\beta \sim Uniform(0, 1 - \alpha)$.

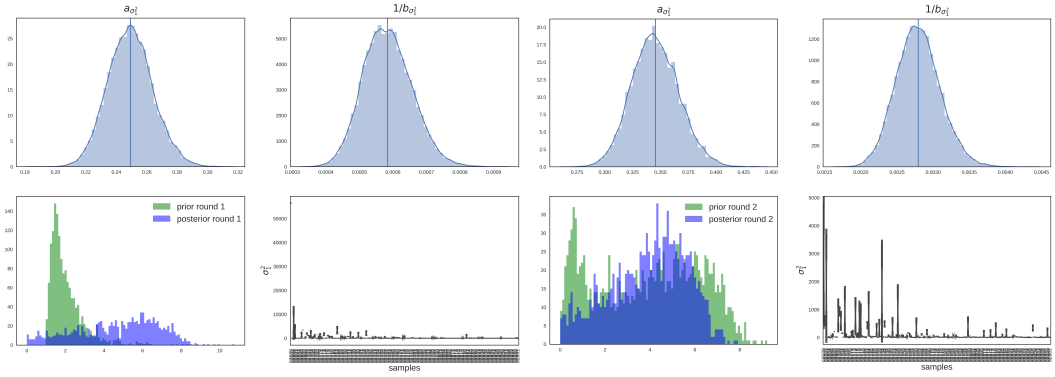
Note that, in the 2NB case (and analogously for 3NB), the definitions of σ_1, μ_2 and σ_2 are chosen so that $\mu_i \leq \sigma_i^2$ $(i = 1, 2)$, that is the success probabilities $p_i = 1 - b_i/d_i$ are less than 1. Moreover, $\mu_2 \geq \mu_1$, meaning that the average number of observed individuals is grater in $NB_2(\mu_2, \sigma_2^2)$ than in $NB_1(\mu_1, \sigma_1^2)$, or better, that $NB_1(\mu_1, \sigma_1^2)$ is associated to rarest species while $NB_2(\mu_2, \sigma_2^2)$ describes more abundant species.
After fitting the round 1 posteriors with Gamma (or Beta) distributions and obtaining their parameters a and b, the round 2 priors were derived with the following transformations, so that to take into account the parameters constraints.

- $\mu_1 \sim Gamma(a_{\mu_1}, b_{\mu_1})$,
- $\sigma_1^2 \sim \mu_1 + Gamma(a_{\sigma_1^2}, b_{\sigma_1^2})$,
- $\mu_2 \sim \mu_1 + Gamma(a_{\mu_2}, b_{\mu_2})$,
- $\sigma_2^2 \sim \mu_2 + Gamma(a_{\sigma_2^2}, b_{\sigma_2^2})$,
- $\mu_3 \sim \mu_1 + \mu_2 + Gamma(a_{\mu_3}, b_{\mu_3})$,
- $\sigma_3^2 \sim \mu_3 + Gamma(a_{\sigma_3^2}, b_{\sigma_3^2})$,
- $\alpha \sim Beta(a_\alpha, b_\alpha)$,
- $\beta \sim Uniform(0, 1 - \alpha)$.

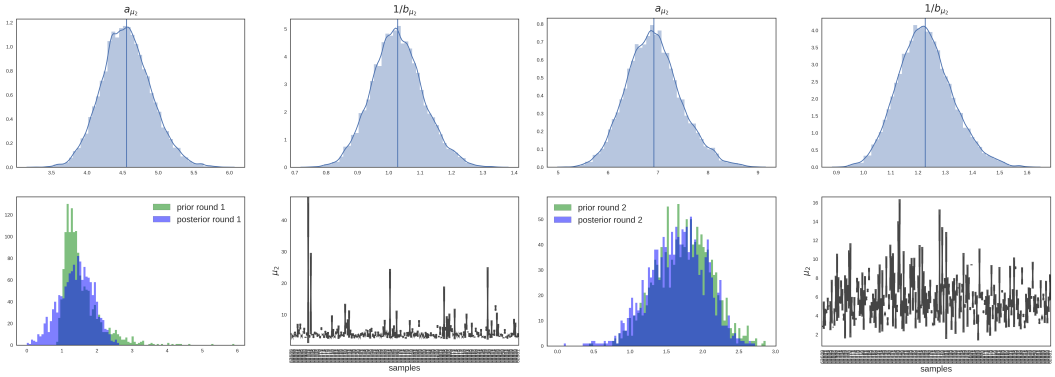
The method to obtain the prior distributions is visually displayed in the following figures. Results are shown for the ELDERMETpart dataset, when the RSA was obtained with a similarity threshold of 97% and was fitted with the 2NB model. Each figure is composed by two subfigures ((a) and (b)), each containing four plots. The left-hand subfigures refer to the first round, while the right-hand ones refer to the second round. The four graphs in each figure refer to the following plots. Bottom-left: prior (green) and posterior (blue) distribution of the parameter for the corresponding round. Note that the priors in round 2 are slightly different from round 1 posteriors because of the transformations detailed above. The posterior distributions were fitted with Gamma (or Beta for α) distributions and the upper-left and upper-right plots show the Gamma (or Beta) parameters a and b found by MCMC fit. Finally, the bottom-right plot shows for each sample the boxplot of the parameters used in this elaboration.
It is clear from the bottom-left plots of all figures in the left column that in the first round the priors (labeled as prior round 1) were chosen pretty wide. This, for instance, allows the new priors (posterior round 1) of σ_1^2 and σ_2^2 to be properly found even if the initial guesses were not peaked on their values. Moreover, the bottom-left figure in the column on the right show that if we iterated the process computing the posteriors of round 2, that would become the priors of an eventual round 3, we would find distributions that are very similar to the priors (prior round 2), meaning that the latter were actually already appropriate. Finally, we remark that our assumption was that the parameters of all samples are drawn from the same prior distributions and thanks to this consideration, overfitting was avoided by considering all samples at the same time, rather than computing the priors for each RSA separately.



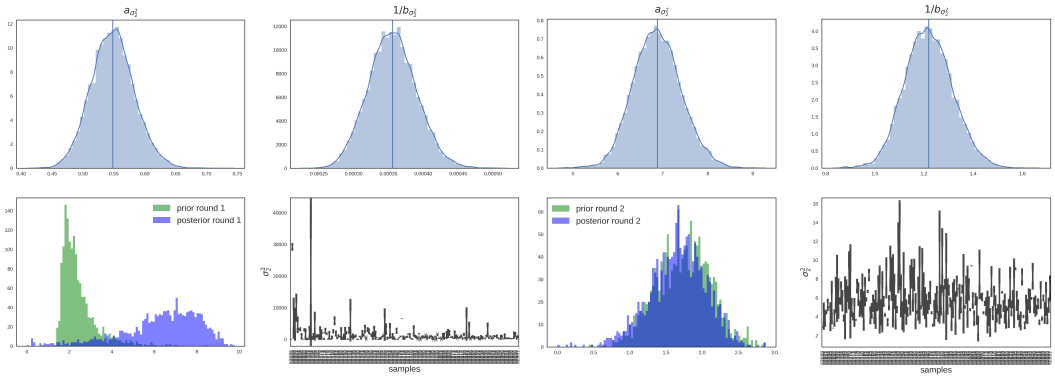
(a) μ_1 - round 1. (b) μ_1 - round 2.



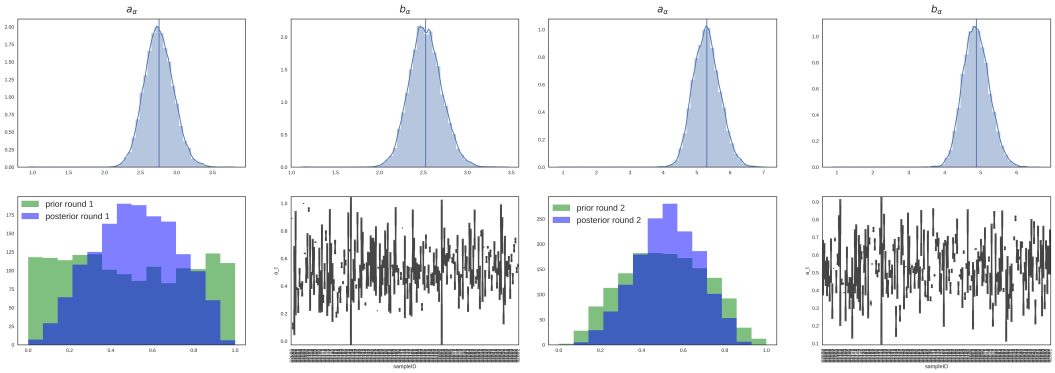
(a) σ_1^2 - round 1. (b) σ_1^2 - round 2.



(a) μ_2 - round 1. (b) μ_2 - round 2.



(a) σ_2^2 - round 1. (b) σ_2^2 - round 2.



(a) α - round 1. (b) α - round 2.

Figure 2.5: Prior and posterior distributions of the 2NB model parameters for round 1 (left) and round 2 (right) of the ELDERMETpart dataset fitting. In each quartet plot, the top graphs are the estimated parameters for the posterior distribution, while the bottom left figure represents the prior (green) and posterior (blue) distributions considering all samples together and the bottom right plot shows the posteriors for each sample separately.

2.5.2 Acceptance criterion

The similarity measure used to compare the generated and observed datasets was ground on the two samples Chi-Squared computed over the Preston plot:

$$\chi^2_{Skellam} = \sum_i \frac{(G_i - O_i)^2}{(G_i + O_i)}. \tag{2.20}$$

Here G_i refers to the value of the $i-$th bin of the Preston plot simulated by ABC, while O_i refers to the data. The variance in the denominator is the Skellam distribution variance $\sigma^2 = G_i + O_i$.
We chose to accept a set of simulated parameters when the $\chi^2_{Skellam}$ associated p-value was less than 0.5. When p-value $= 0.5$, the computed $\chi^2_{Skellam}$ corresponds to the median of the distribution, meaning that the deviations between O and G are comparable with the variance, that is what you would expect statistically. The median value, in fact, corresponds to $\chi^2_{Skellam}/df \sim 1$, that is the case in which for each bin the deviation between O_i and G_i has the same order of magnitude as the theoretical standard deviation. Moreover, we added a further condition on each bin, according to which each bin of the simulated Preston plot has to be at most 30% different from the empirical one. In this way, we allowed a bigger error in higher bins, that, in GM data, refer to the rarest species. These bins are in fact less precise for two reasons: first because sequencing errors and inaccuracies due to the clustering method mainly affect the estimate of the number of species with one or few individuals; secondly because in the Preston plot each bin contains the sum of the number of species that have an abundance in the bin range. So, while in the first bin we only count species with one individual, in the second one we count species with 2 or 3 individuals and in the last one, for example, we may count species with abundance between 2^{11} and 2^{12}. It is then clear how the highest bins, that refer to the left-hand side of the curve, are those with the biggest uncertainty.
Finally, at both fitting rounds, the parameters posterior distributions were computed considering all the accepted values.

2.5.3 Model selection

Model selection was performed computing the Akaike Information Criterion (AIC), in order to take into account the number of parameters in the 3 nested models 1NB, 2NB

and 3NB. For each sample, we considered all the accepted set of parameters and for each of them we computed the corresponding theoretical distribution $G = P_{RSA}(\vec{\theta})$. We obtained the residuals between G and the data O over the Preston plot bins as

$$RSS = \sum_{i=1}^{n} (O_i - G_i)^2, \tag{2.21}$$

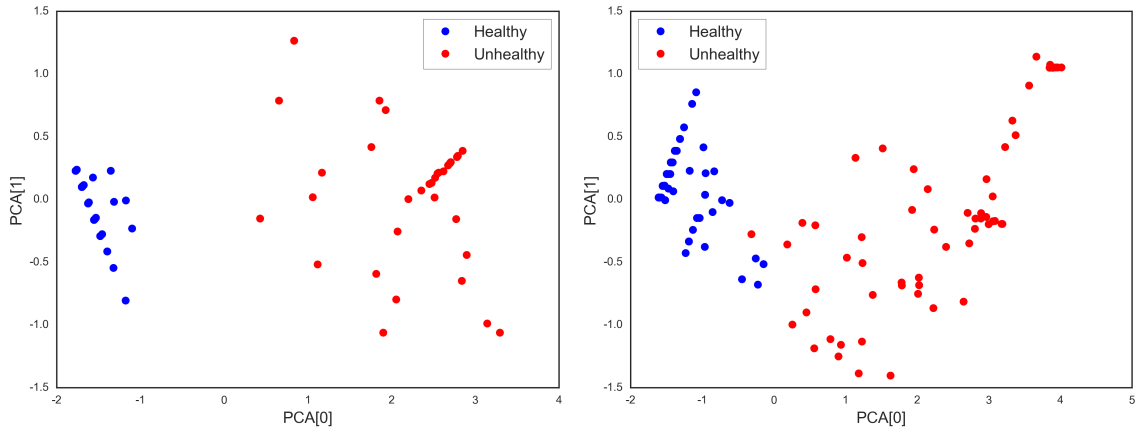
where n are the number of bins. Finally, we derived the AIC as

$$AIC = 2k + nlog(RSS/n), \tag{2.22}$$

where k is the number of model parameters [72]. When no set of parameters was accepted over the 10^7 trials, an AIC equal to the maximum AIC found for the considered model was assigned to the sample.

2.6 Predictive model based on biodiversity

As mention earlier, the elderly subjects may be divided in two groups (healthy and unhealthy) according to their physical and cognitive abilities measured by the FIM, Barthel and MMSE indices. This subdivision also reflects the differences in residence settings and dietary habits as outlined in Sec. 2.1. Running a Principal Component Analysis based on these covariates, shows that the healthy and unhealthy people cluster in fact in two separate groups (see Fig. 2.6), where the healthy one is defined by: Barthel index ≥ 15, MMSE > 24, FIM > 100 and community-dwelling or day-hospital residence setting.



(a) PCA ELDERMET. (b) PCA ELDERMETpart.

Figure 2.6: Elderly subjects represented by the first and second component of PCA. Dots color represent the subject health state according to our classification.

We assessed the discrimination and prediction accuracy of biodiversity, performing a Leave-One-Out Cross Validated (LOO CV) logistic regression. The predicted variable was the health state (healthy or unhealthy) and the covariates were the biodiversity indices.
Moreover, we compared the prediction power obtained using the biodiveristy index from our modeling, i.e. θ_1 and θ_2 for the 2NB model, with the one computed using the first

and second order Hill's numbers

$$Hill_1 = exp\left(-\sum_{i=1}^{S} p_i log(p_i)\right)$$
$$Hill_2 = \left(\sum_{i=1}^{S} p_i^2\right)^{-1} \tag{2.23}$$

that are commonly used biodiversity indices simply based on the species relative abundances, respectively obtained as transformations of Simpson and Shannon indices (see Sec. 0.5.1).
The prediction accuracy was measured using the Area Under the Receiver Operating Characteristic Curve (AUC of ROC or briefly AUC). The LOO CV logistic regression, iteratively use all but one sample as training set, find the regression coefficients for such set and then predict the outcome (health state) for the left out sample. The predicted value, in general, will be a real number so that, in order to assign the test sample to one or the other class, the classifier has to decide a threshold $\tilde{y}$, and assign the sample to one class (i.e. healthy) if the output is higher than $\tilde{y}$, or to the other one (i.e. unhealthy) otherwise. Obviously, if we change such threshold, the number of true and false positives (healthy) will also change. The ROC curve is built considering several possible thresholds and plotting the rate of true positives versus the rate of false positives. Note that since we are considering the true/false positives rates, the x-axis and y-axis of the plot will both range from 0 to 1.
If the binary classifier was random, i.e. it predicted the sample to be healthy with probability 0.5 and unhealthy with probability 0.5, the ROC curve would be the line with slope 1 and the area under it would be AUC= 0.5 (worst scenario). If instead the prediction was perfect, meaning that all and only the healthy elderly people were classified in the healthy class, than even when the number of false positives is zero ($x = 0$), the rate of true positives is 1 ($y = 1$). In this case the ROC curve is a vertical line from 0 to 1 for $x = 0$ and then it becomes the horizontal line $y = 1$. The area under the curve will be in this case AUC= 1 (best scenario).

Chapter 3

Results

3.1 Model selection and goodness of fit

We ran the ABC rejection algorithm (see Sec.2.5) for both the ELDERMET and the ELDERMETpart datasets using three similarity thresholds, 97%, 95% and 93% and we fitted the data with the three models that we want to test: the purely neutral model that predicts a Negative Binomial (1NB), the hybrid two neutral niches model described by a mixture of two Negative Binomials (2NB) and the hybrid three neutral niches model in which we consider 3 non-interacting populations and a mixture a three Negative Binomials (3NB).
In order to compare the performances of the three models we used the Akaike Information Criterion detailed in Sec. 2.5.3. The AIC mean and 95% confidence intervals (CI) for the three models are shown in the following tables for the three similarity thresholds 93%, 95% and 97%.

model	AIC 97%	CI 97%	AIC 95%	CI 95%	AIC 93%	CI 93%
1NB	100.33	(97.62,103.05)	76.00	(74.02,78.00)	70.22	(68.54,71.90)
2NB	83.08	(81.20,84.96)	75.04	(73.28,76.81)	72.11	(70.56,73.66)
3NB	89.69	(87.81,91.58)	83.96	(82.20,85.74)	80.44	(78.86,82.02)

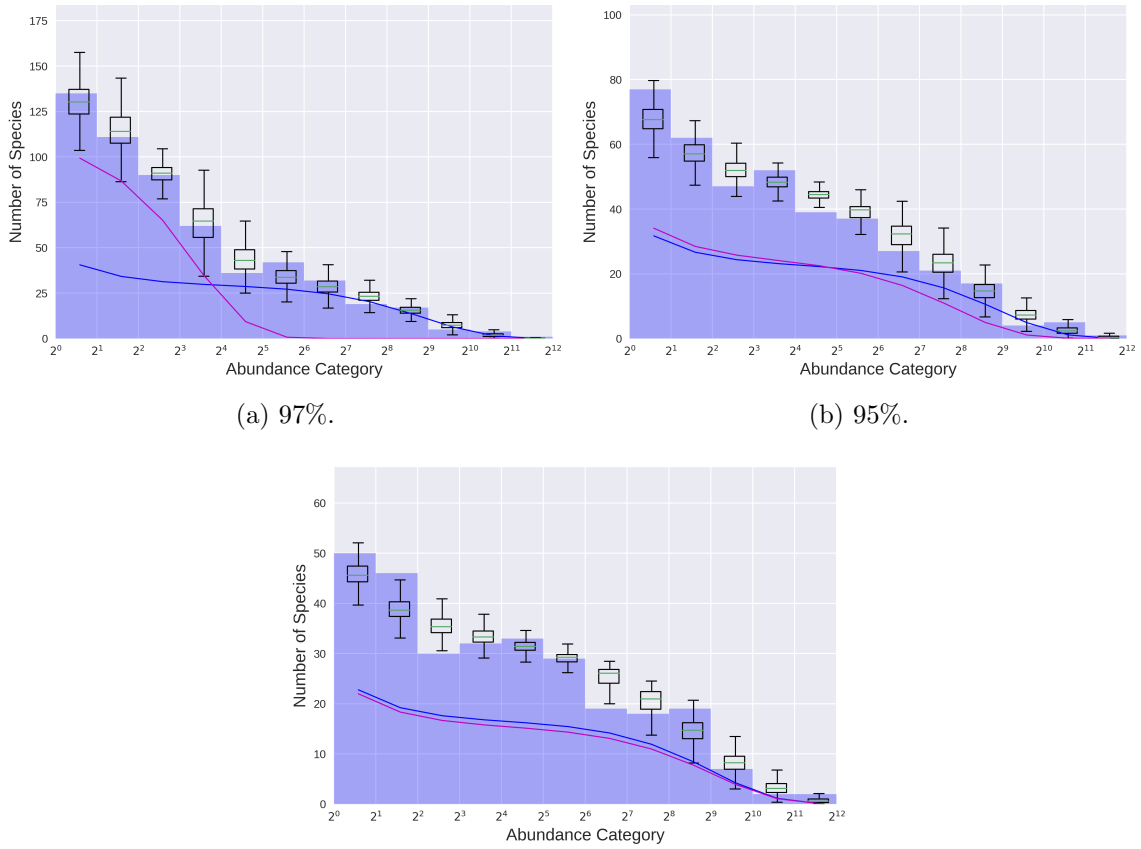
Table 3.1: Models AIC with 97%, 95% and 93% similarity threshold for the ELDERMETpart dataset.

model	AIC 97%	CI 97%	AIC 95%	CI 95%	AIC 93%	CI 93%
1NB	85.30	(83.17,87.43)	74.42	(72.71,76.12)	70.76	(69.02,72.51)
2NB	75.69	(74.24,77.14)	69.96	(68.66,71.25)	66.88	(65.65,68.11)
3NB	82.10	(80.66,83.55)	77.31	(75.99,78.64)	74.38	(73.15,75.60)

Table 3.2: Models AIC with 97%, 95% and 93% similarity threshold for the ELDERMET dataset, considering the three time points together.

The 2NB model is always preferable to the 3NB, having lower AIC. For what concerns the comparison between model 2NB and 1NB, when considering high similarity threshold 2NB is clearly better. At 93% of similarity, 2NB is still found to have better performances than 1NB in the ELDERMET dataset, while in the ELDERMETpart dataset the two models are comparable with a slight preference for model 1NB. This could be explained

noting that at 93% we are considering larger OTUs, that include many different species. Thus, we are somehow observing an average behavior of the ecosystem and this is why the deviation from neutrality that we see at 97% may be masked.
Figure 3.1 shows the results for one ELDERMETpart sample when the RSA was computed with the three similarity thresholds and fitted with the 2NB model. The histogram is the empirical Preston plot. In each bin, we also show the boxplot obtained considering the ABC accepted simulations. The green line is the median, while the box represents the first and third quartiles (interquartile range), that is the likely range of variation. The vertical lines, instead, are drawn so that to include the most extreme, non-outlier data points. In each plot we also show two continuous lines that are the two Negative Binomials of the 2NB mixture obtained with the medians of the parameters posterior distributions. We will label the magenta Negative Binomial as NB_1 and its biodiversity number as θ_1. The blue Negative Binomial, instead, will be indicated as NB_2 and its biodiversity number will be θ_2. Note that NB_1 is the one describing the rarest species, that is the left-hand side of the Preston plot, while NB_2 fits the abundant species that form the distribution tail.



(a) 97%.

(b) 95%.

(c) 93%.

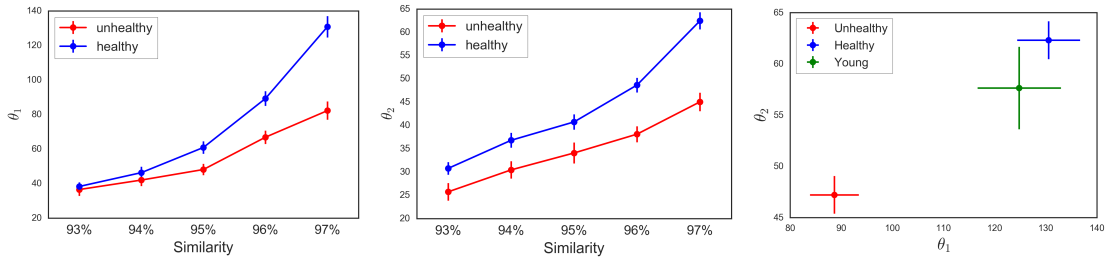
Figure 3.1: Results of ABC fit with 2NB model on a sample from the ELDERMETpart dataset for different similarity thresholds. Boxplots represent the ABC accepted simulations. Continuous lines represent the the Negative Binomials that summed up give the 2NB mixture.

Comparing the three graphs in Fig. 3.1 we may notice that the RSA is skewer for higher similarities (97%). In this case, in fact, the GM species are clustered into thiner OTUs and, as a consequence, we will have more OTUs with few individuals than when considering a smaller threshold. Since biodiversity is related to the number of rare species, besides the total number of species, we may expect it to be greater when the similarity threshold used

is higher, and this is in fact what we will observe in Fig. 3.2 and 3.3. Another difference that comes into eyes is that, while at 97% the two Negative Binomials of the mixture are clearly different, lowering the similarity threshold they become more and more similar. This means that at 97% the two populations are separated, while at 93% the neutrality approximation would be more accurate. As mention before, the reason for this relies in the fact that when using a lower threshold we allow the OTUs to put together more different bacteria that, for instance, may belong to the population described by NB_1 or to the one described by NB_2. In this way, the properties of pretty different species are averaged and the differences in the two populations may be not visible anymore.

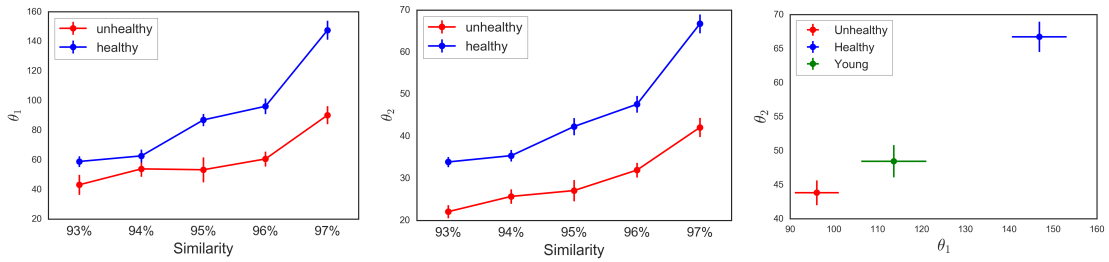
3.2 Healthy aging prediction

Since the 2NB model performs better the the 1NB and 3NB, especially at high similarity thresholds, when most of the biodiversity is captured and greater differences are appreciable (see also Fig. 3.2 and 3.3), we fulfilled the prediction analysis considering only this case.



(a) θ_1 versus similarity. (b) θ_2 versus similarity. (c) θ_1 versus θ_1.

Figure 3.2: Trend of θ_1 and θ_2 with similarity threshold and discrimination between healthy and unhealthy subjects. Results are shown for the ELDERMETpart dataset.



(a) θ_1 versus similarity. (b) θ_2 versus similarity. (c) θ_1 versus θ_1.

Figure 3.3: Trend of θ_1 and θ_2 with similarity threshold and discrimination between healthy and unhealthy subjects. Results are shown for the ELDERMET dataset at $t = 0$.

We associated to each sample the median of its θ_1 and θ_2 and computed their mean for the two classes healthy (blue) and unhealthy (red). Fig. 3.2 and 3.3 (a) and (b) show the trend of θ_1 and θ_2 when varying the similarity threshold. The vertical bars indicate the standard error of the mean. As expected, both biodiversity indices increase with the similarity threshold. Moreover, the two groups are well separated and the difference is higher for higher similarities. This also confirms that at low similarities we are observing a sort of average of the GM ecosystem and biodiversity differences are masked. The (c) graph in Fig. 3.2 and 3.3, instead, show the discrimination of the two classes in a θ_2 versus θ_1 plot. Note that healthy elderly subjects have higher biodiversity, both considering θ_1 and θ_2. Moreover, the young group, represented in green, stands between the two classes.

This may be due to the fact that the elderly classes are extreme cases or that the young group is heterogeneous and may include people that will become part of the healthy or the unhealthy class in their old age.
The predictive performance of θ_1 and θ_2 have been tested through a Leave-One-Out Cross Validated logistic regression, as detailed in Sec. 2.6. In particular, the predicted variable was the health state of the elderly subject and the covariates were θ_1 and θ_2. The logistic modeling was also performed using two others commonly used biodiversity indices that are based only on the relative abundances of the GM OTUs: the first and second order Hill's number $Hill_1$ and $Hill_2$ (see Sec. 2.6 for definitions). Tables 3.3 and 3.4 show the AUC of ROC (see Sec. 2.6) values for the two datasets, considering the different similarity thresholds. For the ELDERMET dataset, results are also shown at different time points. Note that from time $t = 0$ to time $t = 2$, the clinical parameters slightly change, but the health state class remains the same for all individuals. The AUC values are smaller for $t = 1$ and $t = 2$ in part because the number of individuals at $t = 0$ is higher. In the ELDERMETpart dataset, instead, we have only one time point. The AUC values here are smaller than in the ELDERMET dataset probably because the two classes are less separated, as can be viewed from Fig. 2.6 in Sec. 2.6. As expected, the prediction accuracy decreases with the similarity threshold for both datasets. This is also true for the Hill's numbers. However, in all cases, the performances of the indices obtained through our modeling (θ_1 and θ_2) are better then for these common indices, meaning that thanks to the inclusion of stochasticity, that takes into account experimental and ecological noise, we obtain a more suitable and useful measure of biodiversity.

Index	$97\%; t0$	$95\%; t0$	$93\%; t0$	$97\%; t1$	$95\%; t1$	$93\%; t1$	$97\%; t2$	$95\%; t2$	$93\%; t2$
θ_1, θ_2	0.851	0.827	0.815	0.728	0.659	0.660	0.700	0.686	0.649
θ_1	0.831	0.755	0.730	0.733	0.658	0.644	0.638	0.637	0.613
θ_2	0.861	0.831	0.824	0.727	0.670	0.672	0.705	0.706	0.676
$Hill_1$, $Hill_2$	0.753	0.645	0.591	0.657	0.598	0.543	0.639	0.588	0.544
$Hill_1$	0.589	0.552	0.501	0.520	0.535	0.486	0.654	0.632	0.561
$Hill_2$	0.405	0.421	0.298	0.255	0.417	0.371	0.581	0.571	0.507

Table 3.3: AUC results for the ELDERMET dataset.

Index	$97\%; t0$	$95\%; t0$	$93\%; t0$
θ_1, θ_2	0.741	0.694	0.665
θ_1	0.733	0.677	0.614
θ_2	0.745	0.706	0.677
$Hill_1$, $Hill_2$	0.703	0.676	0.653
$Hill_1$	0.639	0.624	0.620
$Hill_2$	0.559	0.555	0.570

Table 3.4: AUC results for the ELDERMETpart dataset.

Chapter 4

Discussion

The composition of the intestinal microbiota in older people (> 65 years) is extremely variable between individuals, and differs from the core microbiota and diversity levels of younger adults [73] [74]. Furthermore, the variation in the intestinal microbiota of older subjects has an impact on immunosenescence and frailty [55]. Claesson et al. [55] found out that community-dwelling elderly subjects and those in long-term residential care (long-stay) were characterized by different Gut Microbiota compositions. Community-dwelling subjects had a higher proportion of phylum *Firmicutes*, genera *Coprococcus* and *Roseburia* (of the *Lachnospiraceae* family) and unclassified reads. Moreover, they clustered with the younger control group. Long-stay microbiota, instead, included more bacteria of the phylum *Bacteroidetes* and genera *Parabacteroides*, *Eubacterium*, *Anaerotruncus*, *Lactonifactor* and *Coprobacillus*.
The authors also propose three diversity measure for the Gut Microbiota, the number of unique OTUs, Shannon index [11] and phylogenetic diversity [75], and show that Microbiota diversity decreases when shifting from diet group DG1 ('low fat/high fibre') to diet group DG4 ('high fat/low fibre'). This also implies that the diversity of long-stay subjects is lower than for the community-dwellers, since the resident setting is highly correlated with diet in their dataset.
Later on, Jeffery et al. [61] asserted that maximal microbiota diversity was not the variable most strongly associated with health but that a particular microbiota composition typifies healthy community-dwelling subjects. They noticed that subjects with low microbiota diversity often have low FIM values. However, their result was ambiguous because the presence of bacteria associated with the long-stay group, that is characterized by lower FIM values, actually increased diversity.
Here, we derived a biodiversity measure based on a stochastic modeling of the population dynamics and showed that a better estimate of the Gut Microbiota diversity can be obtained by taking noise into account.
We proved that the GM ecosystem shows deviations from the neutrality assumption of species equivalence, especially when considering low phylogenetic levels, that roughly correspond to the genus classification. The simplest relaxation of neutrality requires to consider two non-interacting populations, each one characterized by its dynamic rates. This assumption turns out to be enough to obtain a good description of the data. At higher phylogenetic levels, the differences between the two populations are less visible. In fact, if for instance we considered the phylum classification, we will have that a single phyla includes many different genera and species. These, in particular, may belong to both the two populations that we distinguished at lower phylogenetic levels and what we observe will be a summary dynamics of the two populations together, that will not be distinguishable anymore.
Our modeling returns a biodiversity measure based on two indices, that regard the two

populations. We proved here that this measure may be used to predict healthy aging. The average behavior observed at high phylogenetic levels has also the effect to diminish the variability in diversity between different subjects. For this reason, the prediction accuracy of the biodiversity indices for healthy aging, was computed focusing on the lowest phylogenetic level, that is the one obtained cluestering the 16S rRNA sequences into OTU with a similarity threshold of 97%.
We noticed that the elderly population could be divided in two groups, one that is characterized by healthy clinical parameters and the other one that includes unhealthy subjects. Considering these two classes, we performed a logistic regression with Leave One Out Cross-Validation to establish the prediction accuracy of our biodiversity measure. It turned out that we obtain a much better result than using common diversity indices, such as the Shannon index used by Claesson et al., that is simply based on the relative abundance of species. Thus, the stochastic modeling of the Gut Microbiota may solve the ambiguities about biodiversity that arose in previous works [61] and provide a better measure to be investigated when we aim to determine how to improve our life style in order to achieve an healthier aging.

Part II
Clustering 16S rRNA into OTUs: a new parameter-free approach

Chapter 5

Introduction

We have seen in Part I the importance of re-defining the concept of species, particularly when the aim of the analysis is to describe the biodiversity of ecosystems, like the Gut Microbiota. There, we also highlighted that such re-definition should be based on phylogeny rather than taxonomy. From an ecological point of view, in fact, human made taxonomic classifications are not of great significance and may even lead to equivocal results. Moreover, since the vast majority of bacterial organisms is not accessible by traditional clonal culture techniques [76][77], it is necessary to base the analysis on sequencing data, so that to be able to sample the whole microbial community.The recommended approach, then, is to obtain and sequence samples of the DNA present in the ecosystem and to determine the evolutionary distances between the collected bacteria, so that to define species as sets of closely related genomes, that may be detected by comparing and *de novo* clustering highly conserved sequences. The newly defined species are usually called Operational Taxonomic Units (OTUs). It is clear that one of most crucial steps in this procedure is the choice of the algorithm used to cluster sequences into OTUs and this is the topic of the following work. In particular, here, after summarizing some of the most commonly used clustering methods, we will introduce a new approach based on the estimation of the sequences density, that was originally proposed in 2014 by A. Rodriguez and A. Laio [78] and that was made fully unsupervised and parameter-free by M. d'Errico E. Facco and A. Laio in 2016 [79], that also collaborated to this work.

5.1 The concept of OTU

The DNA sequence usually employed for the re-definition of bacterial species is the one that codes for the 16S rRNA, that is a component of the 30S small subunit of prokaryotic ribosomes. This is a highly conserved sequence, meaning that its probability of encountering mutations is relatively low. As a consequence, we would expect two bacteria of the same species to have very similar 16S rRNA, with at least 97% of the bases in common. The choice of 97% of similarity for species definition is an ordinary one and is based on the assumption that rare mutations and possible errors due to the sequencing procedure generate at most a difference of 3%. If, instead, we considered two bacteria that are phylogenetically distant and belong to different species, their 16S rRNA would have evolved independently for a longer time, gathering more mutations, and we would expect their differences to be greater than 3%. It is logical, then, to base the re-definition of species on the differences between 16S rRNA sequences, so that to obtain a classification that reflects the phylogenetic relationships between bacteria. The newly defined species are called Operational Taxonomic Units (OTUs) and are obtained by *de novo* clustering the collected 16S rRNA sequences. With *de novo* we mean that the clustering procedure should not be based on taxonomic databases, but should only rely on the properties of

the sequences. In fact, even if the use of taxonomy may be appealing because it enables to place labels onto sequences, indicating their relationships to previously cultured and characterized microbes, it is not the recommended approach to detect bacterial species. Taxonomic classifications are in fact human made and there are lots of examples of organisms that belong to the same species and have different phenotypes and organisms with the same phenotype belonging to different taxonomic lineages. Moreover, several organisms are still unclassified and the numerous existing different curated taxonomy outlines contain significant conflicts with each other [80]. In order to estimate the evolutionary relationships of organisms, genes, species, etc. the tool to be used is phylogenetic analysis and *de novo* OTUs computation, that despite not enabling to give names to species, has the great advantage of avoiding the loss of information due to a taxonomic classification.

5.2 Clustering sequences into OTUs

Several clustering methods have been proposed to fulfill the task of computing OTUs, and the choice of which one to use is critical. Ideally, the selected algorithm should be able to cluster together similar sequences and separate those that are more different, but this is not a simple requirement for sequencing data, because of their complex and hierarchical structure. Moreover, the method should also optimize computational costs in terms of both time and memory consumption, that may become important when dealing with these kind of data.
Many of the available algorithms are based on fixing a similarity threshold so that two sequences will belong to the same cluster if their distance is less than the threshold and will be placed in two separate clusters otherwise. For instance, hierarchical clustering, that is implemented in mothur [67], requires to first compute the distance matrix of the sequences and then to cluster them at a specific level of sequence dissimilarity according to some criteria (nearest neighbor, furthest neighbor or average neighbor). The computational complexity, in this case, is $O(N^2)$, where N is the number of sequences, and may pose a significant computational bottleneck when processing large-scale datasets [81]. For this reason, greedy heuristic algorithms such as CD-HIT [82] and UCLUST [83] have been developed to assign sequences into OTUs with lower time and space complexity.
However, these approaches still have in common with the previous ones the requirement of a fixed threshold, and this is not necessarily the best option. Specifically, Schloss et al. [80] asserted that it is not possible to define distance-based delineations for different taxonomic levels. They unveiled, in fact, that the genetic distance between the most disparate full-length 16S rRNA gene sequences within a named taxonomic group represented a continuum for each level in the hierarchy. Furthermore, the distances within a taxonomic group are not evenly distributed within the group. Genera such as Bacillus, Bacteroides, Clostridium, and Pseudomonas were very broad, with mean distances ranging from 4 to 9%, while genera such as Bradyrhizobium, Cetobacterium, Pseudoalteromonas and Staphylococcus were much tighter, with average distances ranging from 1 to 3%. Moreover, considerable overlap in the maximum intra-taxon distances between taxonomic levels was observed and groups at every level in the taxonomic outline were found to have maximum intra-group distance less than 1.5%. Similarly, also the variation in phylogenetic diversity represented at each taxonomic level was found to represent a continuum.
As mentioned before, we would not suggest to rely OTUs computation only on taxonomic classifications. However, the remarks pointed out by Schloss et al. indicate that a correct grouping of 16S rRNA sequences should not be based on a fixed similarity threshold, but should preferably rely on the true structure and topography of the data.

Chapter 6

Material and Methods

In the following, first, we review three common methods used to cluster 16S rRNA sequences into OTUs: mothur [67] and UCLUST [83], that were mentioned in the Introduction, and CROP [84], that has the peculiarity of using a soft threshold. Then, we introduce a new method developed by d'Errico et al. [79] (LOCk-NN) and we show this can be adapted to deal with sequencing data. Finally, we will detailed how a simulated dataset was generated in order to test and compare the performances of the above mentioned methods.

6.1 Sequences distance and similarity

Before exploring the clustering methodologies, let us mention that every method that aims to compute OTUs from 16S rRNA data requires the definition of a distance between sequences, or equivalently of a similarity measure. In particular, mothur and LOCk-NN are both preceded by the computation of the dataset distance matrix, and this was performed with mothur [67] using the following procedure. First, sequences were trimmed so that to exclude unreliable reads or reads parts, as we will detail when describing the simulated dataset. Then, preprocessed sequences were aligned against the silva reference database [85]. And finally, the distance matrix was computed using default options. The gap policy was to count a string of gaps as a single gap and to penalize also terminal gaps. Gaps and mismatches were given a penalty of -1 while matches were given a score of $+1$. Using these criteria, similarity was computed as the number of matches divided by the length of the shortest sequence (excluding gap extensions) and distance was obtained as $(1 - similarity)$. Analogously, UCLUST computes the similarity between sequences by first aligning them. Terminal gaps are discarded before identity is calculated, while internal gaps always count as mismatches. Then, similarity is computed using the same scores and procedure as in mothur. The main difference, as we will detail, is that UCLUST does not require to compute all pairwise distances, but these are obtained only for an appropriate subset. Finally, also CROP does not compare all sequences versus all but first splits the dataset in a certain number of blocks in which it computes the distance matrix. The pairwise alignments in CROP are obtained using the Needleman-Wunsch algorithm [86], and the distances are determined with the Quickdist algorithm [87], which ignores terminal gaps and treats gaps of any length as single mismatches.

6.2 mothur

Mothur [67] implements three hierarchical algorithms that compute clusters starting from a distance matrix according to the nearest neighbor, furthest neighbor or average neighbor rule. We chose to use the average neighbor method since it was proven to be the one with

best performances [80]. Consequently, we obtained clusters in which the average distance between their elements was $X\%$, where X is the user defined threshold that we chose to be 7%, 5% or 3%. In the following we will refer to these thresholds in the form of similarities, i.e. 93%, 95% and 97%, respectively.

6.3 UCLUST

UCLUST [83] starts with an empty database in memory, that will contain the OTUs seeds or representative sequences. The algorithm reads the sequences in input order, takes the first one as first seed, and then processes all the others according to the following statement: if a sequence is similar to an OTU representative sequence (database seed) within a predefined similarity threshold, then the query is assigned to its OTU; if a sequence is instead not similar to any seed, then it will become the representative sequence for a new cluster and will be added to the seeds database.
The comparison between the query sequence and the database seeds is performed using the USEARCH algorithm [83]. USEARCH exploits that fact that similar sequences tend to have several short words in common. These words have a fixed length k, and are sometimes called k-mers. Unlike other programs that use k-mer counting, USEARCH does not attempt to estimate sequence identity from the number of matching k-mers. This is because identity correlates only approximately with the word count and does not give an accurate estimate, especially for lower identities. Instead, USEARCH uses the word count to prioritize the database search. For every query, it sorts the database seeds in order of decreasing unique word count, so that, if a hit above the similarity threshold exists, this will be found among the top sequences of the ordered database. The main advantage of this technique is a great improvement in speed, due to the fact that if the top seeds are rejected and no match is found, than the algorithm stops and create a new cluster without investigating the rest of the database.
As mentioned before, UCLUST processes the sequences in input order and this implies that the first sequences in the input file will be more likely to become the cluster seeds. For this reason, the clustering procedure is usually preceded by an appropriate sorting of the sequences. In particular, after trimming the reads at 350 bp, so that to exclude the last bases that are known to be noisy, we sorted them by decreasing length, following the idea that longer sequences are more reliable and hence more useful to be used as cluster seeds.

6.3.1 CROP

Moving towards the idea of a parameter-free clustering method that relies on the structure of the data, Hao et al. [84] proposed a Gaussian mixture model-based algorithm termed Clustering 16S rRNA for OTU Prediction (CROP). CROP adopts an unsupervised probabilistic Bayesian method and uses a soft threshold for defining OTUs, bypassing the setting of an often subjective hard cut-off threshold and aiming to reduce the effects of PCR and sequencing errors in inferring OTUs.
In CROP, a Gaussian mixture model is applied to the 16S rRNA sequcences, that are indicated as $x = (x_1, ..., x_N)$ and are assumed to be independently drawn from a mixture density with k clusters, where k is an unknown parameter

$$p(x|k, \pi, \mu, \sigma^2) = \sum_{i=1}^{k} \pi_i f(x; \mu_i, \sigma_i^2) \tag{6.1}$$

where $\mu = (\mu_1, ...\mu_k)$ and $\sigma^2 = (\sigma_1^2, ...\sigma_k^2)$ indicate the centers and variances of the k clusters, while $\pi = (\pi_1, ..., \pi_k)$ are the non-negative mixture proportions that sum up to 1.

Hao et al. assumed that the probability of a sequence x_i to belong to cluster j, $f(x_i; \mu_j, \sigma_j^2)$, is given by a modified univariate Gaussian distribution with mean μ_j, corresponding to the center of cluster j, and variance σ_j^2:

$$f(x_i; \mu_j, \sigma_j^2) = \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-\frac{D^2(x_i, \mu_j)}{2\sigma_j^2}}, \tag{6.2}$$

where $D(x_i, \mu_j)$ denotes the distance between the sequence x_i and μ_j. The the mixture parameter and the variance are here supposed to have prior distributions

$$\begin{aligned} &\pi|k \sim Dirichlet(\gamma, ..., \gamma) \\ &\sigma_i^2 \sim Inverse - Gamma(\alpha, \beta) \end{aligned} \tag{6.3}$$

The cluster center μ_j, instead, should correspond to one of the sequences of the dataset and is thus chosen randomly from $x_1, ..., x_N$ without replacement.
The parameters (π, μ, σ^2) are determined maximizing the posterior distribution $P(\pi, \mu, \sigma^2|x)$, that is constructed simulating a birth death process and using the Markov Chain Monte Carlo (MCMC) method. If the number of clusters at a certain step of the process is k, in the next step we may have $k+1$ clusters, because of the birth of a new cluster, that occurs with probability $P_B/(P_B + P_D)$, or $k-1$ clusters, due to the death of an existing one, that occurs with probability $P_D/(P_B + P_D)$. The birth probability P_B is chosen to be constant, while the death probability P_D is defined as $P_D = \sum_{j=1}^{k} d_j$, where d_j is the probability for a cluster to die.
If a birth happens, the $(k+1)$-th cluster is generated using a new center μ_{k+1} randomly chosen among all sequences that are not already the center of any cluster. The probability of choosing a certain x_i is given by

$$P(\mu_{k+1} = x_i) \propto \frac{1}{f(x_i; \mu_t, \sigma_t^2)} \tag{6.4}$$

where (μ_t, σ_t^2) are the parameters of the cluster to which x_i is assigned at the current step t. In this way, μ_{k+1} is more likely to be chosen among the sequences that are far away from the center of the cluster to which they currently belong. For other parameters of the new cluster, the priors are given as follows

$$\begin{aligned} &\pi_{k+1} \sim Beta(\gamma, k\gamma) \\ &\sigma_{k+1}^2 \sim Inverse - Gamma(\alpha, \beta). \end{aligned} \tag{6.5}$$

Then, the existing parameters need to be updated setting $\pi_j = \pi_j(1 - \pi_{k+1})$.
The death event of an existing cluster, instead, has probability

$$d_j = \frac{L_{-j}}{L}. \tag{6.6}$$

where L_{-j} is the likelihood without the j-th cluster and L is the likelihood with the j-th cluster:

$$\begin{aligned} L &= \prod_{i=1}^{N} \sum_{l=1}^{k} \pi_l f(x_i; \mu_l, \sigma_l^2) \\ L_{-i} &= \prod_{i=1}^{N} \sum_{l=1, l\neq j}^{k} \pi_l' f(x_i; \mu_l, \sigma_l^2). \end{aligned} \tag{6.7}$$

Also in the case of the death of the j-th cluster, π_j is updated for every $j = 1, ..., k$, $j \neq i$ becoming $\pi_j' = \frac{\pi_j}{1-\pi}$.
In each step of the MCMC, the hierarchical mixture model parameters are updated as follows.

1. Update the cluster to which x_i is assigned, C_i, with probability

$$P(C_i = j) = \frac{\pi_j}{\sqrt{\sigma_j^2}} e^{-\frac{D^2(x_i,\mu_j)}{2\sigma_j^2}} \tag{6.8}$$

2. Update $\pi|... \sim Dirichlet(\gamma + n_1, ..., \gamma + n_k)$, where n_j, $j = 1, ..., k$, is the numbers of data points belonging to the mixture cluster j: $n_j = \sum_{i=1}^{N} 1(C_i = j)$.

3. Update σ^2. The variances of clusters are limited with an upper bound U and a lower bound L, so that to allow clustering results at different phylogenetic levels. Thus, σ_j^2 is restarted using

$$\sigma_j^2 \sim Inverse - Gamma(\alpha, \beta'), \tag{6.9}$$

with β' satisfying $\frac{\beta'}{\alpha-1} = \frac{L+U}{2}$.

4. Update μ by sampling the center of each cluster in the following way. Consider the n_j sequences of cluster j, that are the candidate centers. For two candidate centers μ_j and μ_j', we have

$$\frac{P(\mu_j'|n_j, x_j, \sigma_j^2)}{P(\mu_j|n_j, x_j, \sigma_j^2)} = \frac{P(\mu_j') \prod_{i=1}^{n_j} f(x_{ji}; \mu_j', \sigma_j^2)}{P(\mu_j) \prod_{i=1}^{n_j} f(x_{ji}; \mu_j, \sigma_j^2)} \tag{6.10}$$

Assume that $P(\mu_j) = P(\mu_j')$. Then, for each $x_{ji} \in C_j$, define the probability of choosing x_{ji} as new cluster center as

$$\delta_i = \frac{\prod_{l=1}^{n_j} f(x_{jl}; x_{ji}, \sigma_j^2)}{\sum_{i=1}^{n_j} \prod_{l=1}^{n_j} f(x_{jl}; x_{ji}, \sigma_j^2)} \tag{6.11}$$

and accordingly sample a new center for each cluster.

Hao et al. set $\gamma = 1$, $\alpha = 2$, and $P_B = 1$. The lower and upper bounds for the variances, L and U, are instead determined according to the desired phylogenetic level. In particular, a similarity of roughly 97% is achieved setting $L = 1$ and $U = 6.25$, while 95% is obtained with $L = 2.25$ and $U = 6.25$. These thresholds, in particular, guarantee that at least 95% of the sequences have at least the required similarity.

6.3.2 CROP work flow

In CROP, the dataset is first randomly split into blocks of 100-1000 sequences each. We used 500 blocks in the first round and a block size of 400 in the others, so that the value of the block size multiplied by the maximum sequence length (350) was less than 150000, as suggested by authors.
Then, an independent Bayesian clustering is applied to each block. A distance matrix is generated for each block as outlined in Sec. 6.1. The algorithm, by default, runs 20*(block size) iterations of MCMC, considering the first 10*(block size) iterations as burn-in. We chose to run 3000 iterations. From all the iterations that follow the burn-in, the one with the largest posterior probability is chosen and reported as clustering result for this block. At this point the output is the set of clusters that resulted from all the considered blocks and the algorithm proceeds to merge them, computing for each block a slightly different distance matrix. This is a 'center sequences against clusters' matrix, whose elements (i, j) indicate the distance between the i-th center sequence and the j-th cluster. Such distance is calculated as the average distance between the i-th center sequence and 20 randomly chosen sequences from the j-th cluster, when available.

Then, a weighted Bayesian clustering is performed based on this distance matrix and using the cluster size as weight. This process will continue until one of the conditions noted below is satisfied.

1. The number of clusters is $> 90\%$ of the number of sequences.
2. The number of clusters is smaller than a predetermined threshold.
3. The process has been running for N times, where N is a predetermined threshold.

Finally, one more round of Bayesian clustering is performed on all the remaining clusters to generate the final result.

6.4 LOCk-NN: a new parameter-free approach

The main idea of the method proposed by d'Errico et al. [79] is to estimate the density distribution of the data sequences space, so that to identify clusters as density peaks. Estimating the local density in RNA real datasets, however, is a complex task, due to the fact that these data are characterized by inhomogeneous density maxima organized in hierarchical structures.
The core of the algorithm is an unsupervised density estimator, called LOCal k-Nearest Neighbor (LOCk-NN) and detailed in Sec. 6.4.1. LOCk-NN automatically detects the local length scale of density variation and also provides a measure of the estimate uncertainty and this allows us to learn the position and height of density peaks and of the saddle points between them and to accurately reconstruct complex density topographies.

6.4.1 The LOCal k-Nearest Neighbor

Let us consider a set of N data points in a d-dimensional space. As in the usual k-NN density estimator [88], we suppose that the density at each point can be approximately defined as the number k of data points in a small neighborhood divided by the volume of the neighborhood. Once the neighborhood of a point is defined, and this is the critical step, such volume will be given by the volume of the hypersphere in d dimensions with radius equal to the distance between the point and its furthest (k-th) nearest neighbor. The intrinsic dimension d of the system was inferred using the DANCo algorithm [89]. Then, for each point i we denote $(r_l)_{l=1,..k}$ the sequence of the ordered distances with the first k nearest neighbors and we compute the volume enclosed between two successive neighbors as $\Delta v_l = \omega_d(r_l^d - r_{l-1}^d)$, where the proportionality constant ω_d is the volume of the d-sphere with unitary radius. We conventionally take $r_0 = 0$. It can be proven that if the density is constant around point i, all the Δv_l are independently drawn from an exponential distribution with rate equal to the density ρ at the point i, $\Delta v_l \sim \rho \cdot exp(-\rho \Delta v_l)$. Therefore, given the probability of observing the first k neighbors at distances r_1, r_2,...,r_k, the log-likelihood function of the parameter ρ is

$$\mathcal{L}(\rho|\{\Delta v_l\}_{l\leq k}) = k \cdot log(\rho) - \rho \sum_{l=1}^{k} \Delta v_l \tag{6.12}$$

By maximizing $\mathcal{L}$ with respect to ρ we find

$$\rho = \frac{k}{\sum_{l=1}^{k} \Delta v_l} \tag{6.13}$$

and noting that $\sum_{l=1}^{k} \Delta v_l$ is the volume of the hypersphere with center at i containing k data points, we have obtained the standard k-nearest neighbor estimator ρ_i^{k-NN}, as

previously anticipated. The estimated error of the density is given by

$$\epsilon_{\rho_i^{k-NN}} = \frac{\rho_i^{k-NN}}{\sqrt{k}} \tag{6.14}$$

and gets smaller as k increases. Therefore, the optimal choice of k would be the largest possible value within the neighborhood, $\hat{k}$.
In order to provide an appropriate definition of neighborhood, and hence to determine $\hat{k}$, d'Errico et al. noted that the density in the neighborhood of a point i is expected to be rather constant, while if we consider further points that may not be part of the neighborhood, we would probably observe inhomogeneities. Therefore, the neighborhood is found as the set of points for which the density is approximately constant. The size of the neighborhood, that is the number of points that compose it, is found testing the hypothesis that the density is constant in the region occupied by the k nearest neighbors. This is achieved comparing the k-NN density estimation obtained from the nearest $k/4$ points and the furthest $k/4$ for increasing values of k. When these two values become inconsistent, according to a statistics built on a sample of points drawn from a uniform distribution, the algorithm stops and does not enlarge the neighborhood anymore. The algorithm hence performs the following steps:

1. estimate the density ρ_i^{k-NN} according to Eq. 6.13 for increasing values of nearest neighbors k;

2. compute the rescaled difference between the log-likelihood of the nearest $k/4$ neighbors and the log-likelihood of the furthest $k/4$ neighbors:

$$\Delta\mathcal{L}_k = \frac{1}{k/4}\left(\mathcal{L}_k^{near} - \mathcal{L}_k^{far}\right) \tag{6.15}$$

3. find the largest k for which $\Delta\mathcal{L}_k$ is significantly small and call it $\hat{k}$. Even if the density is constant ΔL_k can take large values due to statistical fluctuations. Therefore k is chosen so that to satisfy

$$|\Delta\mathcal{L}_k| \leq C_k \ \forall k \leq \hat{k} \text{ or } |\Delta\mathcal{L}_{\hat{k}+1}| > C_{\hat{k}+1}, \tag{6.16}$$

   where C_k is defined in such a way that the probability of satisfying this condition would be equal to $p = 0.001$ for a sample of points harvested from a uniform probability distribution.

4. The number of neighbors used for estimating the density of data point i is fixed to $\hat{k}$.

In this procedure, at the exit value $\hat{k}$ the log-likelihood of the nearest neighbors is by construction significantly different from the log-likelihood of the furthest ones, indicating that the density close to the $\hat{k}$-th neighbor is already substantially different from the density close to data point i. The authors corrected this trend, computing the density by a Jackknife procedure aimed at capturing its variation as a function of the distance from each point. The $\hat{k}$ neighbours where separated into M equal subsets. Then, for each subset the average volume was computed as

$$v^{(j)} = \frac{M}{\hat{k}}\left(v_{j\hat{k}/M} - v_{(j-1)\hat{k}/M}\right). \tag{6.17}$$

They assumed the variation of the density to be linear in j, $v^{(j)} \sim v^0 + \beta j$, where v^0 and β are estimated by minimizing $\sum_{j=1}^{M}\left(v^{(j)} - v^0 - \beta j\right)$. Jackknife procedure [90] was used

to estimate the density and the error induced by the fit: each $v^{(j)}$ was, in turn, dropped from the sample and the parameter v^0 was estimated from the reduced sample, providing a set of M estimates $v^0_\alpha, \alpha = 1, ..., M$. The estimator of the density is then

$$\rho_i = \frac{1}{Mv^0 - \frac{M-1}{M}\sum_{\alpha=1}^{M} v^0_\alpha}. \tag{6.18}$$

The error on this estimate is $\epsilon_{\rho_i} = \rho_i^2 \sqrt{\frac{M-1}{M}\sum_{\alpha=1}^{M}(v^0_\alpha - v^0)^2}$. This is considered reliable only if higher than the k-NN estimates, and if not, is substituted by this last one. The procedure was performed for all the values of M between 4 and $\hat{k}/2$, and the value of M that maximizes ϵ_{ρ_i} was chosen.
Summarizing, the method by d'Errico et al. is capable of providing accurate estimates of the density at each point and also of its associated uncertainty and this allows us to recognize genuine features of the probability density distribution, for instance a density peak, and to distinguish them from statistical fluctuations due to finite sampling.

6.4.2 Detection of unreliable points

For a small but significant fraction of points the value of $\hat{k}$ provided by the condition in Eq. 6.15 may be smaller than expected due to the fact that condition 6.16 is violated because of statistical fluctuations rather than to a drift of the density. Those points are by construction affected by a much larger error than their neighbors, leading and are thus classified as unreliable. In order to detect such spikes, d'Errico et al. developed an heuristic criterion that tests the assumption that the density at a point is uniform within the neighborhood of size defined by its $\hat{k}$. They compared the estimated density at a point with the average estimated densities at the $\hat{k}$ nearest neighbors. Denoting by ρ_i the density at data point i and by NN_i the set of its $\hat{k}$ nearest neighbors, they computed $\mu_i = E[\rho_j | j \in NN_i]$ and $\sigma_i^2 = Var[\rho_j | j \in NN_i]$. Then, they classified as unreliable those data points for which $(\rho_i - \mu_i) > \sqrt{\sigma_i^2 + \epsilon^2_{\rho_i}}$. Finally, the values of μ_i and σ_i^2 were recomputed, restricting the averages only to the points that were classified as reliable. This procedure was iterated until the set of unreliable points remains unchanged in two successive iterations.

6.4.3 Reconstruction of density topographies

The estimated density and associated error can be used to find peaks and saddle points within the dataset and this information can be exploited to partition the dataset into separate clusters.
In order to find the density peaks, points are ranked not according to their density, which can be affected by non-uniform errors, but to a function of the probability of their difference in density. Defining the pull, Δ, as the set of the $\Delta_i = \frac{(\rho_i^{true} - \rho_i^{estimate})}{\epsilon_{\rho_i}}$, d'Errico et al. numerically proved that its probability distribution is accurately described by a Gaussian with zero mean and unit variance. As a consequence, the probability that the estimated density at point i, $\rho_i^{estimate}$, corresponds to the true density, ρ_i^{true}, can be estimated by a Gaussian centered on $\rho_i^{estimate}$ and with a variance equal to the estimated error ϵ_{ρ_i}. Then, every point i was ranked estimating the quantity g_i

$$g_i = \prod_{l \neq i} \int_{-\infty}^{\infty} dx \mathcal{N}_x(\rho_l, \epsilon^2_{\rho_l}) \int_{x}^{\infty} dy \mathcal{N}_y(\rho_i, \epsilon^2_{\rho_i}) \tag{6.19}$$

where $\mathcal{N}_x(a, b) = \frac{1}{\sqrt{2\pi b}} e^{-\frac{(x-a)^2}{2b}}$. The quantity g_i is the product over l of the probabilities that point i has a higher density than point l. We here employ g_i as an effective density

instead of ρ_i, and define as cluster centers the local maxima of g_i. Therefore, the centers are the points that, with maximum probability, are surrounded by points with a lower density. Note that the local maxima of g_i coincide with the local maxima of ρ_i when the error is uniform. The double integral in Eq. 6.19 has not a simple analytical form but for computational purposes it can be parametrized as $\left[1 + exp\left(-2\frac{(\rho_i-\rho_l)}{\sqrt{\epsilon^2_{\rho_i}+\epsilon^2_{\rho_l}}}\right)\right]^{-1}$ as numerically verified by authors. Following the idea proposed by Rodriguez and Laio [78], putative centers are considered as those points i for which the distance to the nearest point with higher value of g, $\delta_i = \min_{j:g_j>g_i} r_{ij}$, is greater than the distance with the furthest nearest neighbor $\hat{k}_i$: $\delta_i > r_{i\hat{k}_i}$. Thus, a data point is a center only if all its $\hat{k}$-th nearest neighbors, which contribute to determine the value of its density, have a value of g lower than g_i. In other words, cluster centers are defined as the local maxima of g, i.e. the points that, with maximum probability, are surrounded by points with a lower density.
Afterwards, all points that were not classified as centers, are assigned in order of decreasing g, to the same cluster of the nearest point with higher g.
Finally, clusters that are not statistically separated are merged together. For this purpose, border points between two clusters c and c' are identified according to the following statement: a point i belonging to c is assumed to be at the border between c and c' if its closest point j belonging to c' is within a distance greater than the radius of c and if i is the closest point to j among those belonging to c.
The border density $\rho_{cc'}$ is defined as the density at the point with the highest value of g, among those at the border between c and c', and the border density error $\epsilon_{cc'}$ is equal to the density error of this point. If all the points in a cluster have density values compatible with the border density, taking errors into account, then the cluster can be considered as the result of a statistical fluctuation and merged with another cluster. In particular, cluster c is merged with cluster c' if $\max_{i\in c}(\rho_i - \epsilon_{\rho_i}) < A(\rho_{cc'} + \epsilon_{\rho cc'})$, where A is a parameter that is fixed to 1 and tunes the confidence level of the topography reconstruction, giving more detailed but less reliable topographies for larger values. This condition is checked for all the clusters c and c', in order of decreasing $\rho_{cc'}$. The merging step allows pruning the set of clusters from those corresponding to density maxima that are not statistically robust, thus recovering the topography of the underlying density function.

6.4.4 Recognizing points with ambiguous cluster assignation

For ensuring proper understanding of the data and a high correspondence between the density topography and the taxonomic composition, we implemented a procedure that removes points with ambiguous clustering assignment. In particular, we aimed at removing those points belonging to the background or to a low-populated taxa that are wrongly assigned to the same cluster of a more abundant one.
As before, let δ_i be the distance between i and its nearest point with higher density rank. For each cluster, we considered the distribution of the δs of its points, excluding the center and those points whose density reconstruction has been marked as unreliable by the LOCk-NN algorithm (see Sec. 6.4.2). If spurious local centers are present, those are by definition at a relatively large distance from any point with a higher density rank. Consequently, their corresponding δ is much larger than the typical δs within the cluster and they appear as outliers in the distribution of δ. Then, using the Tukey's test [91], we identified as major outliers those points that lie above the threshold $\delta_{th} = Q3+3(Q3-Q1)$, where $Q1$ and $Q3$ are the lower and upper quartiles respectively. Moreover, for background identification, we defined a density threshold ρ_{th} equal to the lower not-null border density. Finally, we removed points for which at least one of the following relation holds:

- $\delta > \delta_{th}$;

- the closest point with higher rank has $\delta > \delta_{th}$;
- the density value is compatible with ρ_{th} within their errors (only if the density measure is reliable).

6.4.5 Re-clustering halo points

The set of points that were removed in the previous step is called halo and should be re-clustered so that to reconstruct the missing OTUs. Here we describe a preliminary proposal to performed such re-clustering.
After using mothur to compute the distance matrix for the halo sequences only (see Sec. 6.1), we sorted the reads according to their density and distance from the closer point with higher density, following the idea by Rodriguez and Laio [78]. Sequences for which these two values are bigger will more probably correspond to cluster centers and will be processed first. Here, density was computed by counting for each sequence the number of reads within a distance of 0.05.
Starting from the top of the sorted list, we considered, for each sequence x_i, the distribution of its distances with all the other reads that are still in the list. Using the Kernel Density Estimation implemented in the python seaborn package [92], we detected the first local minimum of the distances distribution. If the corresponding distance $\hat{\delta}$ was less than 0.2, we considered x_i and all the sequences within $\hat{\delta}$ as belonging to the same cluster. Otherwise, the sequence x_i was classified as singleton. Here, the 0.2 cutoff was introduced to deal with the situation in which the read is a singleton and the first local minimum is the one that includes the closer cluster to which the read does not belong. If the local minimum is at a distance greater than 0.2, in fact, we can confirm that the detected cluster is not biologically significant because of the highly conservation of the 16S rRNA sequence. Finally, all sequences detected in the cluster were removed from the list and the next sequence, if any, was processed.

6.5 Simulated Data

In order to assess the performances of the new parameter-free approach detailed in the previous section, we simulated a datasets of 16S rRNA sequences, in which we know by construction which sequence belongs to which cluster. In particular, we considered a set of representative sequences derived from the Greengenes database that are known to have a minimum pairwise distance of about 3%.
Data are available at $http://greengenes.lbl.gov$ under the filename $gg\_97\_otu\_6oct2010.fasta$. This database had been generated by clustering 16S rRNA sequences into OTUs with UCLUST at 97% similarity threshold, and keeping only the representative sequences, i.e. clusters centroids. Note that, as priviously detailed, UCLUST is an heuristic algorithm and it may happen that the similarity between two cluster centroids is actually bigger than the threshold. However, as ensured by UCLUST author and also shown in Fig. 6.1, these are rare events in practice and we thus did not take such consideration into account.
To simulate the dataset, we randomly selected 1000 sequences from the Greengenes representative sequences, that we will call reference sequences. The number 1000 was chosen so that to obtain a similar number of OTUs as in the ELDERMET dataset used in Part I, when we used a similarity threshold of 97%. Among these 1000 sequences, 4 will not be part of the simulated centroids so that the true number of clusters will be 996. For each reference sequence, we identified the V4 region by aligning it to the universal primer 'AYTGGGYDTAAAGNG' and trimming it so that to start at the V4 primer.
Fig. 6.1 shows the distribution of the pairwise distances among the sampled 1000 sequences. As anticipated, only a very small number of them turns out to have a pairwise distance

smaller than 0.03 (or 3%) and results located on the left of the red line.

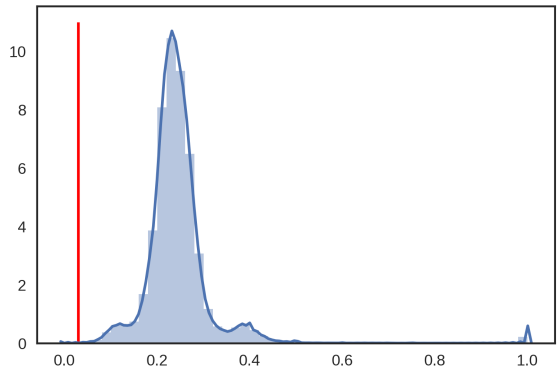


Figure 6.1: Distribution of the pairwise distances between the 1000 reference sequences selected as cluster centroids.

We used the grinder algorithm [93] to simulate 30000 sequences starting from the 1000 references. The idea of the simulation is to generate, for each reference, a cluster of sequences that differ because of some random mutation that may be attributed to biological or experimental noise. Moreover, clusters are produced in such a way that their abundance will satisfy a predefined RSA distribution (see Sec. 0.5.2 for details on the RSA). In particular, we simulated a RSA given by a mixture of two Negative Binomials, that is the distribution that we have hypothesized in Part I. The resulting RSA is shown in Fig. 6.2.

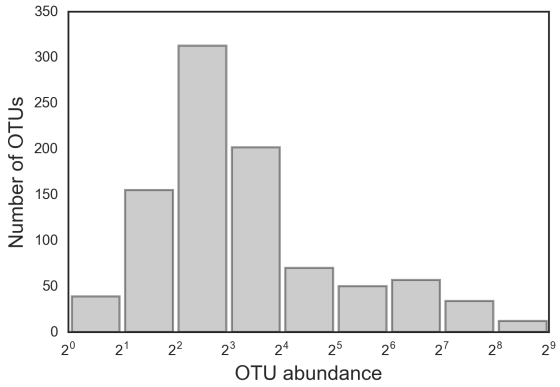


Figure 6.2: Preston plot of the simulated OTUs abundances. The RSA distribution is a mixture of two Negative Binomials.

The simulation was performed imposing that read lengths follow a Normal distribution with mean 250 bp and standard deviation 50 bp, so that to mimic the 454 GS FLX Titanium sequencing platform, that is the one used in the ELDERMET dataset (see Part I). Sequencing errors were introduced both in the form of mutations and homopolymeric stretches. The distribution for the homopolymer length n was chosen to be $\sim \mathcal{N}(n, 0.03494 + n \cdot 0.06856)$, following Balzer's empirical model [94], that was computed for the Titanium technology. For what concerns the mutation distribution, we relied on Gilles' results [95] and used a linear model that sets a 0.53% error rate at the 5' end of reads and 1.07% at the 3' end. The percentage of insertion and deletion among the mutations was set at 54% and 36% respectively. Finally, a probability of 10% was set for reads in the amplicon library to be chimeras, and the default distribution was used for higher order chimeras. Chimeras are reads composed by two or more biological sequences joined together that may be formed during the sequencing procedure. Several methods exist to remove chimera before starting the clustering algorithm, such as for in-

stance UCHIME [96], that we used in Part I as a component of UPARSE. Here we avoid considering this further issue and remove chimeras by checking the number of references from which each sequence has been simulated.
Before proceeding with the clustering, the simulated sequences were preprocessed with the same conditions required for real data in Part I. In particular, we used mothur to trim sequences so that to exclude reads shorter than 150 bp, to cut sequences longer than 350 bp and to cull those with ambiguous bases or with homopolymers longer than 8 bp. The preprocessing step returned 27021 filtered sequences, that reduce to 24289 sequences belonging to 932 OTUs after chimera removal.
Finally, we clustered the simulated sequences into OTUs using the four above mentioned methods: mothur, UCLUST, CROP and LOCk-NN.

6.6 Clustering comparison and evaluation

We compared the results obtained of the different clustering with the ground truth, or true clustering, that is obtained by recalling for each sequence the reference from which it was generated, that will indicate the cluster to which it belongs.
The comparison was performed based on three measures. First, we simply counted the number of clusters detected by each method. Then, we visually compared the resulting RSA distributions obtained by counting the number of clusters with a certain number of elements (see Sec. 0.5.2). This was performed by plotting the RSA cumulative (cum), or more precisely the logarithm of $(1 - cum)$. Finally, we compared each clustering outcome with the ground truth computing the Normalized Mutual Information score (NMI) [97]. We chose to use NMI rather than, for instance, the cluster purity so that to take into account the number of clusters, besides their composition. The NMI index, in fact, penalizes two types of errors: the wrong assignment of sequences with the same species label into different clusters, and the assignment of sequences with different species labels into the same clusters. Purity, instead, considers only the second situation and would wrongly suggest an optimal result even when a big number of very small clusters is detected, as may be the case, for instance, of methods that use a high similarity threshold.
Calling C_1 and C_2 the two clustering classifications, the NMI is defined as

$$NMI(C_1, C_2) = \frac{MI(C_1, C_2)}{\sqrt{H(C_1)H(C_2)}}, \tag{6.20}$$

where $H(C_k)$ is the entropy for the clustering k, $H(C_k) = -\sum_i P_k(i)log(P_k(i))$, with $P_k(i)$ being the probability that an object is picked at random from C_k, and

$$MI(C_1, C_2) = \sum_{i \in C_1} \sum_{j \in C_2} P(i,j) log \left(\frac{P(i,j)}{P_1(i)P_2(j)} \right) \tag{6.21}$$

is the Mutual Information (MI), that depends on $P(i,j)$, the probability that an object picked at random falls in both class i in C_1 and class j in C_2. The NMI score ranges from 0 to 1, with 1 indicating perfect correspondence between the two clustering.

Chapter 7

Results

The simulated dataset contains 24289 non chimera sequences and is a complex sample of 932 hierarchically related species (OTUs) with a population ranging from a few units to thousands of elements. The intrinsic dimension of the dataset estimated by DANCo [89] was 7 and The LOC-kNN clustering returned a total of 132 clusters, from which 2805 sequences were excluded because classified as halo. The average abundance of the reconstructed OTUs was 162 ($\pm$ 135), while the true average size of the simulated OTUs was 28 ($\pm$ 61).
We may already notice that LOC-kNN detects a smaller number of clusters with a general bigger size than expected. In particular, as shown in the following, LOC-kNN returns good estimates for clusters that are not too small, for instance with a population bigger than 50. The reason of such behavior relies mainly in two issues. The first one is that, when a cluster is composed by a small number of elements, it is difficult to give a reliable estimate of its density. The second one, instead, is related to the hierarchical relation between bacteria. Because of this, in fact, small clusters will probably not be isolated but, instead, they may be close to some bigger cluster. In this case, the algorithm will probably merge the two clusters because it will not be able to detect a significant change in the density distribution.
The halo detection step has been added to LOC-kNN as an attempt to detect sequences that may belong to small clusters and were wrongly assigned. As we will show, this enables to improve the LOC-kNN results, but further refinements should be performed in the future to achieve a more truthful clustering.
Among the 2805 halo sequences, 1926 actually belong to clusters with abundance less than 50. However, other 4844 sequences are still wrongly merged to bigger clusters and are part of the 21484 non-halo sequences.
In order to assess and compare the performances of the clustering methods that we considered, we focused on three measures. First, we checked if the number of detected clusters was similar to the truth. Then, we computed the Normalized Mutual Information score comparing each clustering outcome with the ground truth. And finally, we compared the RSA distributions in the form of $log(1 - cum)$, where cum indicates its cumulative. Since we noticed that LOC-kNN has bad performances for very small clusters but may give good results for bigger populations, we performed the analysis considering different thresholds and excluding the sequences that, according to the true classification, belonged to OTUs with abundance less than the threshold. In this way, we were able to evaluate the performances of LOC-kNN for highly populated clusters independently of the presence of the smaller ones.
Moreover, we performed the comparison first excluding the halo sequences from all results and then considering all the sequences and adding the OTUs of the re-clustering step to LOC-kNN.

7.1 Excluding halo sequences

Fig. 7.1 shows the number of clusters obtained when considering only sequences that belong to clusters bigger than increasing thresholds (x-axis) and also excluding those of the halo. Results are shown for all the evaluated methods and the black line indicates the ground truth. As anticipated before, LOC-kNN shows good performances if we exclude small clusters (abundance $\lesssim 50$). These are in fact not detected, as can be also seen from the Preston plots in Fig. 7.3 that compares the obtained RSA distribution with the ground truth in the form of Preston plot. In the other methods, it is clear that the choice of the similarity threshold has great impacts on the performances. Both mothur and UCLUST approximately detect the correct number of OTUs if we do not introduce a threshold in the minimum abundance and if we use a similarity threshold of 93%. However, excluding the rarest sequences, the number of clusters turns out to be always overestimated, even at 93% threshold, and changing the similarity threshold produces worse results. CROP is the method that gives better results, and this is probably due to its soft threshold. In this case, using 95% or 97% of similarity threshold gives similar results that are both in agreement with the ground truth.

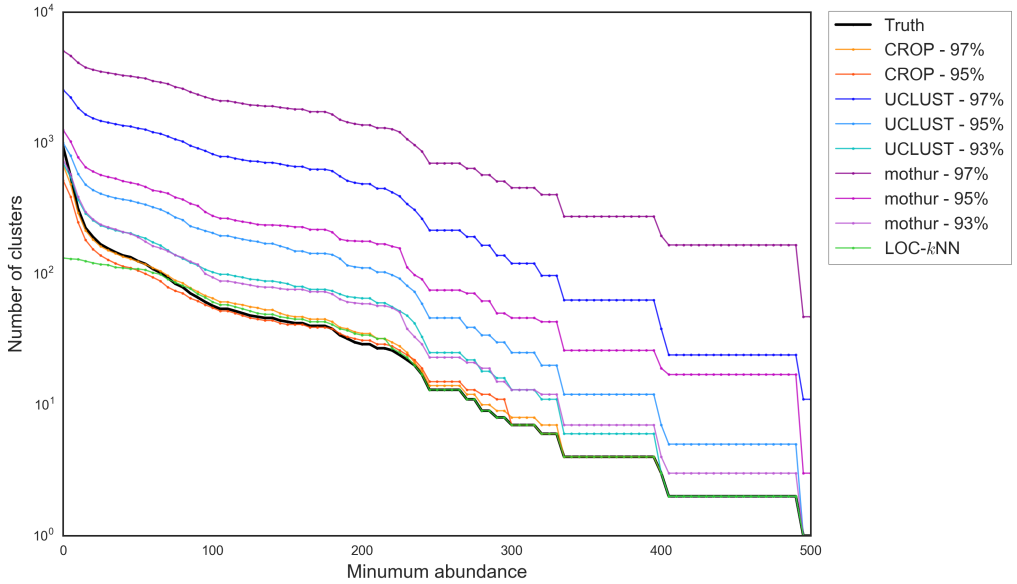


Figure 7.1: Number of clusters (logarithmic scale) when varying the minimum cluster size. Halo sequences were excluded.

The NMI scores are plotted in Fig. 7.2. Again, the x-axis indicates the minimum abundance of the ground truth OTUs included in the comparison. Accordingly to the previous result, LOC-kNN shows good agreement with the ground truth for clusters with more than ~ 50 elements, while mothur and UCLUST outcomes highly depend on the cutoff and achieve the best performances at 93%. The dependence on the similarity threshold is now highlighted also for CROP, whose best NMI scores are achieved at 97% but decrease at 95%.
Finally, Fig. 7.3 compares the Preston plot obtained by LOC-kNN with the ground truth and Fig. 7.4 compares the RSA distributions obtained with all the four methods. Also in the $log(1-cum)$ plots, the black line indicates the ground truth. The reported distributions were all computed considering only sequences that belong to clusters with a minimum size of 50 elements according to the ground truth classification. Halo sequences were also excluded. The RSA distribution of abundant clusters obtained with LOC-kNN shows good agreement with the ground truth and its results are comparable with those obtained

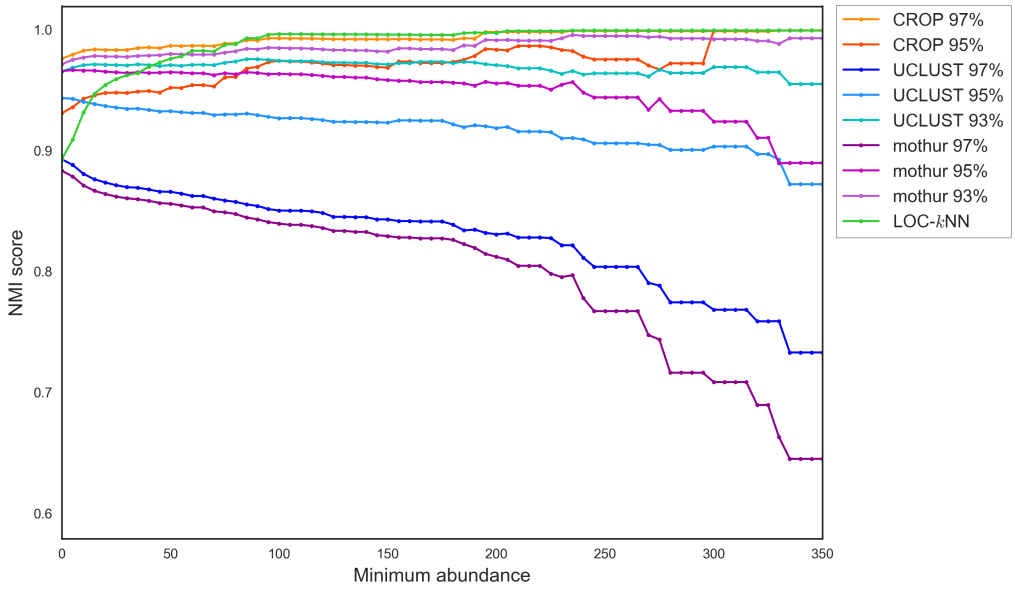


Figure 7.2: NMI index obtained while restricting to the most abundant species, starting from the full sample (population greater than zero). Halo sequences were excluded.

with CROP. UCLUST shows the worst performances and mothur roughly reconstructs the true distribution only with 93% similarity threshold.

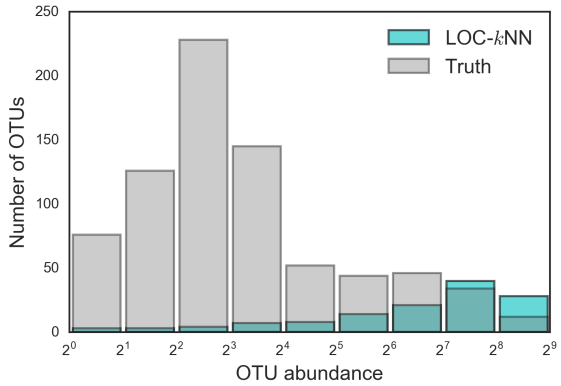


Figure 7.3: Preston plot representing the ground truth RSA (gray) and the abundances obtained with LOC-kNN (light blue). Halo sequences were excluded.

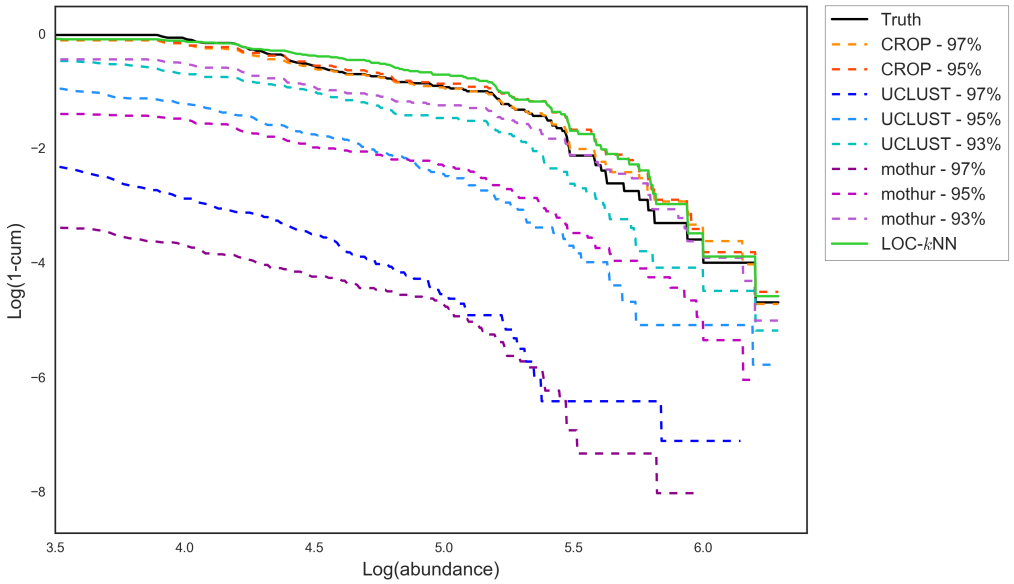


Figure 7.4: RSA distributions obtained for different clustering methods including only species with at least 50 individuals and excluding halo sequences.

7.2 Re-clustering halo sequences

Re-clustering the halo sequences with the proposed procedure (LOC-kNN + KDE) produces a RSA distribution that is comparable with the one that we would obtain considering the true classification, as shown by the Preston plots in Fig. 7.5. This suggests that, if we collected all and only the small clusters sequences in the halo set, we would be able to reconstruct their OUTs.

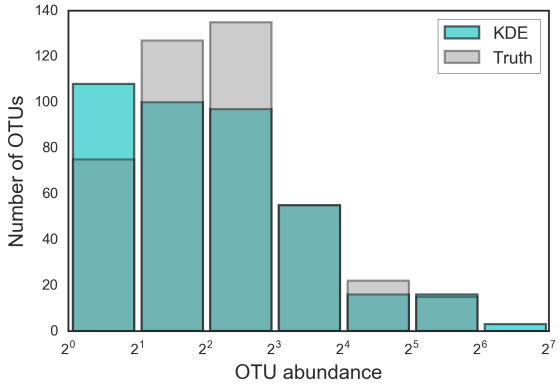


Figure 7.5: Preston plot of the halo sequences for the ground truth classification (gray) and the KDE clustering (light blue).

The NMI scores obtained with the LOC-kNN + KDE method are still comparable with the other methods if we only consider OTUs with at least ~ 50 elements. Since the re-clustering shows good performances, this implies that the problem has to be found in the halo detection step, as suggested before. Note, moreover, that the decrease of the NMI score for high values of the abundance minimum is due to the fact that some sequences of the most abundant clusters have been wrongly included in the halo set, and this is also why we did not observe this trend in Fig. 7.2, where halo sequences were excluded in advance.
Finally, the estimate of the total RSA distribution, without excluding any sequence, is improved by the re-clustering step, in the sense that we are able now to roughly reconstruct

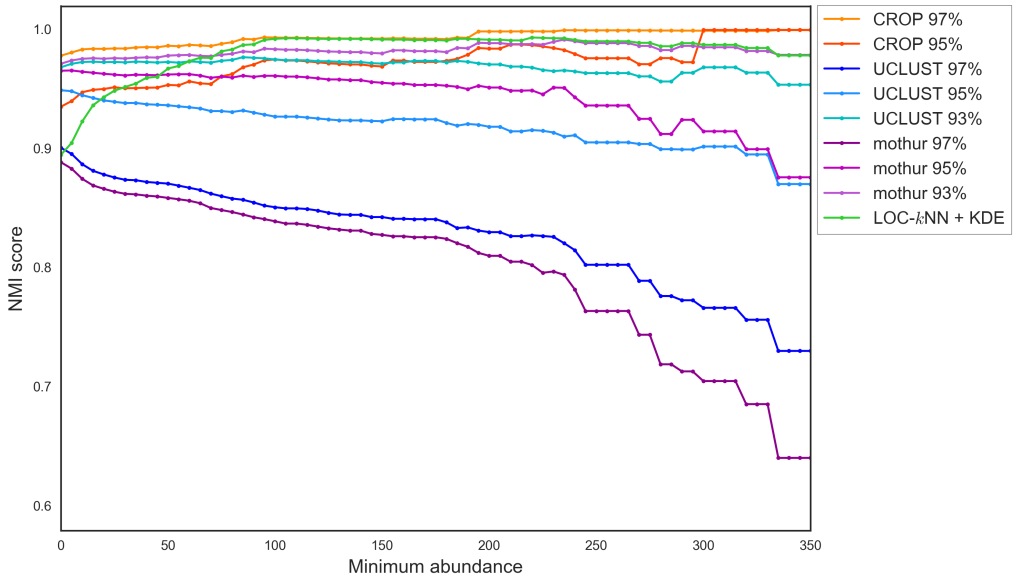


Figure 7.6: NMI index obtained while restricting to the most abundant species, starting from the full sample (population greater than zero).

also the left-hand side of the curve, that is however still underestimated, as shown in the Preston plots of Fig. 7.7 and in the $log(1 - cum)$ plots in Fig.7.8 (a), in which no threshold on the minimum OTUs abundance was applied. Moreover, excluding the OTUs that in ground truth have less than 16 elements, i.e. using a lower threshold then before, we already obtain a RSA that is comparable with the best results obtained with other methods, specifically with CROP and especially for what concerns the distribution tail (see Fig.7.8 (b)).

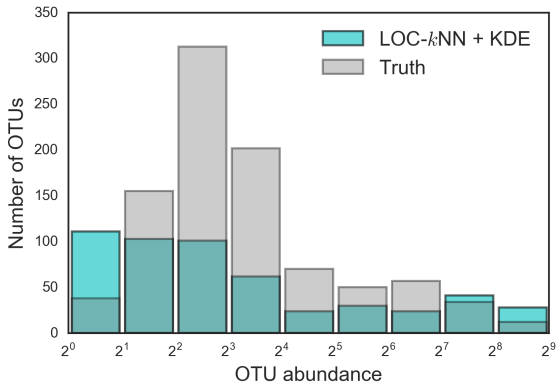
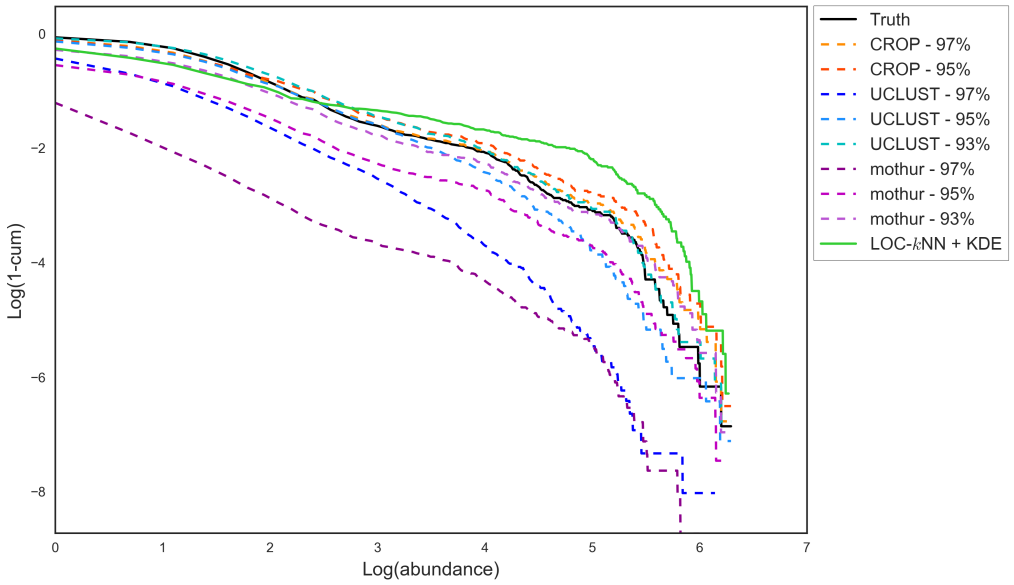
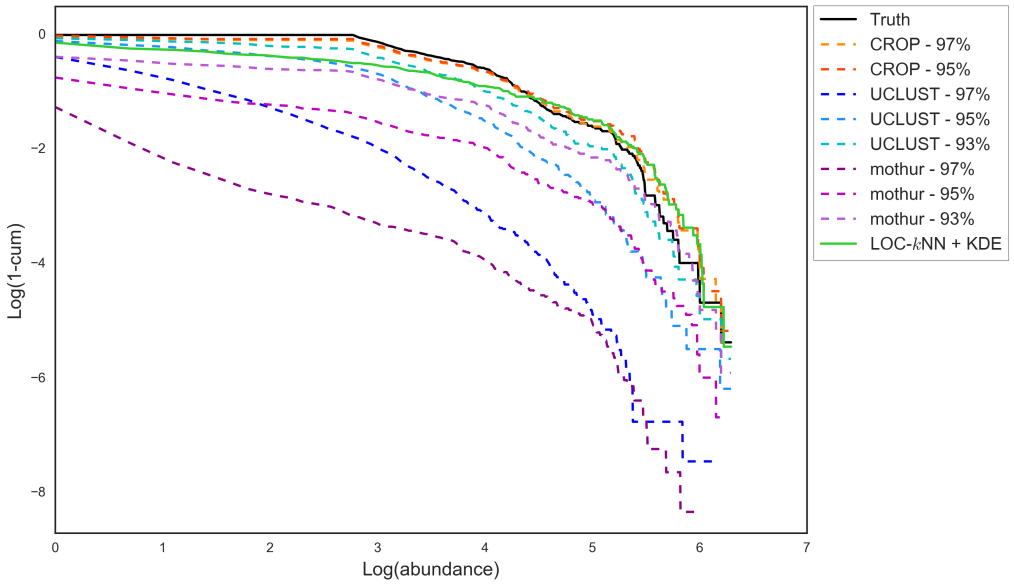


Figure 7.7: Preston plot representing the ground truth RSA (gray) and the abundances obtained with LOC-kNN + KDE (light blue).



(a) RSA distributions computed with different clustering methods without excluding any sequence.



(b) RSA distributions computed with different clustering methods considering only sequences that belong to species with abundance > 16.

Figure 7.8

Chapter 8

Discussion

In this work we proposed to use the recently developed LOC-kNN method to cluster 16S rRNA sequences into OTUs. This is a crucial step in the analysis of bacterial ecosystems such as the Gut Microbiota and its main limitation is usually the dependence on a predefined similarity cutoff that fixes the clusters sizes. The use of a threshold may be appealing according to the idea of studying the system at different phylogenetic levels, as we also did in Part I. However, a parameter-free clustering may reconstruct the data structure in a more reliable way and may thus provide a better tool to re-define the concept of species and learn, for instance, the ecosystem biodiversity. In LOC-kNN, the clustering is performed by estimating the data topography and density distribution. The algorithm is able to correctly detect clusters that are not too small (i.e. with more than 50 elements). The NMI score for LOC-kNN, when excluding OTUs with less than 50 elements, is in fact 0.98, indicating good agreement with the ground truth. The LOC-kNN performances in the reconstruction of abundant clusters are actually better than those achieved with mothur and UCLUST, considering the whole biologically meaningful range of similarity thresholds (93% to 97%), and are comparable with CROP at 97%, with the advantage of not relying on a predetermined cutoff. CROP is the clustering method, among the considered ones, that gives better results and, interestingly, it is based on the use a soft threshold. When this threshold is set to 97%, it is able to reconstruct the true clusters, showing very good concordance with the ground truth, for what concerns both the cluster composition and the RSA distribution. However, results worsen if we consider a different threshold (95%).

In order to reconstruct smaller clusters, that are not detected by LOC-kNN, we proposed a procedure that reveals sequences with ambiguous assignment (halo) and re-clusters them on the basis of the distances distribution. The re-clustering method shows satisfying performances, even if the halo detection step still misses a fraction of wrongly classified sequences and thus needs some further improvement.

Finally, we remark that, even if the use of a fixed similarity threshold to cluster 16S rRNA sequences into OTUs may be appealing, under the idea of reproducing different phylogenetic levels, we think it would be more interesting to detect the true structure of the dataset, avoiding, for instance, the further and incorrect partitions that would be obtained with high similarity thresholds. When dealing with real data, in fact, it would be difficult to know in advance which threshold to use and, probably, the best choice would not even be the same for all species. For these reasons, we conclude that the modified LOC-kNN method is a promising tool, that, with some further improvements, may yield an important contribution to the study of bacterial ecosystems biodiversity.

Part III
Relative Species Abundance of protein domains

Chapter 9

Introduction

9.1 Protein domains and evolution

Proteins are long unbranched polymer chains called polypeptides, created by joining amino acids into particular sequences. Through billions of years of evolution, these sequences have been selected to give proteins useful functions, such as catalyzing reactions, maintaining structures, generating movement, sensing signals, and so on. Studies on the conformation, function and evolution of proteins suggest to consider the protein as a set of independent segments that constitute its organization units, so as its evolutionary components. These are called *protein domains* and are defined as substructures produced by any contiguous part of a polypeptide chain that can fold independently of the rest of the protein into a compact and stable structure [98]. Different domains of a protein are often associated with different functions. They usually contain between 40 and 350 amino acids, and they are the modular unit from which many larger proteins are constructed. The smallest protein molecules contain only a single domain, while larger proteins can contain several dozen of domains, often connected to each other by short, relatively unstructured lengths of polypeptide chain that can act as flexible hinges between domains. The particular amino acid sequences that form protein domains and hence proteins, are determined by the genetic code written in the cell DNA genes and are the result of billions of years of evolution. Thanks to accidents and mistakes in the normal mechanisms by which genomes are copied or repaired when damaged and as a consequence of natural selection and non-random survival, genomes undergo evolutionary processes and this allows the evolution of organisms complexity through the increase in protein repertoires. The mechanisms by which protein evolution occurs are mainly [98][99]: (i) duplication of sequences that code for one or more domains; (ii) divergence of the duplicated sequences by mutations, deletions and insertions to produce modified structures that may have useful new properties and be selected; (iii) recombination of genes that results in novel arrangements of domains (domain shuffling); (iv) relocation of transposable DNA elements, that are parasitic DNA sequences that spread within the genome, often disrupting or altering gene functions and occasionally creating novel genes; and (v) horizontal gene transfer, by which genes are transfered from one species of cell to another, especially in prokaryotes. All these processes can be seen as occurring independently in each protein domain composing the evolving protein, and for this reason protein domains are considered the evolutionary units of proteins. Therefore, many aspects of the evolutionary process that generated the current protein ensemble for each organism, can be investigated by studying the evolution of protein domains. Here, we describe protein domains as elements (or species) of a population and we aim to describe the ecological dynamics that lead their evolution. As for the Gut Microbiota (see 1), we concentrate on the RSA, that is the Relative Species Abundance distribution. This describes how many species have a certain number of in-

dividuals and its form is the stationary solution of the stochastic process that rules the system evolution. We will consider the protein domain population of several Bacteria. Taking into account thousands of bacterial species, we will also be able to compare the differences in the protein domain distributions and to compare them with the classical taxonomic classification.
In the following, we will describe an appropriate stochastic model for the protein domain ecosystem. Such model predicts that protein domains are generated through an inhomogeneous Poisson process with Lognormal success rate. We will see how, for the protein domain ecosystem, it is essential to introduce environmental noise, that we instead neglected when describing the Gut Microbiota, and to consider Gompertzian model for the death rate, according to which that the death probability increases with time. This is particularly realistic for protein domains, since death is caused by mutations that normally accumulates over time, so that the longer a domain is present in the genome the higher is the probability that mutations have transformed it into a new domain or in a sequence that is not a domain anymore. We will explain how, starting from bacterial genome sequences we inferred the protein domain distribution and, finally, we will present the results and some considerations.

9.2 Birth-Death-Innovation models and drawbacks

Studies from the early 2000s, suggested that the distributions of several genome-related quantities follow power law trends. The frequency distributions of proteins or domains in different proteomes appeared in fact to fit the power law: $P(i) \approx ci^{-\gamma}$ where $P(i)$ is the frequency of domain species including exactly i members, c is a normalization constant and γ is a parameter, which typically assumes values between 1 and 3 [100][101][102]. Power distributions arise in an extraordinary variety of contests, from words frequencies in linguistic, to the distribution of links between documents in the Internet, to the number of species that become extinct within a year. Power laws have been especially studied in network theory [103], to describe the distribution of the number of links, or degree, of nodes. In our contest, nodes would correspond to particular kind (species) of protein domains, while the node degree would indicate the species abundance, that is the number of protein domain elements of that particular species. The principal pattern of network evolution that ensures the emergence of power distributions is preferential attachment, whereby the probability of a node (protein domain species) to acquire a new connection (increasing its abundance) increases with the number of connections (abundance) this node already has. A network whose degree distribution is a power law is called scale-free network to underline the fact that the shape of such distribution remains the same regardless of the scaling of the analyzed variable. This property reminds us of demographic noise, that produces a constant increase in the species abundance independently of its size.
Another recent study proposed a more general distribution, claiming that many biological systems are better described by the so-called generalized Pareto function, $P(i) = c(i + a)^{-\gamma}$, where a is an additional parameter [104][105]. Obviously at large i ($i \gg a$), a generalized Pareto distribution is indistinguishable from a power law, but at small i, it deviates significantly, with the magnitude of the deviation depending on a. The importance of the analysis of frequency distributions of domains or proteins lies in the fact that distinct forms of such distributions can be linked to specific models of genome evolution. For this reason, the authors proposed a class of simple evolutionary models, which are called birth death innovation models (BDIM) and include: (i) domain birth, due to duplication; (ii) domain death, as a result of inactivation and loss; (iii) and innovation or emergence of a new species, that may occur, for example, because of extensive modifications of a member of an existing species, horizontal gene transfer or even due to the origin of a new protein

from non-coding a sequence.
This kind of modeling completely disregards the protein domain identity, but gives rise to equilibrium distributions of protein domain abundances that can be compared with empirical data [106]. The authors suggested to consider linear birth and death rates, respectively given by $\lambda_i = \lambda(i + a)$ and $\delta_i = \delta(i + b)$. This model is equivalent to the one proposed by Volkov [24] (see Eq. 78 in Sec. 0.7.5) and the stationary distribution is indeed a Negative Binomial, that tends to the trunkated Log-Series distribution when innovation is small, as already pointed out when discussing Volkov's Negative Binomial RSA (see Eq. 85 in Sec. 0.7.5). Karev et al. showed that the simplest model that results in a good fit to the observed domain abundance distributions for the several prokaryotic and eukaryotic genomes that they analyzed, is the so called second-order balanced linear BDIM, where $\lambda = \delta$ [107]. In this case, the asymptotic distribution turns out to be a power-law. However, in order to obtain more consistent estimation of the genome evolution time, the authors had to introduce a quadratic term, setting $\lambda_i = \lambda(i + a)(i + 1)$ and $\delta_i = \lambda(i + b)i$ and even with this correction, the estimated mean formation time for the protein domain species was $\sim 10^{11}$ years, while it is estimated that the common ancestor to all cellular life may have arisen about $3.8 \cdot 10^9$ years ago [108].

9.3 The Log-Normal hypothesis

Here, we propose an alternative model for the protein domain RSA, that better fits the data. The main peculiarities of this model are the presence of environmental noise, besides the demographic one, and the hypothesis of Gompertzian death. As described in Sec. 0.7.6, under these assumptions the RSA turns out to be a Poisson distribution with Log-Normal rate $\lambda(x)$, that is called Poisson Log-Normal. Note that the Poisson distribution rate $\lambda(x)$ describes the probability of sampling a protein domain species with x individuals in an infinitesimal time interval and for this reason we refer to it as the abundance model.
Log-Normal behavior has been previously hypothesized for the abundances of the chemical components in the cell and it was empirically observed for protein abundances in *Escherichia coli* [109]. Recalling that protein domains are protein subunits, it is clear how the Log-Normal hypothesis for protein domains abundances is consistent with this observation. Furusawa et al. [109] also proposed a chemical reaction model to explain the protein abundance log-normality. Intuitively, log-normality arises when the population growth, besides depending on the number of individuals that are present, is also influenced by some multiplicative noise, such as the environmental one. Biochemical reactions inside the cell consist of a huge number of catalytic reaction processes in which several molecular species participate. Clearly, the probability of interaction between the molecules involved in the reactions depends on the number of available molecules. At the same time, however, all chemical reaction processes are inevitably accompanied by fluctuations arising from the stochastic collisions of chemicals, that act as environmental noise.
The reason for which this kind of process gives rise to a Log-Normal distribution can be understood considering the corresponding population dynamic equation:

$$\frac{dx(t)}{dt} = bx(t) \tag{9.1}$$

where b is the growing rate. The system is subject to environmental noise, that is multiplicative. Thus, we can write $b = \bar{b} + \sigma_e^2$, where $\bar{b}$ is the temporal average of the birth rate, while $\sigma_e^2 \cdot x(t)$ is the noise. Note that the environmental noise does not promote the growth of some species, as the demographic noise would do, but has the same effect on all

species, depending on their abundances. Substituting the expression for b, we obtain

$$\begin{aligned}
&\frac{dx(t)}{dt} = (\bar{b} + \sigma_e^2) \cdot x(t) \\
\Rightarrow &\frac{dx(t)}{x(t)} = (\bar{b} + \sigma_e^2)dt \\
\Rightarrow &d(log(x(t))) = (\bar{b} + \sigma_e^2)dt \\
\Rightarrow &\frac{d(log(x(t)))}{dt} = \bar{b} + \sigma_e^2
\end{aligned} \tag{9.2}$$

The logarithm of the species abundances $x(t)$, thus, follows a Brownian motion and its stationary distribution is expected to be a Normal distribution, so that the stationary distribution of $x(t)$ will be a Log-Normal distribution.

9.4 Population dynamic model and the Poisson Log-Normal distribution

A more general stochastic model considers, besides the environmental noise, σ_e^2, also the demographic one (influx), σ_d^2/x, so that the birth rate is given by $b_x = r + \frac{1}{2}\frac{\sigma_d^2}{x} + \frac{1}{2}\sigma_e^2$. Moreover, a density regulation term (death rate) should also be included to take into account that protein domains may be lost or inactivated. An appropriate model for protein domain death seems to be the Gompertzian one, where the death probability exponentially increases with time [110], so that the abundance decrease is described by

$$-\frac{dx}{dt} = rx;\ r = r_0 e^{kt} \tag{9.3}$$

Integrating in dt we obtain $x = x_0 e^{b(1-exp(kt))}$, where $b = r_0/k$, and deriving again for dt we have

$$\begin{aligned}
\frac{dx}{dt} &= -x_0 e^{b(1-exp(kt))} bke^{kt} \\
&= -kxlog(x/x_0) - b,
\end{aligned} \tag{9.4}$$

given that $log(x/x_0) = b(1 - exp(kt))$. In particular, we write the Gompertzian death rate as $d_x = -\gamma x log(x + \epsilon)$, where $\epsilon = \sigma_e^2/\sigma_d^2$. As mentioned before, the Gompertzian hypothesis is a suitable one for protein domains because their loss, or inactivation, becomes more likely as the number of mutations due to replication errors increases. Then, since mutations accumulate over time, it is reasonable to suppose that the probability of having a protein domain death also increases with time. The stochastic Langevin equation for the proposed model is

$$\frac{dx}{dt} = \left[r + \frac{1}{2}\frac{\sigma_d^2}{x} + \frac{1}{2}\sigma_e^2\right] \cdot x - x \cdot \gamma \cdot log\left(x + \frac{\sigma_e^2}{\sigma_d^2}\right) + \sqrt{\sigma_d^2 x + \sigma_e^2 x^2}\frac{dB(t)}{dt} \tag{9.5}$$

and was derived by Engen and Lande [29] as detailed in Sec. 0.7.6. The stationary distribution turns out to be an inhomogeneous Poisson distribution with Log-Normal rate given by

$$\lambda(x) = \frac{\alpha w_0}{x + \epsilon} e^{-\frac{1}{2}\frac{[ln(x+\epsilon) - r/\gamma]^2}{\sigma_e^2/2\gamma}} \tag{9.6}$$

where $\epsilon = \sigma_e^2/\sigma_d^2$ and

$$\alpha = \frac{2}{\sigma_e^2} e^{\frac{\gamma}{\sigma_e^2}\left[ln(1+\epsilon) - \frac{r}{\gamma}\right]^2}. \tag{9.7}$$

Chapter 10

Material and Methods

10.1 Retrieving data

In collaboration with J. Koehorst, E. Saccenti, P. Schaap and M. Suarez-Diez from the Laboratory of Systems and Synthetic Biology of Wageningen University (Netherlands), we downloaded the genome sequences of 3374 bacteria from the NCBI database [111]. In particular, we discarded the so-called draft genome sequences and retained only the higher quality fully circular genome sequences, that (presumably) do not contain any gap and are retrieved upon extensive manual curation and additional sequencing, if required. GeneBank files containing genome sequences and existing annotations were retrieved from the NCBI database and imported into the Semantic Annotation Platform for Prokaryotes [112] using the EMBL/GBK to RDF SAPP module. *De novo* identification of genetic elements (gene calling) was performed using Prodigal (2.6) [113] with codon table 11. Dedicated SPARQL queries were built to extract proteins and their sequences from the RDF triplestore used by SAPP to store the intermediate results. InterProScan [114] was used to identify protein domains in the corresponding sequences. Due to the high number of distinct protein sequences to be analyzed, the SURFsara GRID was used (Grid reference) to concurrently analyze the sequences. Dedicated SPARQL queries were used to retrieve the identified domains and assign them to the originating protein and bacterial genome. Finally, the matrix generating module from SAPP was used to generate a matrix containing for each of the studied genomes and for each of the identified protein domains the number of instances of the detected domain (domain abundance). Overall 3374 bacterial genomes were analyzed and 13934 distinct domains were identified.

10.2 Fitting the protein domains RSA

Protein domains RSAs were fitted with the Maximum Likelihood Estimation method implemented in R 'sads' package v.0.3.1 [115]. As mentioned in Sec. 2.5, even if the common way to visualize the RSA is in the form of Preston plot, it is better not to use this representation when fitting experimental data. The Preston plot, in fact, is an histogram with logarithmic bins in which every bin averages the information between its minimum and maximum. The loss of information derived by this representation makes it difficult to evaluate the model performances and, more importantly, to distinguish between similar distributions, such as the Poisson Log-Normal and the Negative Binomial. In addition, even if the RSA predicted by the proposed model is a Poisson Log-normal that depends on $x + \epsilon = x + \sigma_e^2/\sigma_d^2$ (see Eq. 9.6 in Sec. 9.4), we chose to fit the stationary distribution in terms of x, neglecting the transposition term, so that to use the likelihood definition derived in [116].
We modeled the data both with a truncated Poisson Log-Normal distribution and with

a truncated Negative Binomial, so that to test and compare different ecological hypothesis. In both cases, truncation was performed to exclude the 0 abundance class, that is indeed not observable in empirical data. Akaike Information Criterion (AIC) [72] and R-squared were computed to assess the models performances and comparison [115] (see Part II Sec. 2.5.3).

10.3 Comparing RSA parameters with taxonomy

One sample was excluded because its Poisson Log-Normal parameters μ and σ were outliers and its RSA was peculiar, having a maximum abundance of 32. For this reason, in the Results we will refer to 3373 bacterial genomes.
In order to compare the distribution of μ and σ with a null model, we randomly shuffled the abundances of protein domains between different Bacteria.
Principal Component Analysis with covariates μ, σ and the total number of protein domains species S was performed to determine whether bacteria belonging to different taxa are characterized by particular model parameters and cluster together. We performed a Ward hierarchical clustering [117] based on the first two PCA components, computing a number of cluster equal to the number of Species. Finally, we compared the clustering in the PCA space with the taxonomic classification at the level of Species using the Normalized Mutual Information (NMI) [97] score, that we already defined in Part II Sec. 6.6.

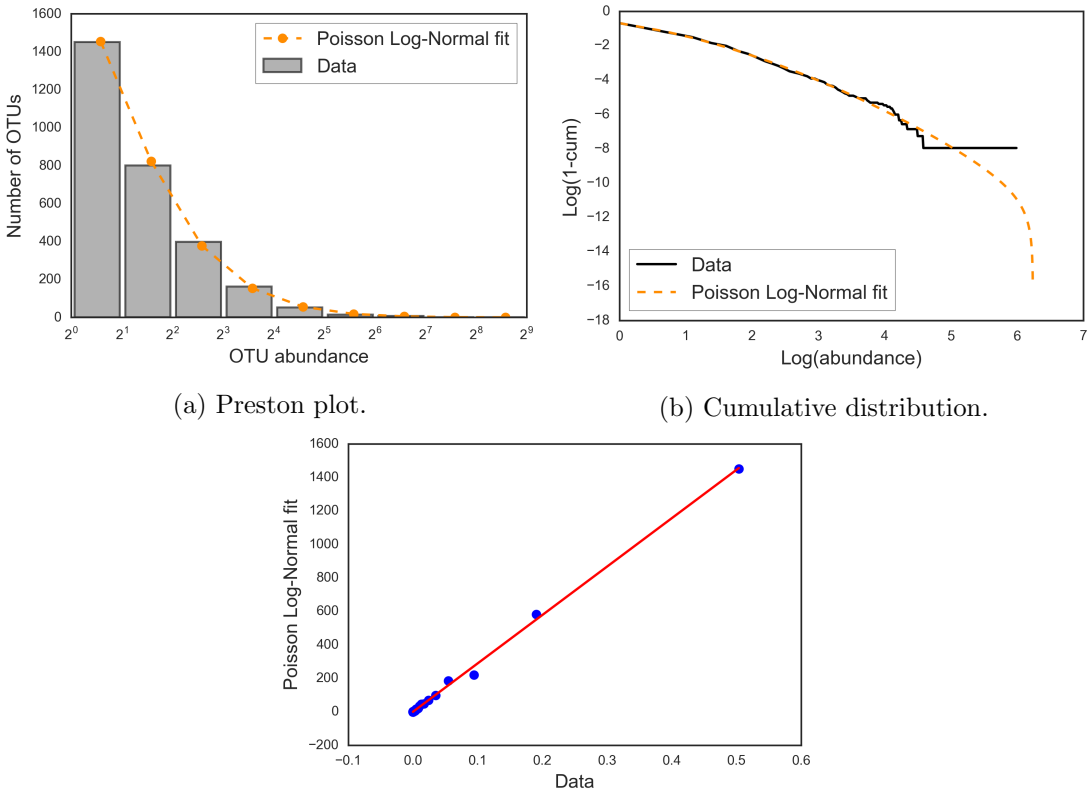
10.4 Comparing RSA parameters with phylogeny

As detailed in Part I and II, taxonomy is a human made classification that not always reflects the evolutionary relations between bacteria. Hence, in order to compare the clustering based on the model parameters with the phylogenetic tree, we retrieved the 16S rRNA reference sequences of the considered bacteria from the silva database [118]. For 248 bacteria, the 16S rRNA sequence was not present, so we considered only the remaining 3124 for the following analysis.
Sequences were aligned using the SINA aligner integrated in the silva web tool [118] and the distance matrix was then computed with mothur [67], as in Sec. 6.1. For each taxonomic level, we considered only bacteria for which the classification was known and we compared the clustering based on the protein domains RSA and the one computed using the 16S rRNA distance matrix. As before, the RSA based clustering was founded on the first two principal components of the Poisson Log-Normal parameters obtained from the fit. In both cases we performed a Ward hierarchical clustering fixing the number of clusters to the number of taxa at the selected taxonomic level. Then, we used the NMI score as measure of clustering agreement. Finally, both clustering outcomes were also compared with the taxonomic classification.

Chapter 11

Results

11.1 Model selection



(a) Preston plot.

(b) Cumulative distribution.

(c) q-q plot.

Figure 11.1: Example of Preston plot (a), cumulative RSA (b) and q-q plot (c) obtained fitting the protein domain population of one of the selected bacterial genomes.

The Akaike information Criterion (AIC) [72] (see Sec. 0.6.1) preferred the Poisson Log-Normal model to the Negative Binomial in $\sim 94\%$ of the 3373 fitted bacterial genomes. The proposed model resulted better also compared to the Log-Series, that depends only on one parameter, having smaller AIC in $\sim 99\%$ of the cases. The mean R-squared for the Poisson Log-Normal model was 0.96 with a minimum value of 0.86. Fig. 11.1 shows one example in which data were fitted with the Poisson Log-Normal distribution. The upper-left figure (a) shows the empirical Preston plot and the fitted curve overlapped in orange. The upper-right plot (b) represents the logarithm of $(1 - cum)$, where cum is the cumulative distribution of the data RSA (black) or of the Poisson Log-Normal

(orange). Finally, the bottom figure (c) is the quantile-quantile plot, that also indicates good agreement between the empirical and predicted data.

11.2 Comparing RSA and taxonomy

Plotting the Poisson Log-Normal location parameter μ as a function of the square of the scale parameter σ^2 obtained for all Bacteria, reveals an inverse relationship between the two parameters (see Fig. 11.2 (left)).

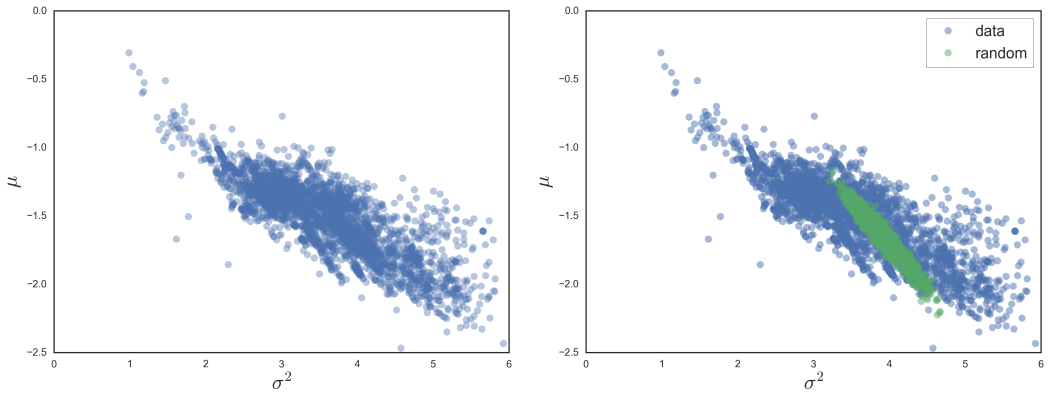


Figure 11.2: Plot of the Poisson Log-Normal parameters μ versus σ^2 obtained for the 3373 bacterial genomes. In the right-hand figure, green dots represent the results for 3373 random populations of protein domains.

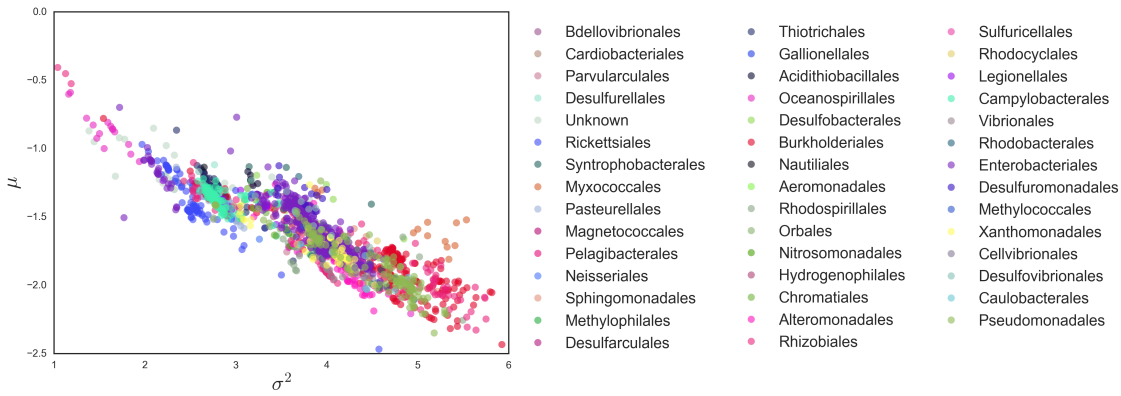


Figure 11.3: Plot of the Poisson Log-Normal parameters μ versus σ^2 obtained fitting the protein domains RSA of bacteria belonging to the Proteobacteria phylum. Different colors correspond to different orders as indicated in the legend .

We may recognize the presence of roughly parallel stripes in the plot, that suggests a cluster structure of the data. If we consider the null model obtained by randomly shuffling the protein domain abundances between different bacteria, the resulting Preston plot is still well fitted by the Poisson Log-Normal distribution and its parameters are distributed as a single stripe in the μ versus σ^2 plot, as shown in Fig. 11.2 (right). In Fig. 11.3, we plotted with different colors different orders of the Proteobacteria phylum and it emerged that the stripes are related to the bacterial phylogeny. The parameters of the various bacterial genomes are not simply sampled from the null model, but are characterized by different dynamic parameters. In Sec. 11.3 we will propose a Principal Component Analysis representation in which the relation between the phylogenetic structure and the model parameters is even clearer.

We may also observe that the location parameter μ has negative values. This seems to be counterintuitive since we defined $\mu = r/\gamma$ (see Eq. 9.6), and both r and γ are supposed to be positive. However, as mentioned before, the theoretical RSA is a Poisson Log-Normal distribution in $x + \epsilon = x + \frac{\sigma_e^2}{\sigma_d^2}$ but, for simplicity, we neglected the translation term during the fitting. However, the negativity of μ suggests that the effect of $\epsilon = \sigma_e^2/\sigma_d^2$ is not negligible compared to x and it is actually greater than r/γ. To understand the shifting effect of ϵ, we can rewrite the exponential term of the Log-Normal in Eq. 9.6, $[log(x+\epsilon) - r/\gamma]$, as $[log(x/\epsilon+1) - (r/\gamma - log(\epsilon)]$, so that the location parameter becomes

$$\mu = \frac{r}{\gamma} - log(\epsilon) = \frac{r}{\gamma} - log(\frac{\sigma_e^2}{\sigma_d^2}) = [\frac{r}{\gamma} - log(\frac{\sigma_e^2}{\gamma}) + log(\frac{\sigma_d^2}{\gamma})] \tag{11.1}$$

Moreover, since $\frac{\sigma_e^2}{\gamma} = \sigma^2$ is the Log-Normal scale parameter, we can understand now the inverse relationship between μ and σ, or better $log(\sigma^2)$.

11.3 Protein domains RSAs and evolutionary distance

As noticed in the previous section, the model parameters for different bacteria seem to reflect their phylogenetic relationship. For this reason, we performed a Principal Component Analysis in which we transformed the data based on μ, σ and the total number of protein domain species S, so that to visualize as much of their variance as possible. Considering the 3251 bacteria for which the Family classification was known, the first two PCA components turn out to be

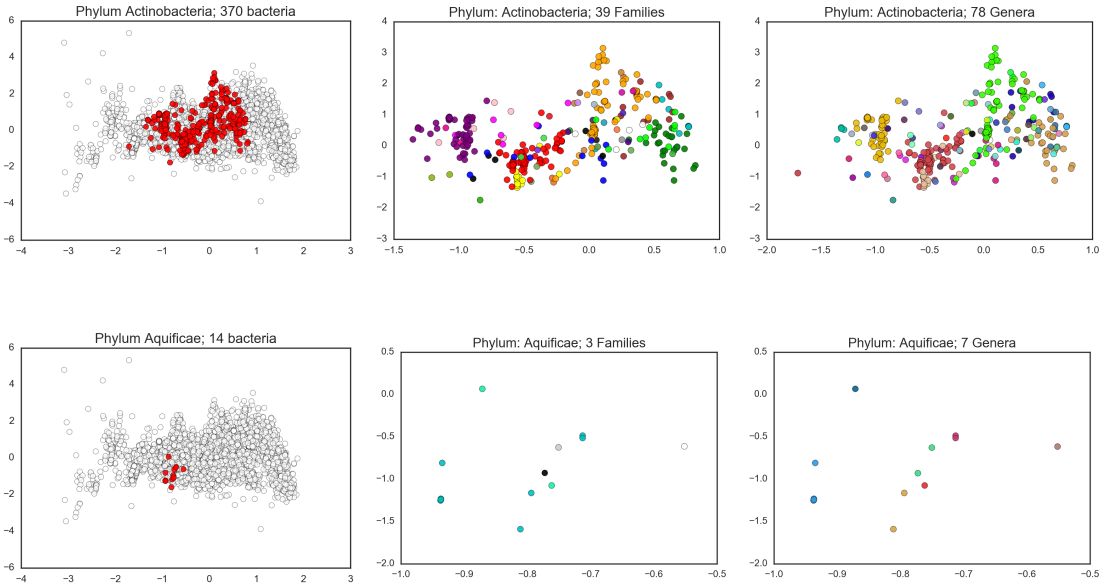
$$\begin{aligned} x &= -0.00014\mu + 0.00016\sigma + 0.99999S \\ y &= -0.88700\mu + 0.46176\sigma - 0.00020S \end{aligned} \tag{11.2}$$

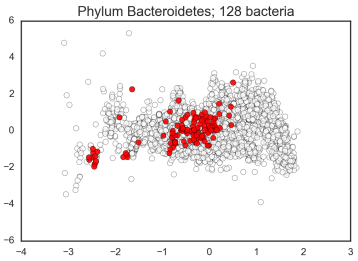
The x-axis is almost entirely determined by S, indicating that what mostly discriminates bacteria is the number of different protein species. This is highly correlated with the total number of protein domains, and finally with the genome length, so that it is comprehensible that it may be related to the phylogeny. However, the y-axis is dominated by μ and σ. The following figures show the 3251 bacteria for which we represented in the PCA components space. We considered the Phyla for which we had at least 10 genomes. In the figures on the left column, we plotted in white the parameters of all 3251 bacteria and in red those of the considered phylum. In the central column, we only considered the selected phylum and plotted bacteria belonging to different families with different colors. Finally, on the right-hand column we used different colors to represent different genera. All plots show that bacteria that have the same taxonomic classification tend to cluster together. Accordingly, the Normalized Mutual Information (NMI) score between the taxonomic classification at the Species level and the hierarchical clustering based on the two PCA components is 0.870, indicating good agreement. As expected, most of the variation that discriminates bacteria in the PCA plots is enclosed in the x-axis, dominated by S. However, some families and genera are separated towards the y-axis. This is for example the case of the red and yellow families of the Actinobacteria phylum and suggests that our modeling is actually capturing the mechanisms of genome evolution.
Finally, Tab. 11.1 shows the NMI scores obtained when comparing the clustering based on the protein domains RSA, i.e. on the Poisson Log-Normal parameters, with the one based on the 16S rRNA distances, that reflects the phylogenetic tree. Here we considered only the 3124 bacteria for which the 16S rRNA reference sequence was available in the silva database and, at each taxonomic level, we also excluded bacteria for which the taxonomic classification was not known. The row 'Number of Bacteria' reports the total

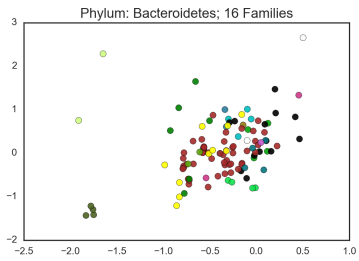
number of bacteria considered. Comparing the rows 'p. domains RSA VS taxonomy' and 'p. domains RSA VS 16S rRNA', we notice that the groups of bacteria determined by the RSAs parameters reflect better the phylogenetic distances rather than the taxonomic classification. We can conclude that the protein domains RSA of bacteria that are close in the evolutionary tree have similar shapes, reflecting the closness with their common ancestor. Moreover, these results confirm what we already stated in Part I and II, i.e. that taxonomy is a biased classification. Although taxonomy and phylogeny are, as expected, strictly related and have the highest NMI scores in Tab. 11.1, protein domains RSAs, that describe the genome composition and evolutionary dynamics, are more congruent with the 16S rRNA based clustering. Lastly, note that the NMI scores decreases when considering higher taxonomic levels (i.e. from Species to Phylum). This may be due to the fact that at the Phylum level, for instance, the number of clusters is much lower than at the Species level, as indicated by the 'Number of Clusters' row. Interestingly, also the agreement between phylogeny and taxonomy decreases at high taxonomic levels, and this may suggest that the effect we observe is connected with the high heterogeneity of the bacterial population at these levels.

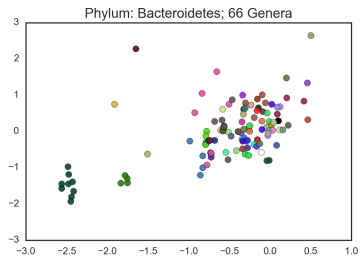
	Phylum	Class	Order	Family	Genus	Species
Number of Bacteria	3123	3043	3079	3003	3096	3124
Number of Clusters	31	52	126	238	655	1523
P. domains RSA VS taxonomy	0.259	0.353	0.530	0.630	0.744	0.876
P. domains RSA VS 16S rRNA	0.428	0.506	0.600	0.660	0.770	0.882
16S rRNA VS taxonomy	0.671	0.790	0.869	0.902	0.928	0.960

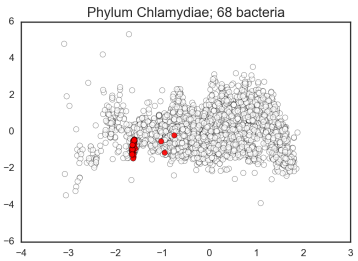
Table 11.1: NMI scores that compare the clustering based on the protein domains RSAs, the one based on the 16S rRNA distances and the taxonomic classification of bacteria. Results are shown at different taxonomic level, as indicated by the column names.

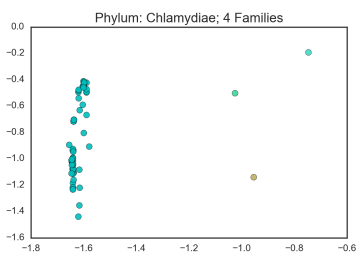


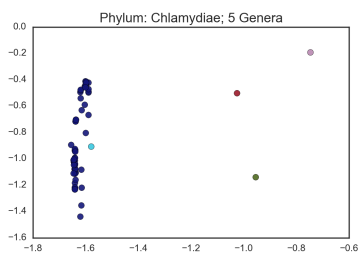

Phylum Bacteroidetes; 128 bacteria

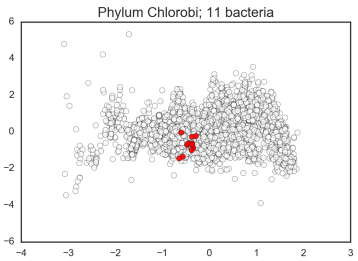

Phylum: Bacteroidetes; 16 Families

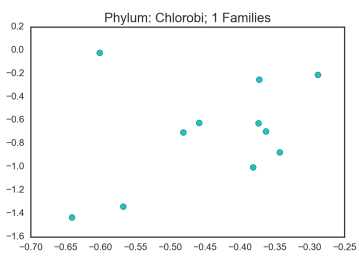

Phylum: Bacteroidetes; 66 Genera

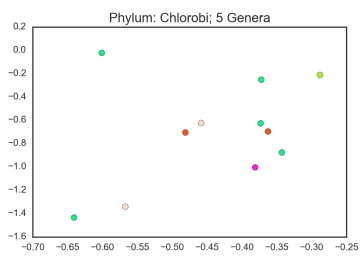

Phylum Chlamydiae; 68 bacteria

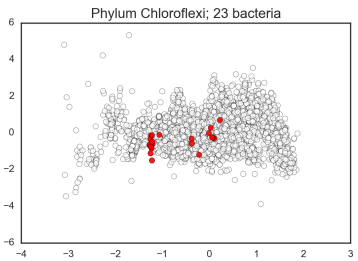

Phylum: Chlamydiae; 4 Families

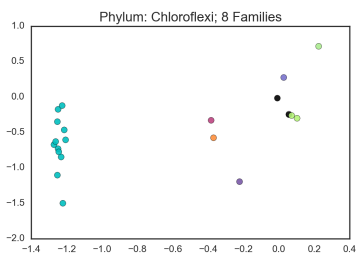

Phylum: Chlamydiae; 5 Genera

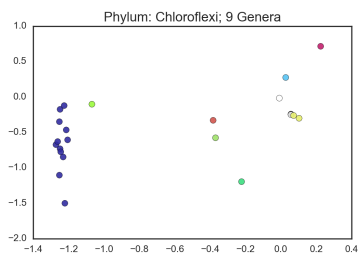

Phylum Chlorobi; 11 bacteria

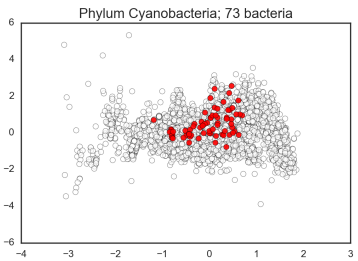

Phylum: Chlorobi; 1 Families

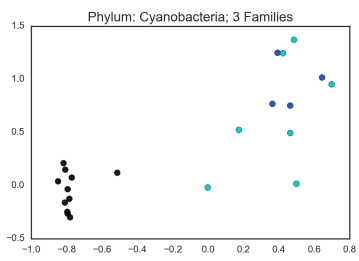

Phylum: Chlorobi; 5 Genera

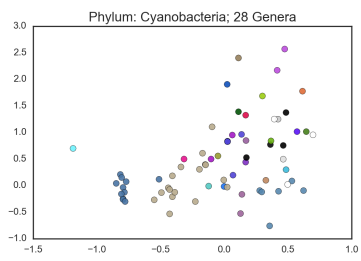

Phylum Chloroflexi; 23 bacteria

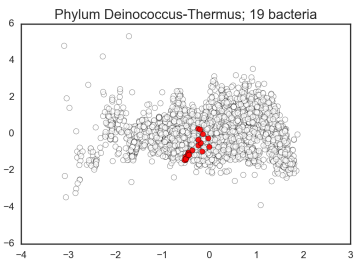

Phylum: Chloroflexi; 8 Families

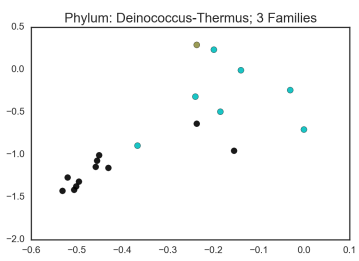

Phylum: Chloroflexi; 9 Genera

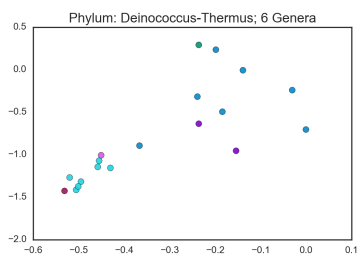

Phylum Cyanobacteria; 73 bacteria


Phylum: Cyanobacteria; 3 Families


Phylum: Cyanobacteria; 28 Genera


Phylum Deinococcus-Thermus; 19 bacteria


Phylum: Deinococcus-Thermus; 3 Families


Phylum: Deinococcus-Thermus; 6 Genera

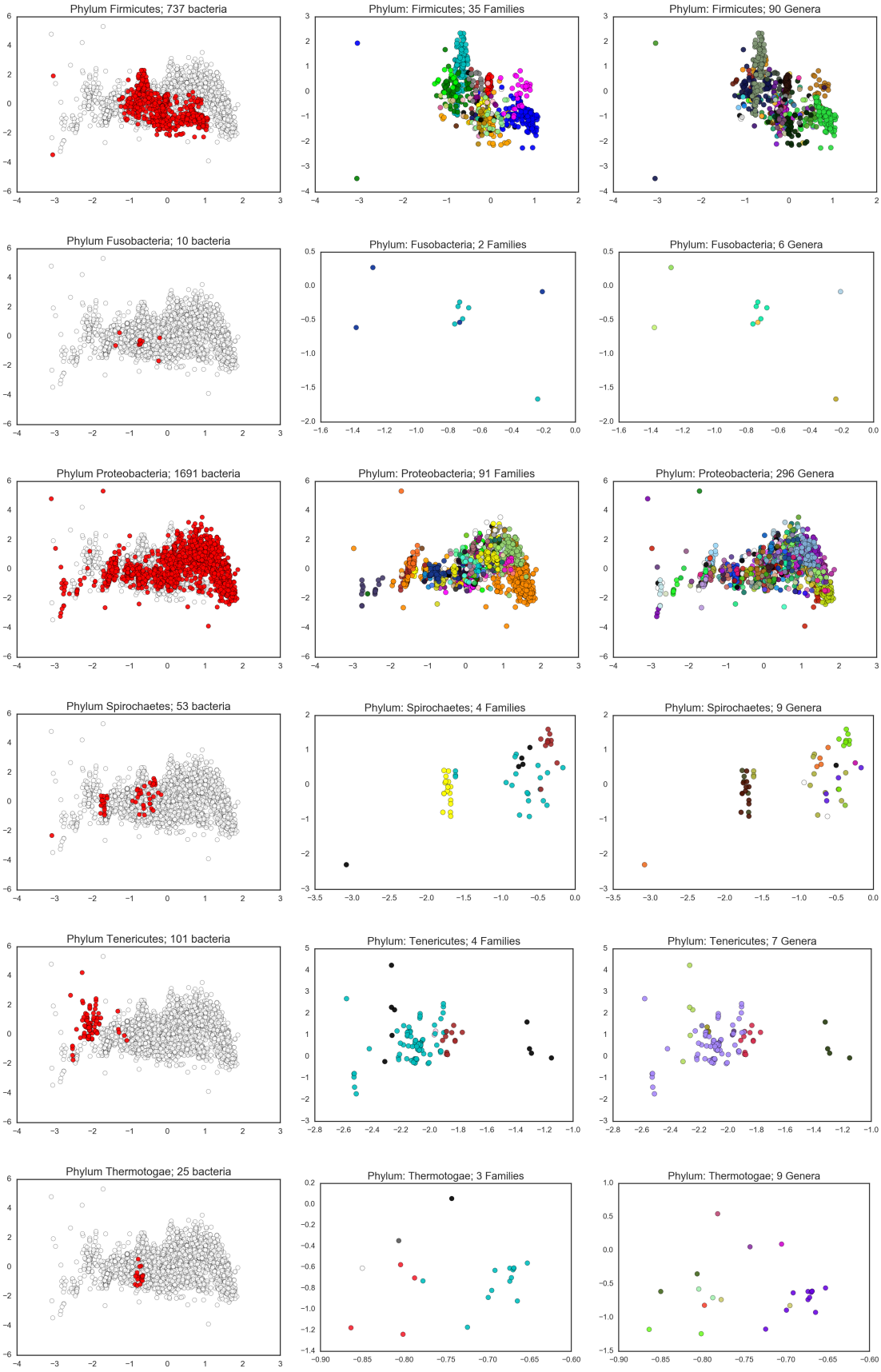


Figure 11.17: Representation of the 3251 bacterial genomes for which the Species classification is known in the space defined by the first two PCA components. Left: red dots correspond to the phylum specified in the title, white dots represent all the other bacteria. Center: different color correspond to different families of the selected phylum. Right: different color correspond to different genera of the selected phylum.

Chapter 12

Discussion

The birth death model proposed by Volkov [24] is not a good assumption for the protein domains ecosystem. Unlike what we found for the Gut Microbiota, the problem here is not the violation of the neutrality hypothesis, but the importance of environmental noise, that was instead neglected by Volkov. Moreover, when dealing with genome evolution, a more suitable model for the death rate is the Gompertzian one, in which the death probability increases with time. Under these new assumptions, the predicted stationary distribution for the RSA is a Poisson Log-Normal [29]. The model now fits well the protein domain abundances, as we verified on more than 3000 bacterial genomes, and has better performances compared to Volkov's Negative Binomial. Our results show that the demographic noise may also be important, even if the environmental one turns out to prevail over the whole term $r/\gamma + log(\sigma_d^2)$. Finally, the differences in the model parameters reflect the evolutionary distances among bacteria. The variable that mostly discriminates among taxonomies is the total number of protein domains in the genome. This is coherent with the fact that genome evolution manifests, for instance, through differences in the genome length, that will be related to the number of protein domains. Nevertheless, some genera that have approximately the same number of protein domains, are discriminated by the other model parameters. This suggests that different taxa are characterized by different dynamic rates. The evolutionary model is the same for all genomes, meaning that the underlying process is the same. However, the parameters have some discrepancies and this indicates a niche structure in the bacterial genomes, where niches correspond to clusters in the phylogenetic tree. To better understand this concept, we may think of considering all the protein domains of all bacteria together, as a single ecosystem. Our results suggest that, in this case, neutrality would not hold but we would find separate niches that, at last, will include protein domains that belong to different taxa (i.e. phyla, genera, species, etc.), reflecting the phylogenetic relationships among bacteria.
The clusters obtained according the RSA parameters show also good agreement with the 16S rRNA based phylogenetic distances, and this confirms that the Poisson Log-Normal model catches the mechanisms of genome evolution. We know, in fact, that bacteria that are close from an evolutionary point of view, inherited their genomes from a close common ancestor, and this is also why they have similar 16S rRNA sequences. Their genomes will consequently be pretty related and their dynamic rates will also be expected to be more similar than those of farther bacteria. It is clear, then, why obtaining a coherence between the phylogenetic tree and the RSA modeling is an important result.
Moreover, in support to our hypothesis of genome evolution, in a previous work by Furusawa et al. [109] it was reported that the abundances of proteins inside the cell are Log-Normal and, in an unpublished work, we also observed that the lengths of the non-coding region of genes follow the same distribution.
To conclude, we mention that an interesting development of this work would be to derive

an estimate for the expected protein domains formation time, as in the work by Karev et al. [107], and to verify whether we obtain a more reliable measure.

Conclusion

In this thesis we aimed to describe the dynamic processes that govern the evolution of two very different ecological systems. First, we considered the ensemble of bacteria that populate the gut (Gut Microbiota), that has been proven to have great impact on human health, being associated, for instance, to several metabolic and immunological diseases. Then, we dealt with the set of protein domains that populate the genome of living organisms. In particular, we recovered the abundances of protein domains for several bacterial genomes with the aim of revealing the main mechanisms of genome evolution.
In general, the neutrality hypothesis, that was proposed by Hubbell as the Ockham's razor for ecology [5], is a good approximation for both the Gut Microbiota and the protein domains ecosystems. In the first case, however, the equivalence of species is not entirely valid and a better description is obtained relaxing the neutrality assumption with the introduction of two non-interacting populations.
In Part I we have proved the importance of the use of a stochastic framework when dealing with biological systems. Taking noise into account, in fact, enabled us to derive a biodiversity index that distinguished between healthy and unhealthy aging. Biodiversity was already supposed to play an important role in the relationship between the Gut Microbiota and the host health. However, using common diversity indices, that are simply based on the relative abundances of species, would not allow us to obtain the same prediction accuracy achieved with our stochastic modeling.
When constructing the empirical distribution of the Gut Microbiota abundances, a fundamental step regards the clustering of particular highly conserved DNA sequences (usually the 16S rRNA gene). This procedure enables to redefine the concept of species, that is now referred to as Operational Taxonomic Unit, so that to rely only on the phylogenetic relationships between bacteria rather then on human made classifications. Many algorithms have been conceived to fulfill this task and one of the main issues that still has not been solved is the requirement of a predefined and mostly arbitrary threshold. For this reason, in Part II we proposed the use of LOC-kNN, an original clustering method that was recently developed by M. d'Errico et al. [79]. This is a totally parameter-free algorithm that aims to reconstruct the data topography and shows promising performances on a simulated dataset, even if some issues will need to be solved in the future to obtain a better estimate of rarely populated species.
While for the Gut Microbiota the simple birth-death-speciation model proposed by Volkov [24] was a good approximation of the data, given the above mentioned relaxation of the neutrality hypothesis, this was not true for the protein domain ecosystem. In Part III we showed that in this case, besides demographic stochasticity, also environmental noise should be taken into account. Moreover, a Gompertzian death model seems more appropriate for protein domains, since their loss or inactivation mainly happens because of mutations or errors that accumulate over time. Following these considerations we were able to model the Relative Species Abundance distribution of the protein domains belonging to the retrieved $\sim$3000 bacterial genomes. Interestingly, we observed that bacteria that are close in the phylogenetic tree, i.e. that have a close common ancestor, also have similar dynamic rates, suggesting that we are capturing in our modeling the fundamental

processes of genome evolution.
To conclude, stochastic modeling is a powerful tool when studying biological systems, in which noise may have important statistical consequences. In particular, in ecology, the inclusion of different sources of noise in the model, enables to postulate and test hypothesis on the dynamics of populations. This is true for more standard ecosystems, such as the Gut Microbiota, but also, for instance, when studying the distribution of the protein domains species in the genome, and allows us to achieve interesting objectives such as the construction of predictive models and a deeper understanding of the systems biology.

List of publications

1. (Submitted in ONCOTARGET) D.F. Durso, M.G. Bacalini, C. Sala, C. Pirazzini, E. Marasco, M. Bonafè, Ì.F. do Valle, D. Gentilini, G. Castellani, A.M.C. Faria, C. Franceschi, C. Nardini, P. Garagnani. Acceleration of leukocytes' epigenetic age as an early tumor- and sex-specific marker of breast and colorectal cancer.

2. (Submitted in PLOSONE) A. De Cesare, Ì.F. do Valle, C. Sala, P. Moniaci, A. Astolfi, G. Castellani, G. Manfreda. Supplementation of serine protease in a low protein diet: effects on taxonomic composition and functional genes in chicken caeca assessed using shotgun metagenomic sequencing.

3. M. Bersanelli, E. Mosca, D. Remondini, E. Giampieri, C. Sala, G. Castellani, L. Milanesi. Methods for the integration of multi-omics data: Mathematical aspects. BMC BIOINFORMATICS. 2016, 17 Suppl 2, pp. 15-25.

4. C. Sala, S. Vitali, E. Giampieri, Í.F. do Valle, D. Remondini, P. Garagnani, M. Bersanelli, E. Mosca, L. Milanesi, G. Castellani, Stochastic neutral modelling of the Gut Microbiota's relative species abundance from next generation sequencing data. BMC BIOINFORMATICS. 2016, 17 Suppl 2, pp. 16 - 25.

5. G. Castellani, G. Menichetti, P. Garagnani, M.G. Bacalini, C. Pirazzini, C. Franceschi, S. Collino, C. Sala, D. Remondini, E. Giampieri, E. Mosca, M. Bersanelli, S. Vitali, Ì.F. do Valle, P. Liò, L. Milanesi. Systems medicine of inflammaging. BRIEFINGS IN BIOINFORMATICS. 2016, 17, pp. 527 - 540.

6. E. Giampieri, D. Remondini, M.G. Bacalini, P. Garagnani, C. Pirazzini, S.L. Yani, C. Giuliani, G. Menichetti, Giulia, I. Zironi, C. Sala, M. Capri, C. Franceschi, A. Burkle, G. Castellani. Statistical strategies and stochastic predictive models for the MARK-AGE data. MECHANISMS OF AGEING AND DEVELOPMENT. 2015, 151, pp. 45 - 53.

7. D. Calçada, D. Vianello, E. Giampieri, C. Sala, G. Castellani, A. de Graaf, B. Kremer, B. van Ommen, E. Feskens, A. Santoro, C. Franceschi, J. Bouwman. The role of low-grade inflammation and metabolic flexibility in aging and nutritional modulation thereof: a systems biology approach. MECHANISMS OF AGEING AND DEVELOPMENT, 2014, 136-137, pp. 138 - 147.

Bibliography

[1] J. Lei, "Stochastic modeling in systems biology," *J. Adv. Math. Appl*, vol. 1, no. 1, pp. 76–88, 2012.

[2] P. S. Swain, M. B. Elowitz, and E. D. Siggia, "Intrinsic and extrinsic contributions to stochasticity in gene expression," *Proceedings of the National Academy of Sciences*, vol. 99, no. 20, pp. 12795–12800, 2002.

[3] V. Shahrezaei and P. S. Swain, "The stochastic nature of biochemical networks," *Current opinion in biotechnology*, vol. 19, no. 4, pp. 369–374, 2008.

[4] R. Lande, S. Engen, and B.-E. Saether, *Stochastic population dynamics in ecology and conservation*. Oxford University Press on Demand, 2003.

[5] S. P. Hubbell, *The unified neutral theory of biodiversity and biogeography*. Princeton University Press, 2001.

[6] S. Karlin and H. E. Taylor, *A second course in stochastic processes*. Elsevier, 1981.

[7] N. Van Kampen, *Stochastic processes in chemistry and physics*. Elsevier, 1981-2007.

[8] F. S. Chapin III, P. A. Matson, and P. Vitousek, *Principles of terrestrial ecosystem ecology*. Springer Science & Business Media, 2011.

[9] B. J. Cardinale, J. E. Duffy, A. Gonzalez, D. U. Hooper, C. Perrings, P. Venail, A. Narwani, G. M. Mace, D. Tilman, D. A. Wardle, *et al.*, "Biodiversity loss and its impact on humanity," *Nature*, vol. 486, no. 7401, pp. 59–67, 2012.

[10] R. H. Whittaker, "Vegetation of the siskiyou mountains, oregon and california," *Ecological monographs*, vol. 30, no. 3, pp. 279–338, 1960.

[11] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. July 1928, pp. 379–423, 1948.

[12] E. C. Pielou, "The measurement of diversity in different types of biological collections," *Journal of theoretical biology*, vol. 13, pp. 131–144, 1966.

[13] E. H. Simpson, "Measurement of diversity.," *Nature*, 1949.

[14] R. Kempton and L. Taylor, "Models and statistics for species diversity," 1976.

[15] R. Fisher, A. Corbet, and C. Williams, "The relation between the number of species and the number of individuals in a random sample of an animal population," *The Journal of Animal Ecology*, vol. 12, no. 1, pp. 42–58, 1943.

[16] F. W. Preston, "The commonness, and rarity, of species," *Ecology*, vol. 29, no. 3, pp. 254–283, 1948.

[17] R. H. MacArthur, "On the relative abundance of bird species," *Proceedings of the National Academy of Sciences*, vol. 43, no. 3, pp. 293–295, 1957.

[18] I. Motomura, "A statistical treatment of associations.," *Japanese Journal of Zoology*, vol. 44, pp. 379–383, 1932.

[19] R. H. Whittaker, "Dominance and diversity in land plant communities," *Science*, vol. 147, no. 3655, pp. 250–260, 1965.

[20] G. Sugihara, "Minimal community structure: an explanation of species abundance patterns," *American naturalist*, pp. 770–787, 1980.

[21] R. MacArthur and E. Wilson, *The Theory of Island Biogeography.* Princeton Univesity Press, 1967.

[22] S. P. Hubbell, "Neutral theory in community ecology and the hypothesis of functional equivalence," *Functional ecology*, vol. 19, no. 1, pp. 166–172, 2005.

[23] R. S. Etienne and D. Alonso, "A dispersal-limited sampling theory for species and alleles," *Ecology letters*, vol. 8, no. 11, pp. 1147–1156, 2005.

[24] I. Volkov, J. R. Banavar, S. P. Hubbell, and A. Maritan, "Patterns of relative species abundance in rainforests and coral reefs.," *Nature*, vol. 450, pp. 45–9, nov 2007.

[25] W. J. Ewens, "The sampling theory of selectively neutral alleles," *Theoretical population biology*, vol. 3, no. 1, pp. 87–112, 1972.

[26] R. C. Griffiths and S. Lessard, "Ewens' sampling formula and related formulae: combinatorial proofs, extensions to variable population size and applications to ages of alleles," *Theoretical population biology*, vol. 68, no. 3, pp. 167–177, 2005.

[27] M. Vallade and B. Houchmandzadeh, "Analytical solution of a neutral model of biodiversity," *Physical Review E*, vol. 68, no. 6, p. 061902, 2003.

[28] R. S. Etienne and H. Olff, "Confronting different models of community structure to species-abundance data: a bayesian model comparison," *Ecology letters*, vol. 8, no. 5, pp. 493–504, 2005.

[29] S. Engen and R. Lande, "Population dynamic models generating the lognormal species abundance distribution.," *Mathematical biosciences*, vol. 132, pp. 169–83, mar 1996.

[30] C. M. Guinane and P. D. Cotter, "Role of the gut microbiota in health and chronic gastrointestinal disease: understanding a hidden metabolic organ.," *Therapeutic advances in gastroenterology*, vol. 6, no. 4, pp. 295–308, 2013.

[31] S. R. Gill, M. Pop, R. T. DeBoy, P. B. Eckburg, P. J. Turnbaugh, B. S. Samuel, J. I. Gordon, D. A. Relman, C. M. Fraser-Liggett, and K. E. Nelson, "Metagenomic analysis of the human distal gut microbiome," *science*, vol. 312, no. 5778, pp. 1355–1359, 2006.

[32] K. Fujimura and N. Slusher, "Role of the gut microbiota in defining human health," *Expert review of anti- ...*, vol. 8, no. 4, pp. 435–454, 2010.

[33] D. Marco, *Metagenomics: Current Innovations and Future Trends.* Horizon Scientific Press, 2011.

[34] A. Mosca, M. Leclerc, and J. P. Hugot, "Gut Microbiota Diversity and Human Diseases: Should We Reintroduce Key Predators in Our Ecosystem?," *Frontiers in Microbiology*, vol. 7, no. March, pp. 1–12, 2016.

[35] P. D. Cani and N. M. Delzenne, "The role of the gut microbiota in energy metabolism and metabolic disease.," *Current pharmaceutical design*, vol. 15, pp. 1546–1558, 2009.

[36] G. Musso, R. Gambino, and M. Cassader, "Interactions between gut microbiota and host metabolism predisposing to obesity and diabetes," *Annual review of medicine*, vol. 62, pp. 361–380, 2011.

[37] J. Qin, Y. Li, Z. Cai, S. Li, J. Zhu, F. Zhang, S. Liang, W. Zhang, Y. Guan, D. Shen, Y. Peng, D. Zhang, Z. Jie, W. Wu, Y. Qin, W. Xue, J. Li, L. Han, D. Lu, P. Wu, Y. Dai, X. Sun, Z. Li, A. Tang, S. Zhong, X. Li, W. Chen, R. Xu, M. Wang, Q. Feng, M. Gong, J. Yu, Y. Zhang, M. Zhang, T. Hansen, G. Sanchez, J. Raes, G. Falony, S. Okuda, M. Almeida, E. LeChatelier, P. Renault, N. Pons, J.-M. Batto, Z. Zhang, H. Chen, R. Yang, W. Zheng, S. Li, H. Yang, J. Wang, S. D. Ehrlich, R. Nielsen, O. Pedersen, K. Kristiansen, and J. Wang, "A metagenome-wide association study of gut microbiota in type 2 diabetes.," *Nature*, vol. 490, pp. 55–60, oct 2012.

[38] L. Zhao, "The gut microbiota and obesity: from correlation to causality," *Nature Reviews Microbiology*, vol. 11, no. 9, pp. 639–647, 2013.

[39] L. V. Hooper, D. R. Littman, and A. J. Macpherson, "Interactions between the microbiota and the immune system," *Science*, vol. 336, no. 6086, pp. 1268–1273, 2012.

[40] P. Bhargava and E. M. Mowry, "Gut microbiome and multiple sclerosis," *Current neurology and neuroscience reports*, vol. 14, pp. 1–8, 2014.

[41] J. Penders, E. E. Stobberingh, P. A. van den Brandt, and C. Thijs, "The role of the intestinal microbiota in the development of atopic disorders," *Allergy*, vol. 62, no. 11, pp. 1223–1236, 2007.

[42] P. D. Scanlan, F. Shanahan, Y. Clune, J. K. Collins, G. C. O'Sullivan, M. O'Riordan, E. Holmes, Y. Wang, and J. R. Marchesi, "Culture-independent analysis of the gut microbiota in colorectal cancer and polyposis," *Environmental microbiology*, vol. 10, no. 3, pp. 789–798, 2008.

[43] J. Vaahtovuo, E. Munukka, M. KORKEAMÄKI, R. Luukkainen, and P. Toivanen, "Fecal microbiota in early rheumatoid arthritis," *The Journal of rheumatology*, vol. 35, no. 8, pp. 1500–1505, 2008.

[44] S. Collins, E. Denou, E. Verdu, and P. Bercik, "The putative role of the intestinal microbiota in the irritable bowel syndrome," *Digestive and Liver disease*, vol. 41, no. 12, pp. 850–853, 2009.

[45] J. Y. Chang, D. A. Antonopoulos, A. Kalra, A. Tonelli, W. T. Khalife, T. M. Schmidt, and V. B. Young, "Decreased diversity of the fecal microbiome in recurrent clostridium difficile—associated diarrhea," *Journal of Infectious Diseases*, vol. 197, no. 3, pp. 435–438, 2008.

[46] C. Lupp, M. L. Robertson, M. E. Wickham, I. Sekirov, O. L. Champion, E. C. Gaynor, and B. B. Finlay, "Host-Mediated Inflammation Disrupts the Intestinal Microbiota and Promotes the Overgrowth of Enterobacteriaceae," *Cell Host and Microbe*, vol. 2, no. 2, pp. 119–129, 2007.

[47] J. R. Goldsmith and R. B. Sartor, "The role of diet on intestinal microbiota metabolism: downstream impacts on host immune function and health and therapeutic implications," *Journal of Gastroenterology*, vol. 49, pp. 785–798, mar 2014.

[48] P. J. Turnbaugh, M. Hamady, T. Yatsunenko, B. L. Cantarel, A. Duncan, R. E. Ley, M. L. Sogin, W. J. Jones, B. a. Roe, J. P. Affourtit, M. Egholm, B. Henrissat, A. C. Heath, R. Knight, and J. I. Gordon, "A core gut microbiome in obese and lean twins.," *Nature*, vol. 457, pp. 480–4, jan 2009.

[49] A. Cotillard, S. P. Kennedy, L. C. Kong, E. Prifti, N. Pons, E. Le Chatelier, M. Almeida, B. Quinquis, F. Levenez, N. Galleron, *et al.*, "Dietary intervention impact on gut microbial gene richness," *Nature*, vol. 500, no. 7464, pp. 585–588, 2013.

[50] A. Tagliabue and M. Elli, "The role of gut microbiota in human obesity: recent findings and future perspectives," *Nutrition, Metabolism and Cardiovascular Diseases*, vol. 23, no. 3, pp. 160–168, 2013.

[51] G. D. Wu, J. Chen, C. Hoffmann, K. Bittinger, Y.-Y. Chen, S. a. Keilbaugh, M. Bewtra, D. Knights, W. a. Walters, R. Knight, R. Sinha, E. Gilroy, K. Gupta, R. Baldassano, L. Nessel, H. Li, F. D. Bushman, and J. D. Lewis, "Linking long-term dietary patterns with gut microbial enterotypes.," *Science (New York, N.Y.)*, vol. 334, pp. 105–8, oct 2011.

[52] C. C. Evans, K. J. LePard, J. W. Kwak, M. C. Stancukas, S. Laskowski, J. Dougherty, L. Moulton, A. Glawe, Y. Wang, V. Leone, D. a. Antonopoulos, D. Smith, E. B. Chang, and M. J. Ciancio, "Exercise prevents weight gain and alters the gut microbiota in a mouse model of high fat diet-induced obesity," *PLoS ONE*, vol. 9, no. 3, 2014.

[53] R. M. Voigt, C. B. Forsyth, S. J. Green, E. Mutlu, P. Engen, M. H. Vitaterna, F. W. Turek, and A. Keshavarzian, "Circadian disorganization alters intestinal microbiota," *PloS one*, vol. 9, no. 5, p. e97500, 2014.

[54] K. P. Scott, S. W. Gratz, P. O. Sheridan, H. J. Flint, and S. H. Duncan, "The influence of diet on the gut microbiota.," *Pharmacological research : the official journal of the Italian Pharmacological Society*, vol. 69, pp. 52–60, mar 2013.

[55] M. J. Claesson, I. B. Jeffery, S. Conde, S. E. Power, E. M. O'Connor, S. Cusack, H. M. B. Harris, M. Coakley, B. Lakshminarayanan, O. O'Sullivan, G. F. Fitzgerald, J. Deane, M. O'Connor, N. Harnedy, K. O'Connor, D. O'Mahony, D. van Sinderen, M. Wallace, L. Brennan, C. Stanton, J. R. Marchesi, A. P. Fitzgerald, F. Shanahan, C. Hill, R. P. Ross, and P. W. O'Toole, "Gut microbiota composition correlates with diet and health in the elderly.," *Nature*, vol. 488, pp. 178–84, aug 2012.

[56] C. De Filippo, D. Cavalieri, M. Di Paola, M. Ramazzotti, J. B. Poullet, S. Massart, S. Collini, G. Pieraccini, and P. Lionetti, "Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 33, pp. 14691–6, 2010.

[57] H. J. Flint, K. P. Scott, P. Louis, and S. H. Duncan, "The role of the gut microbiota in nutrition and health," *Nature Reviews Gastroenterology and Hepatology*, vol. 9, no. 10, pp. 577–589, 2012.

[58] D. Graf, R. Di Cagno, F. Fåk, H. J. Flint, M. Nyman, M. Saarela, and B. Watzl, "Contribution of diet to the composition of the human gut microbiota," *Microbial ecology in health and disease*, vol. 26, 2015.

[59] A. Suau, R. Bonnet, M. Sutren, J.-J. Godon, G. R. Gibson, M. D. Collins, and J. Doré, "Direct analysis of genes encoding 16s rrna from complex communities reveals many novel molecular species within the human gut," *Applied and environmental microbiology*, vol. 65, no. 11, pp. 4799–4807, 1999.

[60] C. Sala, S. Vitali, E. Giampieri, Ì. F. do Valle, D. Remondini, P. Garagnani, M. Bersanelli, E. Mosca, L. Milanesi, and G. Castellani, "Stochastic neutral modelling of the gut microbiota's relative species abundance from next generation sequencing data," *BMC Bioinformatics*, vol. 17, no. 2, p. 179, 2016.

[61] I. B. Jeffery, D. B. Lynch, and P. W. O'Toole, "Composition and temporal stability of the gut microbiota in older persons," *The ISME Journal*, pp. 1–13, 2015.

[62] C. V. Granger, B. B. Hamilton, R. A. Keith, M. Zielezny, and F. S. Sherwin, "Advances in functional assessment for medical rehabilitation.," *Topics in geriatric rehabilitation*, vol. 1, no. 3, pp. 59–74, 1986.

[63] C. D. Wolfe, N. A. Taub, E. Woodrow, and P. Burney, "Assessment of scales of disability and handicap for stroke patients.," *Stroke*, vol. 22, no. 10, pp. 1242–1244, 1991.

[64] V. C. Pangman, J. Sloan, and L. Guse, "An examination of psychometric properties of the mini-mental state examination and the standardized mini-mental state examination: implications for clinical practice," *Applied Nursing Research*, vol. 13, no. 4, pp. 209–213, 2000.

[65] A. Gordon and G. Hannon, "Fastx-toolkit," 2010.

[66] B. Ewing, L. Hillier, M. C. Wendl, and P. Green, "Base-calling of automated sequencer traces usingphred. i. accuracy assessment," *Genome research*, vol. 8, no. 3, pp. 175–185, 1998.

[67] P. D. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, *et al.*, "Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities," *Applied and environmental microbiology*, vol. 75, no. 23, pp. 7537–7541, 2009.

[68] R. C. Edgar, "UPARSE: highly accurate OTU sequences from microbial amplicon reads," *Nature Methods*, vol. 10, no. 10, pp. 996–998, 2013.

[69] R. Edgar, "Uparse otu radius," 2013.

[70] R. C. Edgar, "Search and clustering orders of magnitude faster than BLAST.," *Bioinformatics (Oxford, England)*, vol. 26, pp. 2460–1, oct 2010.

[71] J. Tang and S. Zhou, "Hybrid niche-neutral models outperform an otherwise equivalent neutral model for fitting coral reef data.," *Journal of theoretical biology*, vol. 317, pp. 212–8, jan 2013.

[72] K. P. Burnham, D. R. Anderson, and K. P. Huyvaert, "Aic model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons," *Behavioral Ecology and Sociobiology*, vol. 65, no. 1, pp. 23–35, 2011.

[73] E. Biagi, L. Nylund, M. Candela, R. Ostan, L. Bucci, E. Pini, J. Nikkïla, D. Monti, R. Satokari, C. Franceschi, *et al.*, "Correction: Through ageing, and beyond: Gut microbiota and inflammatory status in seniors and centenarians," *PLoS ONE*, vol. 5, no. 6, 2010.

[74] M. J. Claesson, S. Cusack, O. O'Sullivan, R. Greene-Diniz, H. de Weerd, E. Flannery, J. R. Marchesi, D. Falush, T. Dinan, G. Fitzgerald, *et al.*, "Composition, variability, and temporal stability of the intestinal microbiota of the elderly," *Proceedings of the National Academy of Sciences*, vol. 108, no. Supplement 1, pp. 4586–4591, 2011.

[75] D. P. Faith, "Conservation evaluation and phylogenetic diversity," *Biological conservation*, vol. 61, no. 1, pp. 1–10, 1992.

[76] R. I. Amann, W. Ludwig, and K.-H. Schleifer, "Phylogenetic identification and in situ detection of individual microbial cells without cultivation.," *Microbiological reviews*, vol. 59, no. 1, pp. 143–169, 1995.

[77] P. Hugenholtz, B. M. Goebel, and N. R. Pace, "Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity," *Journal of bacteriology*, vol. 180, no. 18, pp. 4765–4774, 1998.

[78] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.

[79] R. A. d'Errico M, Facco E, Laio A, "Automatic topography of complex data sets by accurate density estimation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 51, pp. 20167–20172, 2016.

[80] P. D. Schloss and S. L. Westcott, "Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis.," *Applied and environmental microbiology*, vol. 77, pp. 3219–26, may 2011.

[81] W. Chen, C. K. Zhang, Y. Cheng, S. Zhang, and H. Zhao, "A comparison of methods for clustering 16s rrna sequences into otus," *PloS one*, vol. 8, no. 8, p. e70837, 2013.

[82] W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, 2006.

[83] R. C. Edgar, "Search and clustering orders of magnitude faster than BLAST.," *Bioinformatics (Oxford, England)*, vol. 26, pp. 2460–1, oct 2010.

[84] X. Hao, R. Jiang, and T. Chen, "Clustering 16s rrna for otu prediction: a method of unsupervised bayesian clustering," *Bioinformatics*, vol. 27, no. 5, pp. 611–618, 2011.

[85] E. Pruesse, C. Quast, K. Knittel, B. M. Fuchs, W. Ludwig, J. Peplies, and F. O. Glöckner, "Silva: a comprehensive online resource for quality checked and aligned ribosomal rna sequence data compatible with arb," *Nucleic acids research*, vol. 35, no. 21, pp. 7188–7196, 2007.

[86] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of molecular biology*, vol. 48, no. 3, pp. 443–453, 1970.

[87] M. L. Sogin, H. G. Morrison, J. A. Huber, D. M. Welch, S. M. Huse, P. R. Neal, J. M. Arrieta, and G. J. Herndl, "Microbial diversity in the deep sea and the underexplored "rare biosphere"," *Proceedings of the National Academy of Sciences*, vol. 103, no. 32, pp. 12115–12120, 2006.

[88] B. W. Silverman, *Density estimation for statistics and data analysis*, vol. 26. CRC press, 1986.

[89] C. Ceruti, S. Bassis, A. Rozza, G. Lombardi, E. Casiraghi, and P. Campadelli, "Danco: An intrinsic dimensionality estimator exploiting angle and norm concentration," *Pattern recognition*, vol. 47, no. 8, pp. 2569–2581, 2014.

[90] J. W. Tukey, "Bias and confidence in not-quite large samples," in *Annals of Mathematical Statistics*, vol. 29, pp. 614–614, 1958.

[91] J. W. Tukey, "Exploratory data analysis," 1977.

[92] M. Waskom, "Seaborn: statistical data visualization," 2012.

[93] F. E. Angly, D. Willner, F. Rohwer, P. Hugenholtz, and G. W. Tyson, "Grinder: a versatile amplicon and shotgun sequence simulator," *Nucleic acids research*, vol. 40, no. 12, pp. e94–e94, 2012.

[94] S. Balzer, K. Malde, A. Lanzén, A. Sharma, and I. Jonassen, "Characteristics of 454 pyrosequencing data—enabling realistic simulation with flowsim," *Bioinformatics*, vol. 26, no. 18, pp. i420–i425, 2010.

[95] A. Gilles, E. Meglécz, N. Pech, S. Ferreira, T. Malausa, and J.-F. Martin, "Accuracy and quality assessment of 454 gs-flx titanium pyrosequencing," *BMC genomics*, vol. 12, no. 1, p. 245, 2011.

[96] R. C. Edgar, B. J. Haas, J. C. Clemente, C. Quince, and R. Knight, "Uchime improves sensitivity and speed of chimera detection," *Bioinformatics*, vol. 27, no. 16, pp. 2194–2200, 2011.

[97] P. A. Estévez, M. Tesmer, C. A. Perez, and J. M. Zurada, "Normalized mutual information feature selection," *IEEE Transactions on Neural Networks*, vol. 20, no. 2, pp. 189–201, 2009.

[98] B. Alberts, A. Johnson, J. Lewis, D. Morgan, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell*. Garland Science, 6th ed., 2015.

[99] C. Chothia, J. Gough, C. Vogel, and S. A. Teichmann, "Evolution of the protein repertoire.," *Science*, vol. 300, no. 5626, pp. 1701–3, 2003.

[100] M. A. Huynen and E. van Nimwegen, "The frequency distribution of gene family sizes in complete genomes," *Molecular Biology and Evolution*, vol. 15, no. 5, pp. 583–589, 1998.

[101] J. Qian, N. M. Luscombe, and M. Gerstein, "Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model," *Journal of molecular biology*, vol. 313, no. 4, pp. 673–681, 2001.

[102] S. Wuchty, "Scale-free behavior in protein domain networks," *Molecular biology and evolution*, vol. 18, no. 9, pp. 1694–1702, 2001.

[103] A.-L. Barabási, *Linked: the new science of networks science of networks*. Basic Books, 2002.

[104] V. A. Kuznetsov, "Distribution associated with stochastic processes of gene expression in a single eukariotic cell," *EURASIP Jouornal of applied signal processing*, vol. 4, pp. 285–296, 2001.

[105] V. A. Kuznetsov, "Statistics of the numbers of transcripts and protein sequences encoded in the genome," in *Computational and Statistical Approaches to Genomics*, pp. 125–171, Springer, 2002.

[106] E. V. E. Koonin, Y. Y. I. Wolf, and G. G. P. Karev, "The structure of the protein universe and genome evolution," *Nature*, vol. 420, no. November, pp. 218–23, 2002.

[107] G. P. Karev, Y. I. Wolf, A. Y. Rzhetsky, F. S. Berezovskaya, and E. V. Koonin, "Birth and death of protein domains: a simple model of evolution explains power law behavior.," *BMC evolutionary biology*, vol. 2, p. 18, 2002.

[108] M. Pagel, "Inferring the historical patterns of biological evolution," *Nature*, vol. 401, no. 6756, pp. 877–884, 1999.

[109] C. Furusawa, T. Suzuki, A. Kashiwagi, T. Yomo, and K. Kaneko, "Ubiquity of Log-normal Distributions in Intra-cellular Reaction Dynamic," *Biophysics*, vol. 1, p. 15, 2005.

[110] D. M. Easton and H. R. Hirsch, "For prediction of elder survival by a gompertz model, number dead is preferable to number alive," *Age*, vol. 30, no. 4, p. 311, 2008.

[111] D. L. Wheeler, C. Chappey, A. E. Lash, D. D. Leipe, T. L. Madden, G. D. Schuler, T. A. Tatusova, and B. A. Rapp, "Database resources of the national center for biotechnology information," *Nucleic acids research*, vol. 28, no. 1, pp. 10–14, 2000.

[112] J. J. Koehorst, J. C. van Dam, E. Saccenti, V. A. Martins dos Santos, and P. J. Schaap, "Sapp: Semantic annotation platform for prokaryotes," 2015.

[113] D. Hyatt, G.-L. Chen, P. F. LoCascio, M. L. Land, F. W. Larimer, and L. J. Hauser, "Prodigal: prokaryotic gene recognition and translation initiation site identification," *BMC bioinformatics*, vol. 11, no. 1, p. 1, 2010.

[114] P. Jones, D. Binns, H.-Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell, G. Nuka, *et al.*, "Interproscan 5: genome-scale protein function classification," *Bioinformatics*, vol. 30, no. 9, pp. 1236–1240, 2014.

[115] M. D. M. Paulo I. Prado and A. Chalom, "sads: Maximum likelihood models for species abundance distributions," 2016.

[116] M. Bulmer, "On fitting the poisson lognormal distribution to species-abundance data," *International Biometric Society*, vol. 30, no. 1, pp. 101–110, 1974.

[117] J. H. Ward Jr, "Hierarchical grouping to optimize an objective function," *Journal of the American statistical association*, vol. 58, no. 301, pp. 236–244, 1963.

[118] C. Quast, E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glöckner, "The silva ribosomal rna gene database project: improved data processing and web-based tools," *Nucleic acids research*, vol. 41, no. D1, pp. D590–D596, 2013.