SISSA

Scuola
Internazionale
Superiore di
Studi Avanzati

Physics area - PhD course in
Theory and Numerical Simulations
of Condensed Matter

Stochastic ecology and evolution in bacterial communities

Candidate:

Lorenzo Fant

Advisor:

Jacopo Grilli

Academic year 2020-2021

MA PER SEGUIR VIRTUTE E CANOSCENZA • SISSA •

Contents

Introduction

Ecology is the study of the interrelationships of living organisms with their physical environment and each other. It considers organisms at individual, community, ecosystems and biosphere levels. It studies different topics:

- The development, both of ecosystems and of adapting individuals
- The movement of materials and energy through communities
- Interactions among species as cooperation and competition
- The abundance and distribution of species among different environments.

A very fertile subject is the one of bacterial communities, providing huge amounts of data that allow for the application of quantitative methods, statistical analysis and mathematical modeling.

Trying to make an estimate of the amount of life that belongs to each kingdom (animals, plants, bacteria, etc) bacteria turn out to be way more important than we would predict by everyday experience. They make up for 35 times the total weight of animals on the planet (for a very nice and surprising representation of life on earth look [60]) and outnumber human cells inside our own bodies by a factor ten [54].

Bacteria appear simpler then other organisms. While many animals and plants are composed by up to 10^{14} cells, bacteria are unicellular. Moreover, they have no nucleus to contain their DNA and organelles, usually keeping their entire genetic information in a single loop of DNA.

Nevertheless, if we look at their evolutionary success we realize that simple does not imply uneffective.

On the contrary, bacteria are among the most adaptable kingdoms, as they are found in every habitat on Earth: from soils, rocks and oceans to very extreme ones as arctic ices an underwater volcanoes. Some live in or on other organisms, plants and animals, including humans.

While some bacteria can cause diseases in animals or plants, most are harmless and are beneficial ecological agents whose metabolic activities sustain higher life-forms. Other bacteria are symbionts of plants and animals, where they carry out important functions for the host, such as nitrogen fixation and cellulose degradation. Without them, soil would not be fertile, and dead organic material would decay much more slowly [8, 1].

We wrote about bacteria as if they were always identifiable as single objects and, spatially, they surely are, being well defined by a membrane that discriminates between what is inside and what outside. But, considering them from an ecological perspective, the image often becomes more complex. In nature, bacteria are found in many different conditions but, usually, they are not found alone. Individuals are often just a component of much more important and diversified system: communities.

Environmental samples reveal ensembles of hundreds or thousands of different bacterial species, living together and mutually interacting. They communicate, harm one another and exchange resources, drastically modifying the environment. Interactions can act at small cellular distances or between individuals kilometers away. Along our intestinal tract there are species interacting at distances of several meters while two low mobility bacteria on two sides of a grain of sand could spend their whole existence without noticing one another. Moreover, small microscopic interactions can have huge macroscopic impacts. For example, the action of photosynthetic cyanobacteria led to the oxygenation of our atmosphere and enabled the evolution of macroscopic organisms, including animals [45, 55].

To better understand how little we know about microbial interaction we can look at one of the main experimental procedures to investigate bacteria: physical isolation and culture. Since its invention in 1860 by Pasteur this has been the golden standard to study bacteria, and still is in many cases. It aims at isolating one bacterial species from the whole community by diluting it on a resource rich surface. If diluted enough there can remain up to just 1 individual that, growing, generates a colony of bacteria all belonging to the same species. Experimentalists can then see a unique and defined morphology of the colony and study its biology and, lately, read its DNA by sequencing the genome. But, looking at the amount of bacteria that have been isolated up to now we get an astonishing result. Just between 2% and 50%, depending on the origin of the sample, of species are considered culturable in a laboratory [61], and these percentages, already small, are decreasing, as we discover new species faster than we isolate them.

On the other hand, bacterial physiology changes upon external conditions. What we measure in the lab, placing them alone on a resource rich surface, can be dramatically different from what is happening in nature, where they are interacting with hundreds of neighbours.

It becomes then clear the importance of being able to study the community as a whole, keeping the setting as it is in nature and focusing on the interplay between external conditions and internal interactions. From this perspective - many interacting objects with properties emerging from collective behaviours - it appears clear the growing interest of physics community towards ecology.

Along the thesis I will focus on two different aspects of community dynamics: the characterisation of the communities stationary state and the effects of interactions on their growth.

To introduce the first it is necessary to define two possible descriptions of a bacterial community: species and functions. The first one is intuitive: determining how many individuals of each species are present in a community. The second is trickier. Bacteria, but life in general, mainly uses proteins to accomplish tasks. Proteins are used for structural and mechanics purposes, to transport molecules and as enzymes to catalyse chemical reactions. They are produced by the cells following the instructions contained in the DNA. Each DNA section encoding the information to produce one protein is a gene.

We can interpret the genome (a filament of DNA in the case of bacteria, the set of chromosomes for humans) as a set of instructions to build tools and associate each set of tools to a function performed by the organism. By abstracting, an organism can be identified by the collection of functions it is able to perform. Similarly, communities can be characterised by the set of functions performed by their individuals. For example, analysing a savanna we could both determine how many antelopes, giraffes and lions there are or how much grass can be eaten per year by all the animals together, independently from the single species contribution to the "eating grass function".

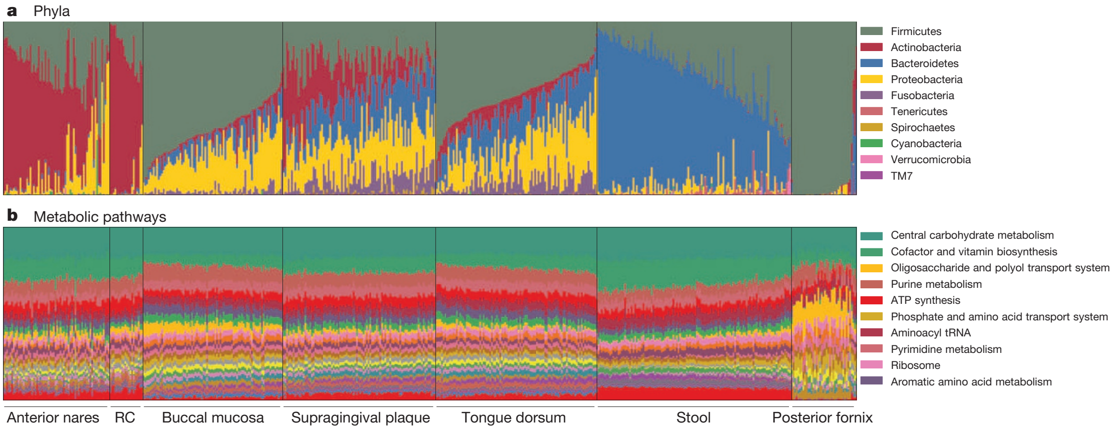


Figure 1: Figure taken from [15] Microbial phyla composition (phyla are one taxonomic category, broader that species) varies while metabolic pathways (functions) remain stable within healthy hosts. Vertical bars represent microbiome samples by body habitat in the seven locations. Bars indicate relative abundances colored by microbial phyla (**a**) and metabolic modules (**b**). A plurality of most communities' memberships consists of a single dominant phylum, but this is universal neither to all body habitats nor to all individuals. Conversely, most metabolic pathways are evenly distributed and prevalent across both individuals and body habitats.

As shown in Fig.1, from the analysis of bacterial communities [15] arise a remarkable behaviour: communities living in similar environments show a stable functional composition opposed to an high variability of species composition. Species vary a lot across samples but overall the functions performed by the community remain constant.

In Chapter 1 and 2, we show how such an experimental observation can be explained by mathematical models and develop a stationary state description of the community using macroscopic observables. We show that in consumer resources models, a widespread representation of bacterial communities, external resources and interactions among bacteria uniquely determine the functional composition of the community while leaving the species composition free to fluctuate. Eventually we show how the quantities characterising the stationary state are combined in a closed set of equations, determining a thermodynamic-like description of communities at equilibrium.

In Chapter 3 we try to deepen our knowledge on functional stability from a data oriented perspective by moving the first steps toward a null model for community composition. The low variability shown in Fig. 1 is, up to now, mainly qualitative. To be able to asses whether what we see is really more stable than expected it is necessary to estimate the functional variability of randomly assembled communities. What are the relevant parameters to consider when assembling a community and which determine a functional variation? We describe the importance of bacteria genome size in a community and develop a new method to obtain an estimate of the genome size distribution from the functional composition data.

The last Chapter (4) deals with a problem quite different from the previous ones, looking at stochastic exponential processes and cooperation.

Stochastic exponential processes are good mathematical descriptions of quantities growing by a multiplicative factor. They represent, among others, ecological processes, as population dynamics, and economical ones, like stock market dynamics. Therefore, while the previous chapters investigate stationary states, we here look for describing the first, exponentially growing, phases of expansion of ecological communities. Particularly, we consider how cooperation can evolve and be stable in such systems.

Cooperation is a long debated subject since the publication of *Origin of Species* by Darwin, where an explanation of the motor of evolution is found

> I use this term in a large and metaphorical sense including dependence of one being on another, and including (which is more important) not only the life of the individual, but success in leaving progeny. Two canine animals, in a time of dearth, may be truly said to struggle with each other which shall get food and live. But a plant on the edge of a desert is said to struggle for life against the drought.... As the mistletoe is disseminated by birds, its existence depends on birds; and it may metaphorically be said to struggle with other fruit-bearing plants, in order to tempt birds to devour and thus disseminate its seeds rather than those of other plants. In these several senses, which pass into each other, I use for convenience sake the general term of struggle for existence.

The debate around the nature of the “struggle for existence” has been intense. The competitive interpretation, spread by the Darwin’s first disciple, Huxley, and resumed in the “Nature, red in tooth and claw” line from Tennyson, faced the cooperative one, strong in the Russian literature and spread in western countries by the Kropotkin’s book *Mutual Aid.*

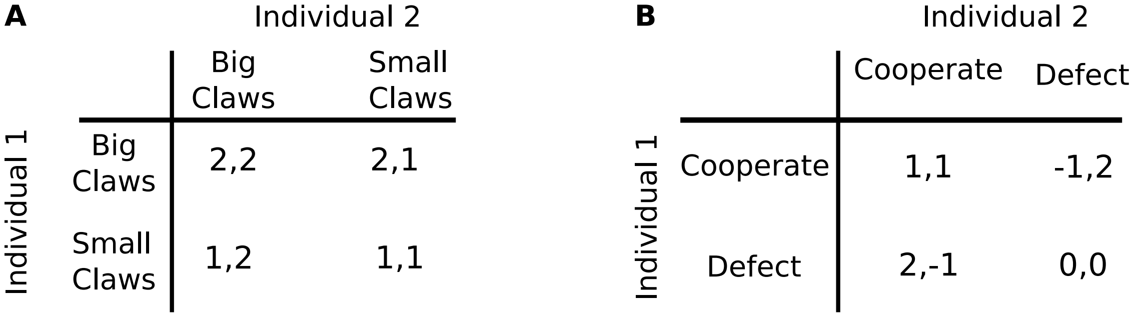
Making a long story short and overcoming the philosophical implications, one of the keys of the discussion lies in the concept of evolutionary stability (if you want the whole story read the very nice article [26]).

Game theory come in helpful to formalise the concept. We can associate each character (of function) expressed by organisms with an expected increase/decrease of reproducing. For example, usually being faster helps to survive.

Imagine life as a game interpreting every possible character as a strategy that organisms can play. In the game, each strategy returns a payoff, depending on its success and individuals reproduce proportionally to the payoff obtained.

Lets consider a simple game , where individuals can adopt one of two mutually exclusive strategies, for example having small or big claws. A a strategy is evolutionary stable if, when present in the population, it cannot be ousted by the competing one.

We can represent the "game" via a payoff matrix. each entry of the matrix gives the payoffs of the two individuals for a given choice of strategy



A

		Individual 2	
		Big Claws	Small Claws
Individual 1	Big Claws	2,2	2,1
	Small Claws	1,2	1,1

B

		Individual 2	
		Cooperate	Defect
Individual 1	Cooperate	1,1	-1,2
	Defect	2,-1	0,0

Figure 2: Two payoff matrices examples. Rows indicate individual 1 strategies while columns individual 2. The entries are the payoff obtained by individual 1, individual 2. **A** the payoff matrix of a simple competitive game with trivial solution with evolutionary stability in "Big claws". **B** The prisoner dilemma payoff matrix. Even though full cooperation is a better setting than full defection for both individuals, the advantage of cheating brings the system towards full defection, a non optimal stable strategy.

In the simple example of Fig.2**A**, having big claws gives an advantage, returning an higher payoff than having small ones. We can see that such an example rapidly brings to the stability of big claws in the population as, if all individuals have small claws, a mutation providing big ones would make an individual reproducing twice as much and transmitting the big claws to the offsprings. On the contrary, in a big claws population, any individual with small claws would give rise to a slow growing genealogy, disappearing in time.

This example shows how a competitive strategy immediately reveals the evolutionary stability. The same procedure is a little less obvious for cooperation.

Cooperating means helping other individuals, spending energies by doing it, with no guaranteed return. The setting, shown in Fig.2**B**, usually states that if two individuals cooperate get an higher payoff that defecting (not cooperating). But, at the same time, the highest payoff is obtained by cheating, defecting while the partner is cooperating. For example if two can choose to share the hunting revenues, by cooperating they have an higher probability to eat meat everyday, while defecting they could get a lot of food one day but starve the next one. Cheating one obtains all the meat he is able to hunt plus half of the partner's.

Such a system, even if for the group cooperation is better than defection, providing to everyone an higher payoff, has a unique evolutionary stable state in defection. In fact in a population of cooperators any defector will have an advantage, growing faster and soon becoming the dominant strategy, while a population of defectors will never be invaded by cooperators that get a lower payoff. In game theory, this type of type of payoff matrix is called *Prisoner Dilemma* where individuals, aware of the payoffs, both end up defecting for the fear of a cheating partner.

Nevertheless, cooperation is observed in nature, alimenting a discussion on the mechanisms that determine its stability, overcoming the above dilemma. Up to now, five different mechanisms have been described [43]. The ingredients required to make cooperation stable span between spatial localization of cooperators to reciprocity, i.e. cooperating just if the partner is doing the same.

In Chapter 4 we show how in stochastic multiplicative environments, where the payoff of one game is a multiple of the payoff of the previous one, cooperation become stable when looking at long term returns. This suggests a 6^{th}, new way, for evolutionary stability of cooperation, opening new interesting possibilities for describing processes and finding new optimal solutions in both ecology and economics.

Chapter 1

Functional Stability and Consumer-Resource Models

This chapter is part of a paper uploaded on the Arxiv [31]

Microbial communities are functionally stable and taxonomically variable: species abundances fluctuate over time and space, while the functional composition is robust and reproducible. These observations imply functional redundancy: the same function is performed by many species, so that one may assemble communities with different species but the same functional composition. The clarity of this observation does not parallel with a theoretical understanding of its origin. Here we study the eco-evolutionary dynamics of communities interacting through competition and cross-feeding. We show that the eco-evolutionary trajectories rapidly converge to a "functional attractor", characterized by a functional composition uniquely determined by environmental conditions. The taxonomic composition instead follows non-reproducible dynamics, constrained by the conservation of the functional composition. Our framework provides a deep theoretical foundation to the empirical observations of functional robustness and redundancy.

1.1 Introduction

The staggering taxonomic diversity of microbial communities parallels with their remarkable functional robustness [10, 35]. At the species and strain level, their taxonomic composition is highly variable across communities with similar environmental conditions and over time. This variability is also observed in microcosmos experiments, under very controlled conditions [24]. On the other hand, the functional composition of communities, estimated for instance using metagenomic data [34, 35], appears to be highly reproducible and stable over time. This — only apparent — contradiction strongly suggests that microbial taxa are highly functionally redundant: since many species can perform the same functions, there exist multiple species combinations corresponding to the same functional profile.

While the replicability of the functional community composition is robust and observed across ecosystems, including laboratory experiments with controlled conditions

[24], its origin is unknown. This lack of understanding is, in part, because our theoretical understanding of ecological models focuses on species composition. Population abundances are the standard degrees of freedom of mathematical models of community dynamics.

Consumer-resource models are the main modeling framework for microbial communities. Their origin goes back to the classic work of MacArthur and Levins [36], which has been extensively studied and discussed in the following decades [59, 14], mostly to describe the coexistence of a handful of species. Recently, these models have been further extended to consider facilitation through cross-feeding [24, 38, 11], where species change resource availability not only by consumption, but also because they release in the environment the waste products of their metabolism. These models qualitatively describe experimental results [24, 17] and have the flexibility to reproduce patterns observed in empirical microbial communities [37].

Once the parameters of the model are set and an initial pool of species is chosen, populations converge for large times to an equilibrium point. Under some mild conditions, identified over decades of theoretical work [13, 29], consumer-resource models are characterized by a globally stable equilibrium: the steady state is independent of the initial population abundances and resource concentrations. The competitive exclusion principle — one of the most fundamental results of theoretical ecology — limits the number of species that can coexist in a stable equilibrium: diversity cannot exceed the number of resources. While this bound is hard, it is often not realized, as only fewer species can coexist [53, 16].

The number and identity of the species coexisting at equilibrium is in fact determined not only by the ecological dynamics, but also by the initial pool of species. This initial pool of species is often interpreted as the metacommunity diversity: the ecological dynamics unfolds in a local community which is coupled to the metacommunity by rare migrations. Most of the recent progresses in understanding the assembly of large ecological communities have been driven by the assumption of “random” species pools [9, 4, 16]. This choice assumes that the parameters characterizing species’ physiological and ecological parameters are independently drawn from some distribution. This assumption implicitly underlies a separation of spatial and temporal scales: the ecological dynamics determining the community composition in the local community occurs independently of the evolutionary processes determining the pool of diversity of the metacommunity.

Instead of assuming a fixed species pool, one can let evolve dynamically individual traits, including the ones specifying their interactions with other individuals and the environment. Classic work in adaptive dynamics [22] has shown how, starting from a clonal population, diversification can evolve under general conditions on frequency-dependent selection. Several works have then studied eco-evolutionary dynamics of interacting populations [3, 18], by allowing individuals’ traits to be subject to mutations and be inherited by the following generation. “Intrinsic” fitness, how fast populations grow in an optimal environment, and niche differences, how the growth of different populations is coupled, both influence community evolution and is their interplay to determine the observed diversity of an evolved community [25].

A key difficulty in interpreting the outcomes of eco-evolutionary dynamics is the fact that there are no natural degrees of freedom to characterize the evolution of the community. The identity of populations, and not only their abundance, is under constant change.

Here we show that the functional composition emerges as the natural variable that characterizes the composition of the community. We consider the broad framework of consumer-resource-crossfeeding models under an explicit eco-evolutionary dynamics, where strains differ in their resources preference and their intrinsic fitness. Higher resource intakes are balanced by a lower efficiency (or equivalently, higher mortality) implemented by a metabolic trade-off [58, 48]. We show that the evolutionary dynamics converge rapidly to a stationary and reproducible functional composition — here defined as the fraction of individuals able to grow on a given resource — which we analytically predict. Interestingly, we show that, once the functional attractor is reached, the strain dynamics is then dominated by fitness differences, implying that functional composition is robust (independent of small fitness differences) and redundant (is obtained under multiple strain compositions).

1.2 Results

The ecological dynamics is defined by standard consumer-resource-crossfeeding equations [38]. In our framework, individuals are characterized by a resource preference vector that determines the intake rate of each of the R resources available in the environment (relative to a maximum). An individual with preference $a_i = 0$ will not consume resource i, while an individual with preference $a_i = 1$ will consume it with a maximum intake rate ν_i [58]. Consumed resources are converted into biomass with finite efficiency (equivalent to an inverse yield). We assume that the yield (or equivalently the death rate, see Materials and Methods) depends linearly on the number of resources consumed: the more resources an individual can grow on, the less efficiently it grows.

Two populations with identical resource preferences can differ in the values of other physiological parameters (e.g., efficiency or mortality). Such differences, which we refer to as intrinsic fitness, determine which population of the two survives when competing. We will use the word 'strain' to identify a group of individuals with equal resource preferences and intrinsic fitness.

Resource dynamics are described explicitly. Resources are introduced in the system with a resource-specific rate h_i and consumed by the individuals present in the community. Their concentrations decrease because of consumption but also vary due to cross-feeding. A fraction $1 - \ell$ of resources consumed by each individual is used for growth, while a fraction ℓ is transformed into different resources and released again in the environment [38]. The cross-feeding matrix, with elements D_{ij}, specifies the relative rates of resource transformation (see Materials and Methods for its parameterization). The per-capita growth rate of a strain μ is a function $g_\mu(\underline{c})$ of the resource concentration $\underline{c}$, which, in turn, depends dynamically on the population abundance because of consumption and cross-feeding. We consider the following choice

$$g_\mu(\underline{c}) = \eta_\mu \left((1-\ell) \sum_{i=1}^{R} a_{\mu i} r_i(c_i) - \frac{1}{\tau} \left(1 + \chi \sum_{i=1}^{R} a_{\mu i} \right) (1-\epsilon_\mu) \right) \ . \tag{1.1}$$

The values of η_μ and τ are arbitrary and their choice does not affect the results. The functional form $r_i(c_i)$ encodes the functional response. Both linear and saturating (Monod-like) functional responses produce the same results (see Fig.A.4). The parameter χ quantifies the fitness cost of consuming one resource and implements the metabolic trade-off. The form of the trade-off generalizes the case considered in [58, 48], which assumes a constant total energy budget devoted to metabolism (i.e. a constant value of $\sum_j a_{\sigma j}$). In our case, the total metabolic energy budget is not fixed to a constant but allowed to vary. In the Materials and Methods, we show that the fixed energy budget scenario [58, 48] corresponds to the limit of large values of χ. Including a constant term in the metabolic cost (set equal to 1 in our framework without loss of generality) is related to a basal cost, related to housekeeping functions. The quantity ϵ_μ determines the intrinsic fitness value.

In our framework, both resource preferences and intrinsic fitness values are subject to mutations and evolution. We consider different implementations of the mutational steps (e.g., including different scenarios for the relative rate of Horizontal Gene Tranfer, see Materials and Methods) which, however, do not affect the results. The timescales between two successive successful mutations is comparable with the ecological timescale, set by the ecological dynamics. We assume that a mutation of the resource preference always corresponds to a mutation of the intrinsic fitness, which is drawn at random from a fixed distribution with width ϵ. The parameter ϵ sets the typical difference of intrinsic fitness values between two individuals. We focus on the case of small fitness differences, and we extensively explore the effect on the eco-evolutionary trajectories of increasing the value of ϵ.

Fig 1.1 shows a sample of eco-evolutionary trajectories resulting from our framework. Starting from a clonal population, a diverse community is rapidly assembled. Strain abundances change abruptly following successful invasion events and keep changing over the whole duration of the simulations.

The final community structure is remarkably simple if, instead of analyzing strain abundances, we focus on its functional composition. We define functional occurrence F_i as the fraction of individuals able to grow on i (i.e., with $a_i = 1$). After a short transient, the functional occurrences and the total biomass N relax to the respective stationary values F_i^* and N^*, which are very reproducible across different realizations.

Two phases characterize therefore the eco-evolutionary dynamics. The first one is an initial-condition-dependent transient, where the community structure is mainly shaped by rapid invasions. In the second phase, conversely, the community has converged to a stable functional composition, which we will refer to as "functional attractor" in the following, and slowly evolves reaching the final strain-level equilibrium.

The total biomass converges to a constant value during the second phase of the eco-evolutionary dynamics, which affects therefore only the relative abundance of strains.

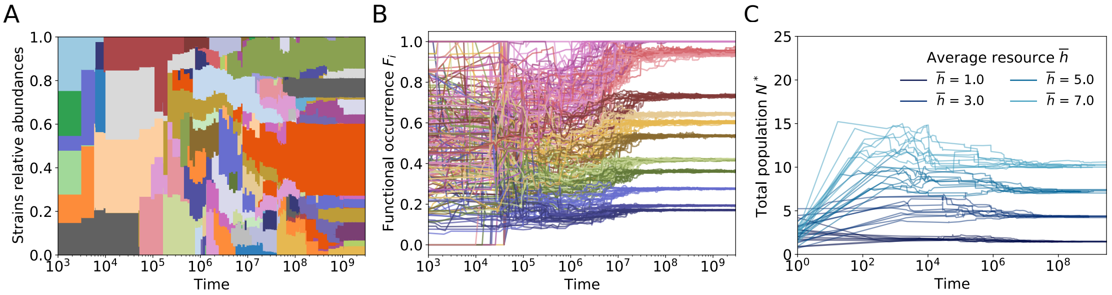


Figure 1.1: Stability of functional occurrences F_i for communities evolving under a consumer-resource model. The system is initialized with a small number of initial random strains, chosen so that each gene is present at least once. The system evolves in a chemostat with fixed resources input. When equilibrium is reached, one mutant is added to the batch. The chemostat then equilibrates to a new fixed point and the procedure is repeated until function and biomass reach stability. **A**: Time evolution of the strains relative abundances for one realization of the system. **B**: Time evolution of functional occurrences for three different realizations of the system. 15 resources are given. **C**: Time evolution of the total biomass of the system. 20 realizations of the system are shown for each value of average resources income.

The sequence of invasions and extinctions of strains is determined by the interplay of fitness differences ϵ_μ and niche differences, related to the dissimilarity of the resource preferences. Importantly, the trajectories of strain abundances are effectively restrained to occur on the low(er)-dimensional space determined by the constraint enforced through the functional occurrences F_i^*. In this second phase the community has thus reached a "functional maturity" and the subsequent evolution only affects strain composition while leaving unaltered the functional one.

The stability and reproducibility of the functional attractor suggest that it is possible to predict analytically its properties. We considered a toy model of the eco-evolutionary dynamics which aims at mimicking the effective exploration of the phenotypic space performed by mutations. In particular, we consider only the ecological dynamics, initialized with an infinitely large species pool, which encompasses all the possible strains (i.e., the 2^R possible resource preferences). A similar approach has been considered to study a simpler version of the model [58] (corresponding to the limit $\chi \gg 1$ and no cross-feeding). The toy model further postulates a timescale separation between resource and population dynamics [58, 48], which is not assumed in the full eco-evolutionary dynamics.

The consumer-resource-crossfeeding model with infinite pool of diversity and no intrinsic fitness differences can be analytically solved. In Appendix A.2 we show that the stationary functional occurrences F_i^* and the total biomass N^* are given by

$$F_i^* = \min\{\frac{h_i^{eff}}{\chi}\frac{1}{N^*}, 1\} \tag{1.2}$$

and

$$N^* = \frac{\sum_i h_i^{eff}}{R(1+\chi \sum_i F_i^*)} \; . \tag{1.3}$$

The parameter h_i^{eff} is the effective resource inflow in the system, which is given by the combination of resources that are externally supplied and the ones produced via cross-feeding. This quantity is in simple linear relation to the inflow rate of externally provided resource h_i through the cross-feeding matrix D (see Materials and Methods).

The analytical calculations are based on many simplifying assumptions (infinitely large pool of diversity, not explicit resource dynamics, absence of fitness differences) which do not strictly hold for the more complex setting of the eco-evolutionary model. Nevertheless, Figure 1.2 shows that the predictions of eq. 1.2 and eq. 1.3 accurately describe the outcomes of the eco-evolutionary dynamics.

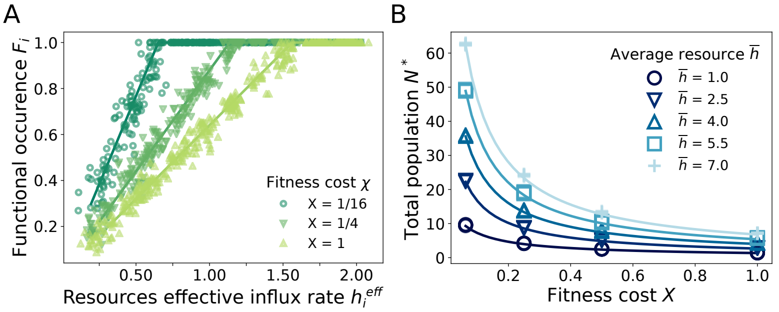


Figure 1.2: The theoretical predictions given by eq. A.10, A.11 (solid lines) are reproduced by numerical integration of eq. A.1, A.2 (markers). **A**: Occurrence of the phenotypes F_i as a function of resource income rates h_i. According to equation A.10 the most abundant resources (core) are consumed by all strains (F_i=1) while the remaining ones only by a fraction $F_i = \frac{h_i}{N^*\chi}$. Notice that increasing χ reflects in a decrease of the number of core resources. **B**: Dependence of the total equilibrium population N^* (biomass) on the value of χ. Here we find a dependence on the average value of the resources incomes $\overline{h}$, which is absent for quantities in panel **A**. In each figure are represented 20 different noise realizations solutions of the system for each χ (**A** and each $\overline{h}$ (**B**)

Resources can be partitioned in two groups based on their effective influx rate h_i^{eff}. If the influx rate is larger than a critical value h_c^{eff}, then the ability to metabolize that resource is a "core" function, shared by all the individuals in the community (i.e., $F_i^* = 1$). The value of h_c^{eff} depends on both the spread of the effective influx rate (the variability among the h_i) and the metabolic cost χ. The higher the metabolic cost and the variability, the higher the critical influx rate threshold h_c^{eff} and, consequently, the fewer the core resources.

At equilibrium, the resources with an influx rate below the critical threshold (i.e., the non-'core' resources) are consumed only by a fraction of the individuals. A linear relation links the functional occurrence F_i^* with the effective resource influx rate h_i^{eff}. The slope of this relation is simply linked to the metabolic cost and the total biomass, being equal to $(\chi N^*)^{-1}$ (see Materials and Methods). Combining equation 1.2 and 1.3, one can obtain an explicit expression for N^*. Figure 1.1B shows that the analytical expression for N^* (as a function of the metabolic cost χ) correctly matches the observations of the eco-evolutionary dynamics.

An emerging feature of the present framework is that the functional composition of communities is extremely robust to fitness differences. We further explore this aspect by considering community response to variation in intrinsic finesses. This variation mimics the temporal or spatial heterogeneity of environmental factors that influence growth, such as abiotic factors (temperature, pH, salinity, etc.) or phages with different host ranges.

We consider two complementary scenarios, which aim at exploring cross-sectional (across communities) and longitudinal (over time) variation. In the former case, we compare the eco-evolutionary outcomes of several communities that share the same resource input but have independent intrinsic finesses. Two individuals with the same resource preference will have uncorrelated intrinsic fitnesses in two different communities. The latter case assumes instead that intrinsic fitness fluctuates over time with a typical autocorrelation timescale (see Materials and Methods). Over time ranges shorter than the autocorrelation timescale, intrinsic fitness is approximately constant. Over times larger than the autocorrelation timescales, the intrinsic fitness decorrelates and becomes an independent variable.

Figure 1.3 shows the strain and functional composition of communities in the two scenarios described above. The strain composition strongly differs across communities or overtime, being highly sensitive to small intrinsic fitness differences. On the other hand, the functional profile is left largely unaffected by fitness variation. These observations clearly show that functional redundancy naturally emerges in complex consumer-resource-crossfeeding models, closely reproducing the phenomenology observed in microbial communities [35].

1.3 Discussion

Our results shed light on the composition of large ecological communities. When the pool of diversity is not a-priori constrained but is instead allowed to evolve, the complex ecological dynamics can be decomposed in a fast, predictable, phase and a slow one, contingent on the (small, yet relevant) fitness differences. The community composition rapidly converges to a set of solutions, determined by resource availability. The following dynamics are constrained on that subspace of solutions and is governed by the difference in relative fitness. Remarkably, this separation of fast and slow components directly map into functional and taxonomic composition: the former is robust and governed only by effective resource influx rates, the latter is constrained by function, but free to move along functionally equivalent directions.

The functional robustness and functional redundancy are the direct consequences of the existence of the two dynamics phases that directly map onto taxonomic and func-

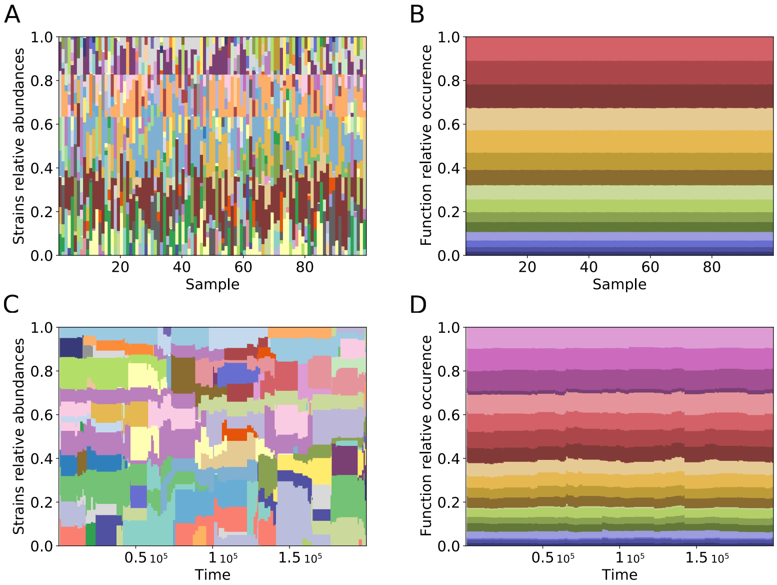


Figure 1.3: Fitness differences allow to demonstrate how functional stability is encoded within the model. While the populations/taxonomies becomes highly variable and heterogeneous, the functional composition is preserved and unaffected by stochasticity. **A**,**B** show a collection of equilibrium configurations of systems with different realisations of a static noise. **C**,**D** show the evolution of the composition of a system with a dynamically varying noise .

tional variation. Functional robustness, the observation that functional composition is stable over time and across communities, originates from the existence of a functional attractor of the eco-evolutionary dynamics. While intrinsic fitness differences are small, they are not negligible as they determine the taxonomic composition within the functional attractor. Variation of the intrinsic fitness leads to functional redundancy, high taxonomic variability with conserved functional profile.

The assumption of small intrinsic fitness differences is critical for the observations of functional robustness and redundancy. Increasing the magnitude of fitness differences affects also the functional profile. Typical intrinsic fitness differences of 1% do not alter substantially the functional composition (see Figure A.1). Larger differences (of the order of 10%) disrupt the structure of the functional attractor: the functional composition is determined by the resource preferences of the individuals with the largest intrinsic fitness and the functional profile becomes largely decoupled from the resource input. Differences of 0.1% and smaller are indistinguishable from the analytical prediction and the functional profile closely matches the one predicted by the resource influx rates.

The existence of these different regimes, where the functional composition is or is not affected by intrinsic fitness differences, strictly relates to the identification of the limiting factors shaping the communities. As mentioned before, in fact, the *intrinsic* differences could be due to abiotic factors, but also to limiting factors (such as phages) other than resource availability. If resources are limiting, we can expect that other factors have a minimal effect on strains' success. On the contrary, if resources are not the limiting factors and other mechanisms that determine strains' growth and decline, the distribution of the functional preferences in the population will not be robust, as it will be subject to the fluctuations of the other limiting factors. Our framework could be extended to explicitly include the factors responsible for intrinsic fitness differences (e.g. phages).

The metabolic trade-off is an essential ingredient of our framework. We implemented it as a fitness cost that is linear in the total rate of resource consumption. This choice generalizes models with fixed total rate [58, 48] by including a basal maintenance cost. This basal cost becomes negligible if the cost per gene, relative to the basal cost, becomes very large. The presence of a non-zero basal cost determines the existence of core resources, whose consumption is shared by all the individuals in the community. The form of the functional attractor is a mathematical consequence of the linearity of the metabolic trade-off. For linear trade-offs, the functional attractor is fully specified by the functional composition and is, in the limit of negligible fitness differences, independent of how functions are distributed across species. Non-linear trade-offs [12] could, in principle, affect the properties of the eco-evolutionary attractor. In the Materials and Methods, we explicitly consider both super-linear and sub-linear trade-offs and show that our results are left qualitatively unvaried. The taxonomic composition is largely affected by fitness differences, while the functional composition is robust. The stable functional composition display core and non-core resources, which are (at least approximately), linearly related to the effective resource influx rates.

A remarkable aspect of our framework is that functional composition — as opposed to taxonomic composition — naturally emerges as the relevant, reproducible degree of freedom suited to characterize ecological communities. Our results demonstrate in fact that the emergence of a stable and reproducible functional composition is a universal feature of consumer-resource-crossfeeding models [24]. This property is likely to hold more generally and not to be restricted to consumer-resource systems or microbial communities. We expect that a similar approach could be developed to study mutualistic communities or pathogen dynamics.

Chapter 2

An Equation of State for Ecological Communities

We can even take a step further and try to create a thermodynamic-like description of the community. The stable state can in fact be described by few quantities, univocally determined by the external condition and put in reciprocal relation by closed equations. These resemble the thermodynamic equations of state, creating a simple relations between the number of species S, the total population N and the functional abundances F_i.

2.1 Introduction

We consider the broad framework of consumer-resource-crossfeeding models, which describe the joint dynamics of metabolites and populations. Individuals extract energy from the environment by consuming externally provided resources. Their metabolisms modify resource concentrations also by secretion and leaking of metabolites. As we have shown in Chapter 1, under some mild conditions, these models are characterized by a globally stable equilibrium. Once the parameters characterizing the community are set, the system converge for large times to the same equilibrium point. If we start with a set of species $\mathcal{S}$ and a set of resources R, some species will go extinct and we will end up with a subset of $S \leq R$ coexisting species characterized by some population abundances, independently of the initial population densities and resource concentrations.

What is extremely challenging in the analysis of these models, which fundamentally hinders our ability to connect theory to empirical data, is to characterize in generality the properties of the equilibrium point. To fully characterize a community with S species and R resources one would need more than $R(S + R)$parameters, which are impossible to measure reliably in realistic settings. It is therefore natural, and of paramount importance, to focus on the properties of these equilibria that are typical, being robust across parametrizations.

The quest for general properties of equilibria has paralleled with the idea that such general properties should become more and more apparent as the number of species (and resources) becomes larger and larger. This idea has been pioneered by May [39], who showed that large, randomly interacting, communities respond to perturbations all in the same way: of the S^2 parameters characterizing the interaction between S species, only an handful of statistical properties of those interaction do actually matter. These results also apply to more realistic scenarios, which include different interaction types [6, 56], complex network structures [7, 28], and the effect of population abundances [23].

In the context of consumer-resources, powerful methods borrowed from disordered systems have allowed to characterize the equilibrium point in the regime where the number of resources R and the initial pool of species $\mathcal{S}$ is large [16].

The diversity of microbial communities strongly backs the idea that ecology should "go big" [5] and theoretical efforts should be directed in understanding models with large number of species. While undoubtely ecology should go big, there are multiple ways to do so. The vast majority of efforts have been focused in the regime where the size of the initial pool of species is comparable to the number of resources ($|\mathcal{S}| \sim R \gg 1$). While competitive exclusion principle [21] inevitably bounds — in absence of fine-tuning — the final number of species S, which is constrained not to exceed the number of resources R, there is not an a-priori constraint on the effective size of the pool of species which could potentially be part of a community. Here we challenge this assumption $|\mathcal{S}| \sim R$, and consider the regime in which the pool of species is much larger than the number of resources and the final number of species ($|\mathcal{S}| \gg R \geq S$).

Biochemistry operates in an high-dimensional space. While no estimate exists for bacteria, it is estimated that in humans there are of the order of 10^5 metabolites [62]. This motivates $R \gg 1$. On the other hand, microbial evolution could allow — a-priori — individuals to perform any combination of functions and grow on any combination of substrates. The exploration process driven by evolution allows biology to operate combinatorially in the biochemical space, suggesting $|\mathcal{S}| \gg R$. The observed diversity in a given community reflect that process of selection operated by the environment, which effectively select a small number S of these potential variants.

In the simplest setting we consider, individuals of a species either consume or not consume a given resource. The increase in growth rate caused by consuming more resources is paralleled with an higher metabolic cost in being able to do so. For each resource consumed, an individual pays a fitness cost χ necessary to maintain that function. Resources are externally provided and depleted by consumers. The total energy content of available resources H is not equally partitioned between resources. Each resource i is characterised by its own quality q_i, which represents the amount of energy than an individual could extract from resources of i in a lifetime, in absence of competitors and metabolic cost (see Materials and Methods). The community composition depends on the initial pool of species $\mathcal{S}$. Following the assumption $|\mathcal{S}| \gg R$, we consider every single 2^R species obtained by considering every possible combination of consumed and not consumed resources. We have therefore $2^R + R$ equations describing the interacting dynamics of population abundances and resources.

2.2 Results

2.2.1 Ecological Linkage Decoupling

In this section we rewrite our model in a form that allows to solve for its stationary state.

With the choices and assumption explained in section A.1, and with individual fitness $\epsilon_\sigma = 0$, the model can be written as

$$\frac{dn_\sigma}{dt} = \eta_\sigma n_\sigma \left(\sum_{i \in R} (1 - \ell_i) w_i \nu_i a_{\sigma i} r_i(c_i) - \frac{1}{\tau} \left(1 + \sum_{j \in R} \chi_j a_{\sigma j} \right) \right) \ . \tag{2.1}$$

$$\frac{dc_i}{dt} = h_i(c_i) - \frac{1}{w_i} \sum_j \left(\delta_{ij} - \ell_j D_{ij} \right) w_j \nu_j r(c_j) \sum_\sigma n_\sigma a_{\sigma j} \ , \tag{2.2}$$

The values of η_σ contribute only to determine the time-scales of the process but do not affect the fixed point or its stability. Therefore, we limit the calculation to the case $\eta = 1$.

Let us introduce the relative species frequency $x_\sigma = \frac{n_\sigma}{N}$, where $N = \sum_\sigma n_\sigma$ is the total biomass. We also introduce the functional occupancies

$$F_i = \sum_{\sigma \in \mathcal{S}} x_\sigma a_{\sigma i}, \tag{2.3}$$

that is, the fractions of individuals able to consume each metabolite i. The F_i for $i = 1, \dots, R$ represent the functional profile of the community.

We can now replace the equation for the species biomasses n_σ with the equations for the relative abundances x_σ and for the total biomass N:

$$\frac{dN}{dt} = \sum_\sigma \frac{dn_\sigma}{dt} = N \left(\sum_j (1 - \ell_j) w_j \nu_j F_j r_j(c_j) - \frac{1}{\tau} (1 + \sum_j \chi_j F_j) \right) , \tag{2.4}$$

$$\frac{dx_\sigma}{dt} = \frac{1}{N} \frac{dn_\sigma}{dt} - \frac{n_\sigma}{N^2} \frac{dN}{dt} = x_\sigma \left(\sum_i \left[(1 - \ell_i) w_i \nu_i r_i(c_i) - \frac{\chi_i}{\tau} \right] (a_{\sigma i} - F_i) \right) . \tag{2.5}$$

The equation for the resource densities becomes

$$\frac{dc_i}{dt} = h_i(c_i) - \frac{N}{w_i} \sum_j B_{ij} w_j \nu_j r(c_j) F_j, \tag{2.6}$$

where we named $B_{ij} := \delta_{ij} - \ell_j D_{ij}$.

Noticing that eqs. (2.4) and (2.6) depend only on the F_i, and not on the relative abundances x_σ, one would be tempted to write an equation for the F_i. However, if we do that, we discover that such equation is not closed, as it depends on F_{ij}, the fraction of individuals that consume both resource i and j (see Appendix B.1).

A set of closed equations can instead be obtained if we introduce the "community structure function"

$$G(\{k\},t) = \log\left(\sum_\sigma x_\sigma(t) e^{\sum_i k_i a_{\sigma i}}\right). \tag{2.7}$$

G is the generating function of the functional occupancies F_i, in fact by deriving over the parameters k_i it is possible to obtain their moments. For instance,

$$\frac{\partial}{\partial k_i} G(\{k\},t)\big|_{k_i=0} = \frac{\sum_\sigma x_\sigma a_{\sigma i}}{\sum_\sigma x_\sigma} = \sum_\sigma x_\sigma a_{\sigma i} = F_i. \tag{2.8}$$

The dynamics of the community structure function is described by

$$\frac{\partial G(\{k\},t)}{\partial t} = \frac{\sum_\sigma \dot{x_\sigma} \exp(\sum_i k_i a_{\sigma i})}{\sum_\sigma x_\sigma \exp(\sum_i k_i a_{\sigma i})} = \sum_i \left(\frac{\partial G(\{k\},t)}{\partial k_i} - F_i\right)\left((1-\ell_i) w_i \nu_i r_i(c_i) - \frac{\chi_i}{\tau}\right), \tag{2.9}$$

where we substituted Eq. (2.5) and noticed that

$$\frac{\partial G}{\partial k_i} = \frac{\sum_\sigma x_\sigma a_{\sigma i} \exp(\sum_j k_j a_{\sigma j})}{\sum_\sigma x_\sigma \exp(\sum_j k_j a_{\sigma j})}. \tag{2.10}$$

The set of equations for the dynamics of G, N and c_i fully describes the dynamics of the system, and in the next section we show that they can be used to find the stationary state, which will only depend on the functional profile of the community and not on the species abundances x_σ.

2.2.2 Attractor with no fitness differences

Let us give a description of the stationary state of the system by imposing that eqs (2.4), (2.6) and (2.9) are equal to 0. From eq. (2.9), we have

$$0 = \sum_i \left(\frac{\partial G(\{k\},t)}{\partial k_i} - F_i^*\right)\left((1-\ell_i) w_i \nu_i r_i(c_i^*) - \frac{\chi_i}{\tau}\right) \quad \forall \{k\}. \tag{2.11}$$

This equality must be true for all values of $\{k\}$. However, the term $\frac{\partial G(\{k\},t)}{\partial k_i} - F_i^*$ depends, in general, on $\{k\}$. As a consequence, for resources i for which that term depends on k, the other term $(1-\ell_i) w_i \nu_i r_i(c_i^*) - \frac{\chi_i}{\tau}$ must be equal to zero. For these resources, at stationarity,

$$r_i(c_i^*) = \frac{\chi_i}{(1-\ell_i) w_i \nu_i \tau}. \tag{2.12}$$

Then, there can be resources for which $\frac{\partial G(\{k\},t)}{\partial k_i}$ is independent of $\{k\}$. Given the expression in eq. (2.10), this is possible if and only if all the species have the same value of $a_{\sigma i}$, that is, $a_{\sigma i} = a_i \ \forall \sigma$. This happens if a resource is consumed by all species or by none of them. For these resources, then,

$$\frac{\partial G}{\partial k_i} = a_i = F_i^* \ \in \{0, 1\}. \tag{2.13}$$

In summary, if we neglect the resources that no species consume, we can distinguish two types of resources:

- Core resources R_c, consumed by all species, with $F_i^* = 1$
- Non-core resources R_{nc}, for which the stationary concentration is given by eq. 2.12.

Setting eq. 2.6 to zero, we find

$$F_i^* r_i(c_i^*) = \frac{\sum_j B_{ij}^{-1} w_j h_j(c_j^*)}{N^* \nu_i w_i} =: \frac{q_i \chi_i}{N^* \nu_i w_i (1 - \ell_i)\tau} \ , \tag{2.14}$$

where we defined the resource quality q_i as

$$q_i = \frac{\tau(1 - \ell_i) \sum_j B_{ij}^{-1} w_j h_j(c_j^*)}{\chi_i} \ . \tag{2.15}$$

By imposing stationarity in equation 2.4 we obtain in turn

$$0 = \tau \sum_i (1 - \ell_i) \nu_i w_i F_i^* r_i(c_i^*) - 1 - \sum_j \chi_j F_j^* = \frac{\sum_i \chi_i q_i}{N^*} - 1 - \sum_j \chi_j F_j^* \ , \tag{2.16}$$

from which we obtain

$$N^* = \frac{\sum_i \chi_i q_i}{1 + \sum_j \chi_j F_j^*} \tag{2.17}$$

For non-core resources we can use equation 2.12 together with equation 2.14 to obtain

$$F_i^* = \frac{q_i}{N^*} \text{ if } i \in R_{nc}, \tag{2.18}$$

Motivated by the fact that the functional abundances are bounded from above to 1, we finally obtain

$$F_i^* = \min\{1, \frac{q_i}{N^*}\} \ , \tag{2.19}$$

Equations (2.14), (2.17) and (2.19) describe the stationary state of the community dynamics when $\epsilon_\sigma = 0$. As anticipated, this stationary state is characterized by its functional profile.

The results of the numerical simulations of the model with $\epsilon \to 0$ (see Appendix B.2 and B.3) confirm the theoretical predictions for the stationary values of F_i, N and S (Fig. 2.1).

The value of q_i is determined by solving eq 2.14 for both core and non-core resources. An efficient algorithm to solve these coupled equations computationally is presented in Appendix B.3.

Considering $\epsilon \sim 0$ is a fair approximation in the limit of a large number of resources R as the distance from the functional manifold caused by ϵ vanishes in the limit $R \to \infty$.

Note that if resources are externally supplied without dilution (or if dilution of resources is negligible compared to their consumption), i.e., if $h_i(c_i^*) = h_i$ equation 2.15 becomes a definition and there is no need to find the stationary concentration of resources to determine the resource quality q_i.

2.2.3 Thermodynamic description

In this section we investigate the relationship between the total community biomass N^* and its diversity S^* when some key model parameters vary. In particular, we consider the parameters related to the resources: their number, quality, metabolic cost and heterogeneity. We introduce the normalized qualities and costs

$$\eta_i = \frac{q_i}{\bar{q}}, \quad \gamma_i = \frac{\chi_i}{\bar{\chi}}, \tag{2.20}$$

and let $p(\eta, \gamma)$ be their joint probability distribution. It is also useful to introduce

$$K = R\bar{q}, \quad X = R\bar{\chi}. \tag{2.21}$$

K is the total energetic content of the resources, while X measures the fitness difference between a generalist and a specialist.

To be able to write analytically the dependence of S^* and N^* on these parameters, we consider the limit of a very large number of resources, $R \gg 1$. In this limit, the resources qualities q_i and metabolic costs χ_i as they can be considered continuous quantities characterized by their probability distributions.

In the continuous limit, the quality of the first core resource can be written as

$$\eta_c = \frac{X \int_{\eta_c}^{\infty} d\eta p(\eta) \eta \langle\gamma\rangle_\eta}{1 + X \int_{\eta_c}^{\infty} d\eta p(\eta) \langle\gamma\rangle_\eta}, \tag{2.22}$$

where $p(\eta) = \int_0^\infty d\gamma \; p(\eta, \gamma)$ is the marginal distribution of normalized resource qualities and $\langle\gamma\rangle_\eta = \int_0^\infty d\gamma \; \gamma p(\gamma|\eta)$ is the average normalized cost of resources conditioned to their normalized quality. Additionally, the total system biomass can be expressed as

$$N^* = \bar{q}\eta_c. \tag{2.23}$$

The derivation of eqs (2.22) and (2.23) can be found in the Material and Methods section. Additionally, since the number of surviving species is equal to the number of non-core resources plus one, we have

$$S^* = 1 + R \int_0^{\eta_c} d\eta p(\eta). \tag{2.24}$$

Equations (2.22–2.24) express the dependence of N^* and S^* on the properties of the resources, and are valid for any joint probability distributions of qualities and costs. However, in this general case, it is not possible to write an explicit expression of η_c from eq. (2.22). Therefore, we make two assumptions that allow us to write η_c explicitly. The first assumption is that the normalized resource qualities η are distributed uniformly in $[1-\sigma/2, 1+\sigma/2]$. The parameter σ quantifies the heterogeneity of resources qualities, and we have $\sigma \in [0,2]$ as η must be positive. The second assumption is that there is a linear relationship between resources costs and qualities, that is,

$$\langle\gamma\rangle_\eta = 1 + \lambda\sigma_\gamma \left(\frac{\eta-1}{\sigma}\right), \tag{2.25}$$

where $\lambda \in [-1,1]$ measures the strength of this relationship and σ_γ quantifies the heterogeneity of the resource costs. A positive λ describes a positive correlation between costs and qualities, i.e. a high quality resource has also a high metabolic cost, while a negative λ describes a negative correlation. Given that $\langle\gamma\rangle_\eta > 0$, we must have $|\lambda| < 1/\sigma_\gamma$. Using these two assumption in eq. (2.22), we find that it has a solution under the condition $X > \frac{12-6\sigma}{6\sigma+\lambda\sigma_\gamma\sigma}$. We denote this solution as $\eta_c(\lambda\sigma_\gamma, \sigma, X)$, see Methods for more details. Then, we finally have

$$\begin{aligned} N^* &= \bar{q}\ \eta_c(\lambda\sigma_\gamma, \sigma, X) = \frac{K}{R}\eta_c(\lambda\sigma_\gamma, \sigma, X), \\ S^* &= 1 + \frac{R}{\sigma}\left(\eta_c(\lambda\sigma_\gamma, \sigma, X) - 1 + \frac{\sigma}{2}\right). \end{aligned} \tag{2.26}$$

Now, we can consider different transformation of the parameters characterizing the resource distributions, and examine the changes of N^* and S^*. The parameters whose change affects N^* and S^* are σ, K, X, R and $\lambda\sigma_\gamma$. Let us explore some interesting transformations.

Varying the heterogeneity of resources quality

We first consider the transformation where the heterogeneity of the resources quality σ varies, while the number of resources R and their average qualities $\bar{q}$ remain constant. In this case, N^* and S^* both increase with increasing σ (see Figure 2.1A). The behavior is the same for positive, null or negative correlation λ (respectively, squares, triangles and circles in the Figure).

Varying the average resource quality

If we vary the average resource quality $\bar{q}$, keeping all the other parameters constant, we obtain that the total energy content of resources K varies. This makes N^* vary proportionally to K, while S^* does not change, as it does not depend on K.

Varying the average metabolic cost

If we vary the average metabolic cost $\bar{\chi}$, keeping all the other parameters constant, we obtain that the fitness difference between generalists and specialists varies. Increasing X makes η_c increase, that is, there are less core-resources. As a result, both S^* and

N^* increases at the same pace (see Figure 2.1B). While it might sound counter-intuitive that the total biomass increases when the average metabolic cost of resources increases, what happens is that each species metabolises a smaller number of resources, therefore the total metabolic cost it pays is smaller, although the cost per resource is higher.

Varying number of resources

There are several possibilities to vary the number of resources.

First, we can have a transformation where only R varies, while $\bar{q}$ and $\bar{\chi}$ are constant. Consequently, the total energetic input K and the fitness difference between generalist and specialist X would vary proportionally to R. In this situation, the increase of R makes both S^* and N^* increase (Fig. 2.1C). The behavior is qualitatively the same for all λ.

Secondly, we can consider a transformation that mimics the protocols of experiments such as [17], where the number of resources R is varied but the total energetic input K remains constant. This would be obtained varying the average quality $\bar{q}$. This situation creates an interesting trade-off between biomass and diversity (Fig. 2.1D), where increasing R makes S^* increase but N^* decrease. The behavior is qualitatively the same for all λ.

If instead we change R keeping X constant, only the number of species is affected, and changes proportionally to R, while the biomass remains constant. This means that if we have more, simpler resources we would have more species that if we had less, more complex ones.

Varying the correlation between cost and quality or the cost heterogeneity

If we increase the value of $\lambda\sigma_\gamma$, all other parameters being constant, both N^* and S^* increase, as can be seen in all panels of Fig. 2.1. This means that a strong positive correlation between cost and quality of resources allows more species to coexist.

For all transformations, the behavior predicted analytically is confirmed by numerical simulations performed with $\epsilon \to 0$ (colored markers in Fig. 2.1).

As we can notice in the figure, we observe a small shift of the total biomass values towards the right. Any value of $\epsilon \neq 0$ determines the existence of a set of species with individual fitness $\sim -\epsilon$. In the limit $S \to \infty$ also the number of species in the set tend to infinity. The final community, as the fittest survives, is then composed by all species with individual fitness $\sim -\epsilon$. The community pays an overall cost lower than the one predicted by χ, allowing for the survival of more individuals and thus returning an higher biomass.

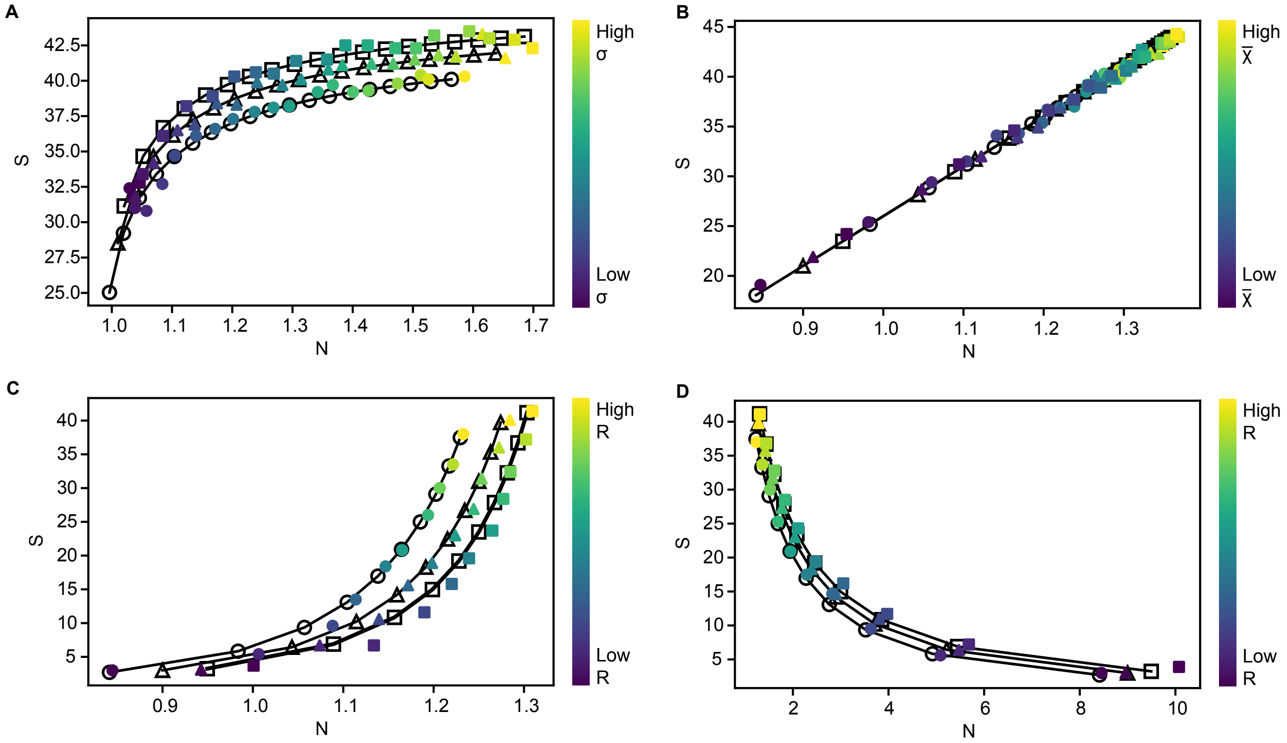


Figure 2.1: Variation of the total biomass N^* and of the number of species S^* under different transformations changing resource-related parameters. Black curves and markers represent analytical prediction, while colored markers are the results of numerical simulations with $\epsilon \to 0$. For each transformation, three possible values of $\lambda\sigma_\gamma$ are shown: -0.8 (circles), 0 (triangles) and 0.8 (squares). A) Transformation varying the heterogeneity of resources qualities σ from 0.2 to 2. The other parameters are fixed at values: $R = 50$, $\bar{\chi} = 1$, $\bar{q} = 1$; B) Transformation varying the average metabolic cost $\bar{\chi}$ from 0.1 to 2. The other parameters are fixed at values: $R = 50$, $\sigma = 1$, $\bar{q} = 1$; C) Transformation varying the number of resources R between 5 and 50. The other parameters are fixed at values: $\sigma = 1$, $\bar{q} = 1$, $\bar{\chi} = 1$; D) Transformation varying the number of resources R from 1 to 10 while keeping the total energy input K constant at a value of 50. The other parameters are fixed at values: $\sigma = 1$, $\bar{\chi} = 1$. All combinations of parameters are chosen to satisfy the constraints. The marker color, from blue to yellow, indicates the increase of the parameter that is varied.

2.3 Discussion

We showed that consumer resource models predict stationary states well described by few macroscopic parameters, the total biomass N^*, the number of surviving species S^* and the functional abundances F_i^*. Such parameters depend on the externally provided resources and are mutually dependent through a closed set of equations.

By considering the limit of a very large number of resources $R >> 1$, it is possible to depict a thermodynamic like description of communities, identifying a set of transformations that smoothly change the system parameters.

We characterized such transformations and showed how analytical calculations are confirmed by numerical simulations.

Chapter 3

Estimating the Genome Length Distribution from Community Functional Composition

We saw how function abundances are regulated by the environment. Consumer-resource models well represent that equilibrium states are functionally characterized at the community level while species are free to fluctuate. This parallels with experimental result [15] shown in Fig.1. Species composition drastically differ among samples while function composition appears stable. Nevertheless, Such a stability has been analysed just qualitatively, by comparing the variability of species and functions. To reach a quantitative conclusion we should compare the observed variability with the typical one.

By determining the key parameters that drive the composition, we can develop a null model to predict the composition of a randomly assembled community. Only by comparing the observed variability with the one predicted by the null model we can reach a quantitative statement, understanding if what we see is more stable than expected.

One of the main factors to take into account is the limitedness of the existing pool of species. Living beings are not randomly assembled as their genomes composition evolved following robust rules (Sec.3.2.1). If we simplify bacteria to be collections of functions, these rules imply that evolution generates species belonging to a small subset of all possible functional combinations. Consequently, there are constraints on the explorable community functional compositions, as the community is contains only genomes belonging to this subset. (For a mathematical description see App. **??**)

For example there exist functions appearing in almost all the existing strains, making it hard to find an abundance variation in the community, independently from the species composition. It is thus important to understand whether the observed variability is actually lower than the one expected by chance, randomly assembling a community.

In the following chapter, going towards the design of a null model for communities, we will focus on one particular property of bacterial genomes, the correlation between genome length and composition. In Sec.3.2.1 we present such a dependence and show how it affects the community composition. The genome length distribution of a community is an important parameter to be understood and evaluated. We want here to exploit this relation to infer information on the species composition from the community functional composition.

3.1 Introduction

Bacterial communities are extremely complex objects, made by thousands of different interacting species. The effective environment surrounding those species is determined by both external conditions and the neighbouring species, making it hard to understand the effective living condition in which they are growing.

The main experimental technique to characterise the species community composition has been for a long time the physical isolation of the strains. It proceeds by diluting the communities, streaking them on an agar plate, until a clear and unique morphology appears. The main requirement for the effectiveness of the method is that the species are able to grow alone on an agar plate but, given our scarce knowledge on the interactions among species, such a requirement is found to be extremely hard to meet.

Amplicon sequencing techniques have been game changing tools in this regards. They do not rely on the physical isolation of strains, allowing for analysing environmental samples, directly identifying the species via the recognition of specific highly conserved marker genes. Specifically $16S$ gene is a section of bacterial genome present across all species and highly conserved, allowing to identify the species and tracing a taxonomic tree. Once the species are identified one can reconstruct the community by the use of databases containing the characterisation of the detected species.

Nevertheless the latter procedure is far from being perfect. Relying on previously measured genomes, the marker genes can belong to a strain never analysed before. This can lead to unusable data or to identify individuals at an high taxonomic level (species, family, ..). Moreover, it relies on the assumption that marker genes are always able to identify different strains. The former assumption is not always true as bacteria evolve so rapidly that the same $16S$ gene could be associated to different strains. It is thus important to find alternative solutions to characterise the individuals with the available data. Particularly, we want to develop a method that does not rely on databases.

Metagenomics is an experimental technique that takes a completely different approach to the problem, not trying to identify single species but considering the community as a whole. The genomes present in a sample are analysed together, identifying the genes but not the species they belong to. The outcome is a list of genes and their abundances describing the whole community. The latter list can be connected to the functional composition of the community as genes are one way that life found to encode the information to perform functions.

On the other hand we have information on the genomes composition. Genomes are ensembles of genes that, as shown in Sec.3.2.1, are built following specific and robust rules. Genome composition has a strong dependence on the total number of genes it contains, i.e. its length. The existence of these rules put some constraints on the possible functional compositions of the communities and leave marks of the length distribution of the genomes on the community functional composition, giving us a way to extract useful information from the latter data.

There exist various tools trying to extract the average length of the genomes in a community from the functional composition data [51]. They mainly rely on the existence of genes that appear in a fixed number of copies in all the strains. They proved to be reliable in this regards, returning accurate estimates of the average genome length but cannot provide any further information on the distribution. Moreover they are under exploiting the dataset by using only a very small fraction of the available genes.

We here develop a new method, relying on the regularity of all the genes, looking for extracting higher moments of the length distribution.

First of all, using a database, we derive the laws determining such regularities. The former databases cluster genes into families that encode similar information and collect thousands of genomes expressing the number of copies of all gene family present in each genome. Among the existing databases we use the PFAM [41]. Using these databases we determine two relevant relations between the family composition and the genome length.

3.2 Main

3.2.1 Genome Length and Composition

Looking at the genomes composition in terms of gene families, one very robust observation is the relation between the number of copies of a gene family and the genome's length [42]. Longer genomes not only add new families to their repertoire but also increase the number of copies of the one owned by shorter genomes. As shown in Fig. 3.1**A**, the copy number follows a power law as a function of the genome length, with exponents raging from 0 to more than 2. This means that some genes have an average copy number independent from the length of the genome while others scale super-linearly. Therefore, relative abundances, the number of copies over the total number of gene families in the genome, also depend on the genome length. Long genomes likely have an higher ratio between high and low exponent genes than short genomes, thus appearing functionally different.

While the dependence of family copy number on the genome length is a very robust result, proved on thousands different strains and genes, it is derived under the bias of the presence of the gene. The power laws dependence is obtained considering just those species that have at least one copy of the gene. Anyway, genes are not always present and could be characterized also by their occurrence over the existing strains. There exist genes with occurrences ranging from 0 (never occurring) to 1 (present in all the examined strains).

It is important to notice that occurrences are not evenly distributed along genome lengths. That means that a gene with a given occurrence o_i over the existing strains, won't be missing evenly across genome lengths but will more probably be missing among short/long genomes.

We did a preliminary study of the distribution of occurrences and found a clear dependence of their distribution on the average occurrence o_i . As shown in Fig. 3.1**B**, genes with high occurrence (blue), present in almost all the strain considered, are more often missing in shorter genomes while lowering the occurrence the vacancies become more evenly distributed across lengths.

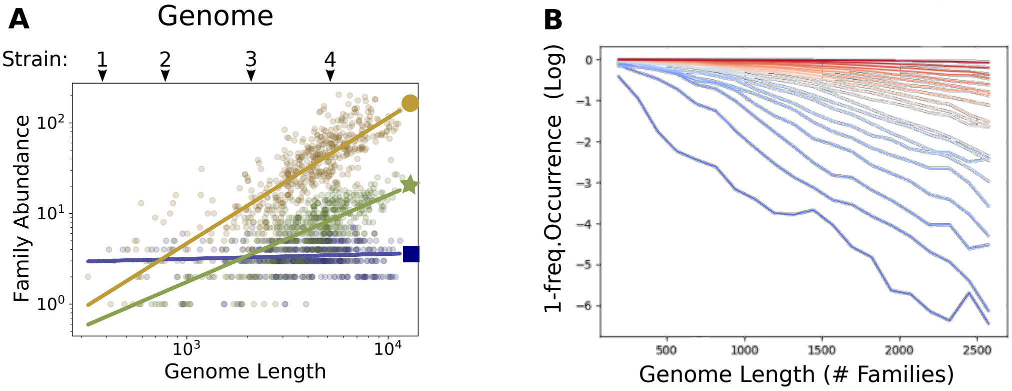


Figure 3.1: Two laws determining the family composition of genomes. **A**: Family abundance has a non-linear dependence on genome length (PFAM domains). The panel shows example PFAM family data. The dependence can be expressed as a power law. Linear, sub-linear and super-linear cases are found in data [42]. This implies that the abundance ratio between two families changes with the length of the genome. **B**: The occurrence of genes among genomes of comparable size depend on the size itself and on the overall occurrence of the gene. We plot the frequency of the vacancies, how many genomes miss the gene. Genes that are rarely present (red) are missing equally across lengths while genes that are almost always present (blue) are mainly missing among short genomes.

These two properties put some constraints on the community functional composition and determine its dependence on the genome length distribution of the strains. As pictorially shown in Fig. 3.2, combining strains with shorter genomes brings to different functional composition than the one obtained with long genomes.

Knowing the genome length distribution of a community becomes thus important when considering its functional composition stability. The former distribution determines what to expect from an average community of that size, allowing the recognition of peculiar fluctuations. On the other hand, not knowing the distribution we could misinterpret the data. Functions that seem stable (fluctuating) across samples could simply be the result of a constant (varying) average length.

While knowing the strains or even simply their genome length easily allow us to design a null expectation for the functional composition of the community, the inverse is not as straightforward. We present an algorithm that, relying on the two laws above presented, is able to infer the genome length distribution of the bacteria in the community from the community functional composition data.

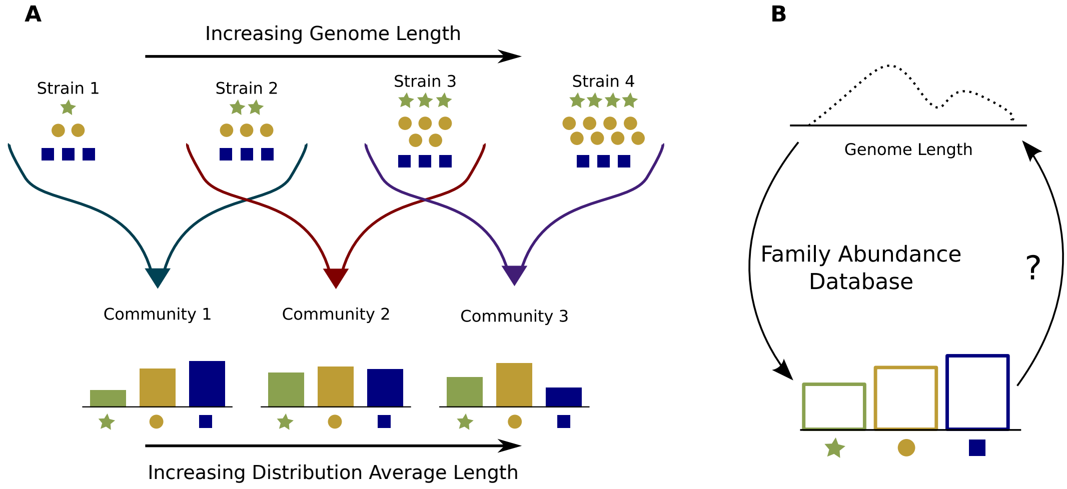


Figure 3.2: The dependence of gene families abundances on genome length has consequences on the community functional composition. **A** illustrates how family scaling at the genome level impacts on family abundance at the community level. Ensembles of short genome strains are functionally diverse from those containing long genomes strains. For example blue squares represent a family with constant copy number across genomes, implying relative abundance decreases with genome length. **B**: Given the genomes compositions, from the genome length distribution we can create a null expectation for the functional abundances of the community. The scope of this work is to derive a method for the inverse problem, extrapolating the genome length distribution from the family composition of a community.

3.2.2 Algorithm

We look for a description of the average composition of genomes in function of their length. We could be tempted to directly use the power laws' parameters as predictors of the average copy number at a given length, obtaining an analytical expression. Unluckily, the non trivial behaviour of occurrences strongly modifies the average composition of genomes at a given length predicted by the power laws.

As shown in Fig.3.3 to overcome this problem, we decided to bin genome lengths and calculate the average composition of genomes inside each bin. In this way, obtaining a discrete description of genome composition, we are sure to take into account both the laws we presented and any other existing property connecting genome length and composition.

Practically, we take the PFAM database D_{sf}, saying the number of copies of family f are present in the strain s. For each strain we calculate its genome length $\ell_s = \sum_f D_{sf}$, the total number of families it contains. We then divide genome lengths in n bins b_l, each centred around ℓ_l, and calculate the average composition of genomes inside each bin

$$M_{lf} = \frac{\sum_{s|\ell_s \in b_l} D_{sf}}{\sum_{s|\ell_s \in b_l} 1} \tag{3.1}$$

M_{lf} is thus the average number of copies of family f present in a genome of length ℓ_l.

It is important to notice that, even though we used a database to derive M_{lf}, the final result represents a general property of genomes composition, independently from the single strains contained in D_{sf}.

Having a way to connect genome lengths and family abundance without relying on databases we now look for determining the genome length distribution of a community. The procedure aims at designing a sample community, created by the use of M_{lf}, functionally as close as possible to the data. We will proceed by successive approximations, beginning from an initial guess, calculating its functional composition and comparing it to the one of the data. The comparison is done via the Kullback-Liebler divergence, a way to quantify the information loss that results from representing the data with the sample.

Consider a bacterial community with an unknown genome length distribution $P(\ell)$. It can be analysed via metagenomics techniques, obtaining a set of family abundances R_f, i.e. how many copies of each gene family f are detected.

To find $P(\ell)$, we create a sample community with its own length distribution $\tilde{P}(\ell)$. We discretise $\tilde{P}(\ell)$ into a vector d_l, encoding how many bacteria are present for every genome length ℓ_l. We then calculate the average functional composition of the sample community by using M_{lf}: family abundances are obtained as $F_f(d_l) = \sum_l M_{lf} d_l$. To compare them to the family abundances R_f we define the relative abundances $f_f(d_l) := \frac{F_f(d_l)}{\sum_{f'} F_{f'}(d_l)}$ and $r_f := \frac{R_f}{\sum_{f'} R_{f'}}$.

We can eventually look for the Kulback-Leibler divergence between the two relative abundances $C(d_l|r_f)$, that is how much information are we loosing by representing the data with our sample community

$$C(d_l) := -\sum_f r_f \log_2 \left(\frac{f_f(d_l)}{r_f} \right) \tag{3.2}$$

By modifying the vector d_l we look for minimizing such a quantity. We do it by successive approximations, as sketched in Fig.3.3

0: Consider an initially empty virtual community $d^0 = \underline{0}$

1: Take the vector $d^1 := \hat{l}$, equal to 1 along l and zero elsewhere, that minimizes $C(d^1)$. Genomes of length ℓ_l the ones functionally closest to the whole community

2: Take $\hat{l}'$ such that $d^2 = d^1 + \hat{l}'$ minimizes $C(d^2)$. That is, if $l \neq l'$, d^2 equals 1 in l, l' and 0 elsewhere, if $l = l'$, d^2 is 2 in l and 0 elsewhere)

n: Take $\hat{l}^n$ so that $d^n = d^{n-1} + \hat{l}^n$ minimizes $C(d^n)$

The value of $C(d)$ decreases during the process finally reaching a plateau, oscillating around a finite value. From d_l^n we can easily obtain the desired length distribution $P(\ell_l) = \frac{d_l^n}{\sum_k d_k^n}$.

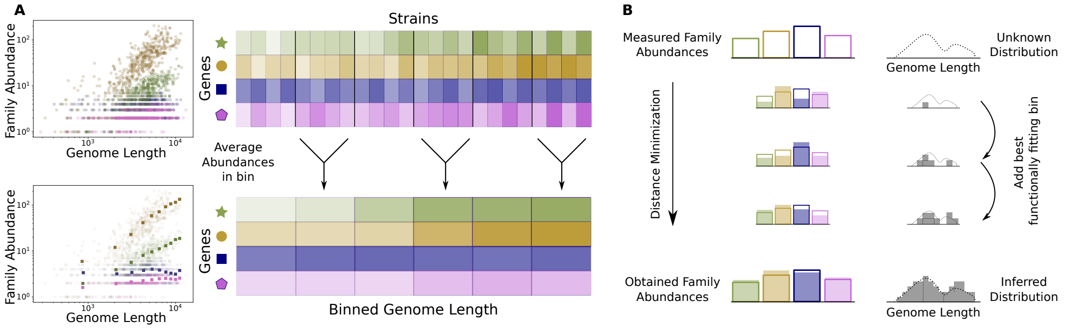


Figure 3.3: Illustration of the algorithm. The dataset is binned and averaged to obtain the average functional composition at a given length. The obtained database is used to reconstruct the genome size distribution of the community. **A**: The $s \times f$ data table containing the family copy number of each strain is averaged over genome length bins. The obtained $l \times f$ data table, $l << s$, encodes the average functional composition at a given genome length. **B**: Given a measured functional composition of a sample, the algorithm aims to infer the genome length distribution by successive approximations. The first iteration selects a community c_1 of length l_1, the functionally closest to the sample. The second iteration creates the community c_2, half of length l_1 and half l_2, and so on. The latter is chosen to minimise the functional distance with the sample. These iterations are repeated up to the desired degree of convergence.

3.3 Results

3.3.1 Virtual Communities

The first test of the algorithm effectiveness was performed on virtual communities, created by randomly selecting strains from the PFAM database. To make test more reliable we decided to split in two the database, the first half was used to generate the virtual communities while the second to infer the genome length distribution with our method. In this way there is no strain in common between the two, avoiding possible biases arising by guessing a distribution knowing the exact strains it contain. This is surely a more likely situation happening when analysing environmental data, given the fact that only 2% of existing strains have been isolated and characterised up to now.

We thus create, with the procedure explained in App.C.2, a community with genome length distribution $P_v(\ell)$ and gene family abundances R_f.

Given the family abundances R_f we run the algorithm and compare the results with the exact distribution of the virtual community.

As shown in Fig. 3.4, the mean, standard deviation and the shape of the distribution are well inferred by the algorithm. The errors on the mean and the standard deviation are of 2% and 15% and the shape properly identifies the relevant peaks. This outperforms the previously existing methods that infer the mean of virtual communities with a 2% error but do not give any estimate neither of the standard deviation nor of the shape of the distribution.

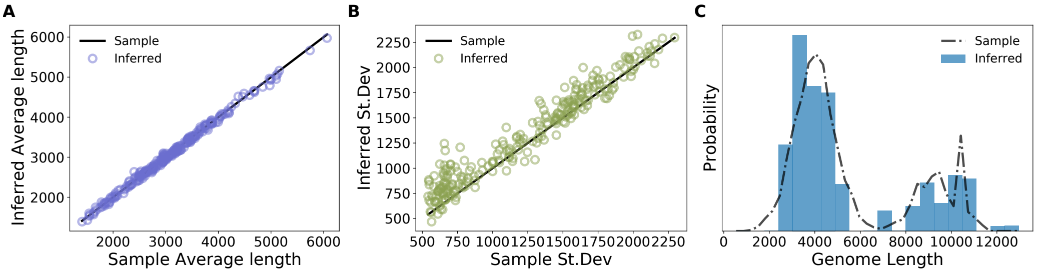


Figure 3.4: The algorithm detects precisely the mean genome size and its variability in virtual communities. Each virtual community was created from a prescribed genome size distribution with 500 random strains, none of which was present in the table of reference genomes used by the algorithm. The genome lengths used for the virtual communities shown in panels **A** and **B** are Log-normally distributed. **A**: The inferred means (violet circles, each symbol corresponds to a different virtual community) precisely represent the mean (solid line). Relative errors are of the order of 2%. **B**: The inferred standard deviations (green circles) compare well to the reference ones (solid line). Relative errors are of the order of 15%. **C**: The algorithm can detect details of the genome size distribution, such as multimodality. Example of comparison of the inferred distribution (blue histogram) to the sample distribution (dotted line) for a virtual community with many peaks.

3.3.2 Mock Communities

The second test was performed on mock communities, laboratory assembled communities, created by mixing fixed amounts of well known strains. They provide a great tool to test our method since we can compare our results to the ones obtained from the known abundances of the strains in the sample.

We used two different mock community experiments: [30, 52]. They both are a mixture of bacteria (~ 20) and eukaryotes, viruses, archaea in smaller amounts. The fact that also non bacterial genomes are in the samples tests the algorithm in a more realistic configuration.

For samples [52] the results are obtained in two ways, by the algorithm previously explained and by its generalisation to different domains. In the latter the database used for inferring the distribution is one containing all 4 different domains present in the samples. This allows the algorithm to identify also genomes behaving differently from the bacterial ones, with genome lengths way smaller (viruses) or longer (eukaryotes).

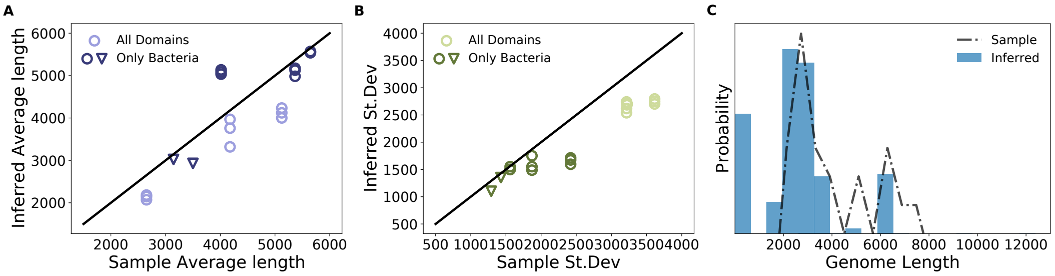


Figure 3.5: The algorithm infers the average and standard deviation of genome length probability distributions in Mock communities. Two different reference mock communities are present. in **A**, **B** Circles represent the 9 samples of [52] while triangles the 2 samples of [30]. among circles darker color shows the results obtained whit an algorithm using a bacteria only database. In this case also the reference average length and standard deviation are calculated on the samples' bacteria alone. Lighter color represents the result obtained by the algorithm using a database with 4 domains (Eukaryotes, Archaea, Viruses and Bacteria) and the reference values are calculated using all the strains in the samples. **C** shows the distribution obtained (blue histogram) for one of the samples in [30] compared to the actual distribution of lengths in the sample (dashed line).

As shown in Fig. 3.5, also for the mock communities the trend of the average length and distribution's standard deviation were properly identified, together with the relevant peaks of the distribution.

At the same time we can notice a way worse precision of the estimates, compared to the ones obtained on virtual communities. We can explain such a difference by one important feature of these samples. While natural communities (and the virtual communities we created) are composed by hundreds or thousands of different strains, mock communities are smaller, composed by tens of strains. This makes single strain variability way more relevant. The strains specificities arise in the community functional composition and it becomes more difficult to detect the average behaviours expressed in Sec.3.2.1.

3.4 Discussion

The increasingly important world of metagenomics is proving to be an incredibly powerful tool to analyse complex bacterial communities. It can provide huge amounts of data on the communities functional composition.

The existing procedures to characterise the species composition from the analysis of 16S marker genes, even if powerful, depend on the databases and rely on the ratio between the mutation rate of the target genes and the other parts of the genome. Being able to derive information on the individuals directly from the functional abundances, without the need of species specific information, could thus increase the effectiveness of the analysis.

The developed algorithm is conceived in this context, trying to obtain the genome length distribution of the community directly from the functional composition. The use of the databases, even though necessary, doesn't look for information on the single strains, as the algorithm only needs the average behaviour of genomes. This is a substantial element as long as most strains in the environment have not been isolated and characterised yet.

The existing methods [51] that try to follow the same path rely on single copy genes to derive the average genome length, calculating the number of individuals in the community from their abundance. The former procedure proved to be reliable to find the average length of virtual communities, obtaining a 2% error on the result, but cannot further explore the distribution as no information on the higher moments can be inferred.

We used the copy number regularity as a function of the genome length to link the functional composition to the genome length distribution. This allows to infer higher momenta of the distribution and to lower the sensibility to sampling errors by using all the genes instead of just core, constant ones. The obtained precision is comparable to [51] for the mean while also providing the standard deviation and the shape of the distribution, as shown in Tab. 3.1.

	MicrobeCensus	Our Method
Average	Yes: 2%	Yes: 2%
St. Dev.	No	Yes: 15%
Distribution	No	Yes

Table 3.1: Comparison with the state of the art method. We show whether each method is able to determine different properties of the genome size distribution and the respective relative error on the estimate when analysing virtual communities.

We expect this algorithm to be useful to better understand functional stability across samples, giving a way to determine a null model to predict the expected average composition of a community with given genome length distribution.

Chapter 4

Evolutionary Stability of Cooperation in Stochastic Multiplicative Environments

This chapter is part of a paper uploaded on the Arxiv [19]

By cooperating, an individual performs an action with an immediate negative payoff but positively affecting others. As [47] shows, in stochastic multiplicative processes groups of agents should favour cooperation over selfishness. The effect of cheating in this system was considered: is taking advantage of a cooperating partner without returning anything back the winning strategy? We analyze the evolutionary stability of cooperation in stochastic multiplicative processes at different time horizons and characterize the equilibrium point. Surprisingly, cooperation is found to be evolutionary stable: cheaters, after a first fast-growing transient, slow down their growth rate. The effective payoff matrix shows a transition as the time horizon stretches, making cooperation a Nash equilibrium. The cooperation dilemma thus disappears in the context of agents maximizing their long term return. This evidence suggests possible explanations for existing phenomena and new optimal cooperative strategies.

4.1 Introduction

The emergence and the stability of cooperation is a central problem in biology, sociology, and economics [43]. Cooperation produces an advantage for the group, through the creation and sharing of social goods, but is inherently unstable to cheating and to the tragedy of the commons, where individual agents benefit from the social good without contributing to its creation [49]. The dilemma of the evolution of cooperation can be solved in presence of one or more specific mechanisms [43], which lead to the emergence and long-term stability of the cooperative trait.

One key aspect in common between multiple systems is that the environment is subject to fluctuation and stochasticity. A paradigmatic example, which has applications in both economics and population biology, is given by the geometric Brownian motion, which describes the stochastic dynamics of a variable $x(t)$ as $\dot{x} = \mu x + \sigma x \xi(t)$, where $\xi(t)$ is a delta-correlated white noise. In biology, x could represent the abundance of a population, in economics x is the value of an asset, while in game theory is the wealth accumulated by a gambler. In general, this equation describes growth under a stochastic multiplicative process. An essential feature of multiplicative growth is that it lacks ergodicity [47, 44], as the time-average behavior differs from the ensemble average. The latter grows exponentially in time with rate μ, while the former grows with rate $\mu - \sigma^2/2$. This difference parallels the difference between arithmetic mean (which corresponds to the ensemble average) and geometric mean (which converges to the time average), and it is the deep reason why the latter is a natural quantity to optimize for agents aiming at maximizing their future profits or growth. In the context of gambling, the Kelly criterion defines the optimal size of a bet based on optimization of the geometric mean [57, 32]. In evolutionary biology, under varying environmental conditions, natural selection favors traits on the basis of their geometric mean fitness [50, 27]. An important consequence of the fact that the geometric fitness determines the optimal solution is that not only the average environment but also the amplitude of its fluctuations, determine its values, as the geometric average grows with rate $\mu - \sigma^2/2$. Reducing fluctuations, i.e., reducing the value of σ, has, therefore, a positive effect and should be expected to produce better strategies and be advantaged by natural selection [2].

In the context of growth under fluctuating conditions, we introduce the possibility of cooperation between G agents, by generalizing the setting of [20, 47, 33]. The value f_i of agent i grows as

$$\dot{f}_i(t) = \mu f_i(t) + \sigma f_i(t) \xi_i(t) + \frac{1}{G} \sum_{j \neq i} (\alpha_j f_j(t) - \alpha_i f_i(t)) \tag{4.1}$$

In the beginning, we will consider the case of white uncorrelated noises. We will generalize this setting to colored noise with the arbitrary correlation between $\xi_i(t)$ and $\xi_j(t)$. The terms proportional to α_i represent the effect of sharing. At each time-step, each of the individual share a fraction $\alpha_i \in [0, 1]$ of its value f_i as a public good. The public good is then instantaneously divided equally among the agents. A value $\alpha_i = 1$ represents full cooperation, where an individual shares all its value. While $\alpha_i = 0$ represent defection.

If all the agents follow the same strategies (i.e., if $\alpha_i = \alpha$), one can obtain an exact solution of the trajectories $f_i(t)$ [20, 47, 33]. In particular, one can obtain the growth rate of the geometric average of the values $g_i = \lim_{t \to \infty} \langle \log f_i(t) \rangle / t$ equals $\mu - \sigma^2(1 - \alpha(G-1)/G)/2$. It is easy to see that g_i is a monotonically increasing function of both α and G. The full defector scenario $\alpha = 0$ corresponds to the original Geometric Brownian motion solution $g_i = \mu - \sigma^2/2$. The full cooperation case $\alpha = 1$ leads instead to an higher growth rate $g_i = \mu - \sigma^2/2G$. The intuition behind these results is that, in this context, cooperation produces an advantage as it reduces effectively environmental variability. By sharing their values with others, agents effectively diversify their investments, making their values less subject to fluctuations, therefore, leading to faster growth.

This result shed the light on the importance of cooperation in fluctuating environments: cooperation screens individuals from the negative effect of variability. This result does not explain however how cooperation emerged and why it could be stable to defection. Also in the simple context of the prisoner dilemma, cooperation produces an individual advantage over defection, when all agents cooperate (i.e., cooperation is Pareto optimal). The dilemma is, as well known, that cooperation is not stable (given that all the other agents are cooperating is advantageous for the individual to defect) while defection is (if all the agents are defecting there is no advantage in starting cooperating). In this chapter, we explore the stability and origin of cooperation in fluctuating environments, using the setting of eq 4.1. We show that the maximization of the individual long-time return leads to the emergence and stability of cooperation. We further explore the robustness of these results to correlated fluctuations, colored environmental noise, and finiteness of time-horizons, showing that for long auto-correlation times and short time horizons a phase transition is observable, returning an intermediate or vanishing value of stable cooperation. Finally, we develop an evolutionary algorithm that confirms the crucial role of the time horizon. We find that depending on its value a population of individuals evolves towards a fully cooperative or defecting configuration.

4.2 Main

In order to make analytical progress on eq. 4.1 it is convenient to introduce $q_i(t) := \ln(f_i(t))$. The quantity that agents optimise is simply $g_i = \lim_{t\to\infty}\langle q_i(t)\rangle/t$. The dynamics of q_i can be obtained from eq. 4.1 using Itô calculus. In the case of two agents ($G = 2$) one obtains

$$\langle \dot{q}_1 \rangle = \mu - \frac{\sigma^2}{2} - \frac{\alpha_1}{2} + \frac{\alpha_2}{2}\langle \exp\left(q_2 - q_1\right)\rangle(t) \; . \tag{4.2}$$

In the case $\alpha_i = 0$ one recovers the original case with $g = \mu - \sigma^2/2$. In all the other scenarios, the growth rate of the value geometric mean of agent i, in presence of another agent with resource sharing ratio α_j, $g_{\alpha_i|\alpha_j}$ will therefore depend on both α_i and α_j. In the simple case of two agents, we can treat $g_{\alpha_i|\alpha_j}$ as the entries of a payoff matrix.

In Appendix D.1, we show that the dynamics of $\exp\left(q_2 - q_1\right)$ — the only non trivial term in eq. 4.2 — is ergodic with a well defined stationary distribution, with first moment $\langle \exp\left(q_2 - q_1\right)\rangle_{eq}$. The growth rate $g_{\alpha_i|\alpha_j}$ will therefore be equal to $g_{0|0} + \left(\alpha_j\langle \exp\left(q_2 - q_1\right)\rangle_{eq} - \alpha_j\right)/2$. In appendix D.1.1, we obtain an analytical result for $\langle \exp\left(q_2 - q_1\right)\rangle_{eq}$ that well reproduces the numerical numerical simulations (see D.1).

We aim at finding the (pure-strategy) Nash equilibria and the evolutionary stable strategies. In our context, the relevant question is: given a strategy of the second player α_2, what is the optimal value of α_1? The first non-trivial original result of our letter is that the value of resource sharing that maximize the growth rate $\alpha_1^*(\alpha_2) := \text{argmax}_{\alpha_1} g_{\alpha_1|\alpha_1}$ for a given strategy of the other agent α_2 is always larger than the latter: $\alpha_1^*(\alpha_2) > \alpha_2$ (see Fig. 4.1).

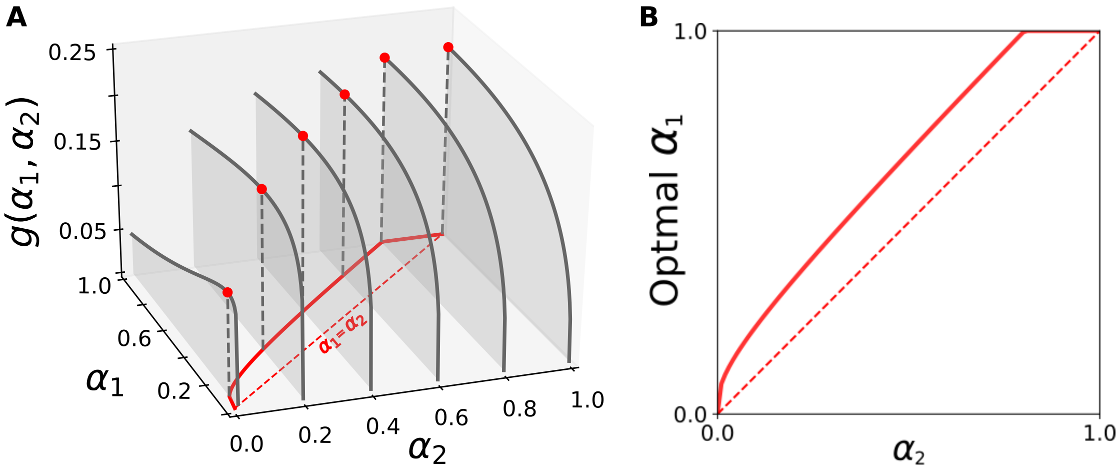


Figure 4.1: The growth rate of individual 1 is maximized for a value of α_1 bigger than the partner'sα_2 for any value of the latter. Panel **A** shows the behaviour of the growth rate of individual 1 as a function of α_1. Every line corresponds to a different value of α_2. The red dashed line marks the diagonal, i.e. $\alpha_1 = \alpha_2$. The solid red line the projection of the maxima of the above curves(red dots) on the plane. It is clear that it always lies above the dashed line, impling $\alpha_{opt} > \alpha_2$. Panel **B** shows the bottom plane of panel **A**, i.e. the optimal value of α_1 as a function of $\alpha - 2$. The, above-diagonal position of the optimal curve implies that in a optimization, evolutionary or learning process both α rapidly converge to 1, making full cooperation evolutionary stable

This mathematical result implies that, contrarily to the mechanism in the tragedy of the commons, each agent has an individual advantage in sharing *more* than the other agent. As a consequence, the evolutionary, adaptive, or learning dynamics maximizing the growth g should lead to a larger and larger level of cooperation, up to the theoretical maximum of full cooperation $\alpha_i = 1$.

The intuition behind this result is that, in fluctuating environments, sharing is akin to investment diversification. As already mentioned, sharing screens the agent from the detrimental effects of fluctuations. In the long-time horizon, the return from this investment (the term $\alpha_2 \langle \exp(q_2 - q_1) \rangle_{eq}/2$) repays its cost (equal to $\alpha_1/2$).

4.2.1 Time and stability transitions

We studied the robustness of this result over four key assumptions: uncorrelated noises (by introducing a non-zero noise correlation equal to ρ), white environmental fluctuation (by introducing a non-zero noise auto-correlation time τ), groups of two agents (by considering arbitrary group sizes G) and infinite time-horizons (by computing the expected growth rate over a finite time horizon T).

By using the unified-colored noise approximation, we obtain in appendix D.1.2 an analytical approximation for the case of infinite time-horizons and arbitrary values of τ and ρ. Positive values of ρ reduce the advantage of increased values of cooperation, but, for any $\rho < 1$ is always more advantageous to share more than the partner, provided that $\tau = 0$. Similar to the effect of correlation, increasing group size G does not alter our result. On the other hand, for $\tau > 0$, full cooperation is not stable anymore and the agents converge to a value $0 < \alpha_{opt} < 1$.

The case of finite time horizons T is not amenable to analytical treatment and it requires relying on numerical simulations to evaluate the average log-return $\langle q_i(T) \rangle / T$. Fig 4.2 shows that two regimes appear separated by a critical time horizon T^*. For $T > T^*$, the system behaves qualitatively as in the infinite time-horizon case: the individual optimizations of the log-average return lead agents to converge to a value $\alpha_{opt} > 0$. For short time horizons ($T < T^*$), defection is more advantageous than cooperation and log-return optimizations lead agents to converge to $\alpha_i = 0$. This result sheds light on the mechanism producing cooperation in our modeling setting: for long time-horizons, the investment in the other agents that cooperation effectively determines has time to returns that overcompensate the costs of cooperation.

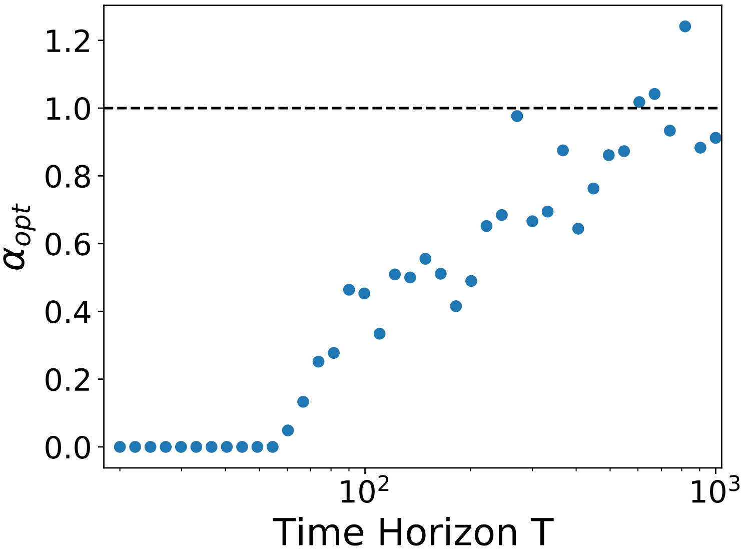


Figure 4.2: Numerical simulations show a phase transition in the optimal α as the time horizon T stretches. For short time horizons $\alpha_i^* < \alpha_j$, bringing the system towards a zero optimal α. Above a threshold T^* a finite value of $\alpha_{opt} > 0$ is stable. It grows by increasing T. Values of $\alpha_{opt} > 1$ are an artifact of the procedure, simply consider them $\alpha_{opt} = 1$.

The results presented above provide a clear mathematical mechanism for the emergence and stability of cooperation in a fluctuating environment.

4.2.2 Evolutionary Alogorithm

We now focus on explicit evolutionary dynamics in a finite population(see Fig. D.4 in the Appendix for a pictorial representation). We consider a population of N agents reproducing with non-overlapping generations at discrete time-steps. Each agent i is characterized by a sharing probability α_i, which is the trait undergoing mutations and selection. Before reproduction, individuals are paired in groups of two and their fitnesses f_i are calculated by integrating eq. 4.1 over a finite time-interval T with initial condition $f_i(0) = 1$. The fitness of each individual is therefore a stochastic variable that depends on the values of α_i of both individuals in the pair. After this step, the pairs are broken up and each individual reproduces proportionally to its fitness value f_i.

As expected from previous results of population genetics in fluctuating environments [40], evolution drives the population to traits that maximize the expected log-fitness. Fig. 4.3 shows the population average values of resource sharing probability α over time. For a short time horizon, defection dominates and the distribution of α is peaked close to 0, with some variance given by mutations and genetic drift. Conversely, when the time horizon is large enough, the vast majority of individuals cooperate, and α peaks close to one.

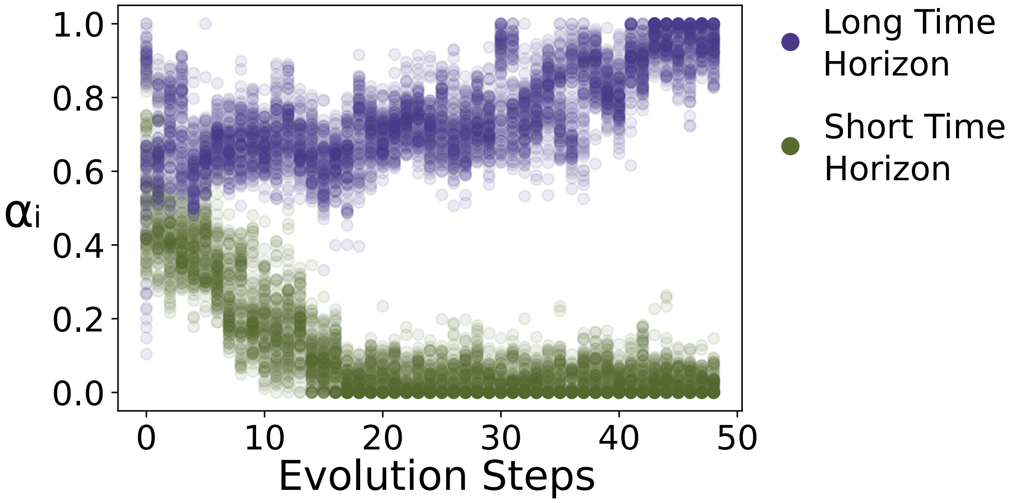


Figure 4.3: The Wright-Fisher model numerical simulation shows the phase transition as the time horizon stretches. Two similar populations, with an initial α distribution with average 0.5, are let evolve with different time horizons ($T_{short} = 20, T_{long} = 1000$). The population with short time horizon evolves towards a distribution peaked in $\alpha = 0$ ehile the long time horizon one, in the opposite direction, towards a $\alpha = 1$.

4.3 Discussion

The emergence and stability of cooperation is a widely discussed topic in ecology, economics and sociology. While it is often observed in nature it is not immediate to theoretically justify its stability. Cheating, in fact, often drives the system away from the cooperative phase -the best for the group - towards a defecting one.

Five mechanisms are known that can overcome the drive towards defection [43]. They mainly rely on the memory of individuals, that behave accordingly to the behaviour of the partners, or on spatial localization and group selection.

We showed that a sixth mechanism exists when the system is representable from a stochastic multiplicative process. Here, when looking at long term returns, the fully cooperative behaviour proves to be the best for growth. As time goes on, in fact, the advantage provided by the of fluctuations become bigger than the one provided by cheating on the partners. We tested this result by exploring different settings, characterizing the stability over noise correlations among individuals, arbitrary group sizes, coloured noises and finite time horizons. We discovered that, while the group size and noise correlations do not change qualitatively the result, time plays an important role in the stability. As time horizon shorten or the auto-correlation time increases we observe a phase transition towards defection, in a prisoner-dilemma like situation.

The existence of this new mechanisms gives new perspectives on the emergence of cooperation. It opens up new possible explanations to existing phenomena and gives the possibility to find new optimal solutions both in ecology and economics.

Appendices

Appendix A

Supplementary Informations to Chapter 1

A.1 Model

We consider a consumer-resource model in presence of cross-feeding, which describes the dynamics of population biomasses n_σ (for $\sigma \in \mathcal{S}$) and resources concentration c_i (for $i \in R$). Changes in population abundance are defined by

$$\frac{dn_\sigma}{dt} = n_\sigma \left(\eta_\sigma \sum_{i \in R} \mathcal{E}^g_{i\sigma} - \delta_\sigma \right) \ . \tag{A.1}$$

where δ_σ is a death term and η_σ is the efficiency of the conversion of energy into biomass. $\mathcal{E}^g_{i\sigma}$ is the energy flux used for species σ to grow from metabolite i. The total energy flux into a cell of type σ is given by $\mathcal{E}^{in}_{i\sigma} = \mathcal{E}^g_{i\sigma} + \mathcal{E}^{out}_{i\sigma}$, where $\mathcal{E}^{out}_{i\sigma}$ are the energy fluxed of secreted metabolites. The associated dynamics of resource concentration c_i is defined by

$$\frac{dc_i}{dt} = h_i(c_i) - \frac{1}{w_i} \sum_{\sigma \in \mathcal{S}} n_\sigma \mathcal{E}^{in}_{i\sigma} + \frac{1}{w_i} \sum_{\sigma \in \mathcal{S}} n_\sigma \mathcal{E}^{out}_{i\sigma} \ , \tag{A.2}$$

were w_i defines the conversion between energy and concentration of resource i. The function $h_i(c_i)$ specify the dynamics of resource concentration in absence of consumers.

We assume that the energy fluxes used for growth are a fraction $1 - \ell_i$ of the total ones: $\mathcal{E}^g_{i\sigma} = (1 - \ell_i)\mathcal{E}^{in}_{i\sigma}$. The energy fluxed from secreted metabolites is given by $\mathcal{E}^{out}_{i\sigma} = \ell_i \sum_{j \in R} D_{ij} \mathcal{E}^{in}_{j\sigma}$. The crossfeeding matrix element D_{ij} defines energy conversion between resource j and resource i. Energy conservation implies $\sum_i D_{ij} = 1$.

The energy flux $\mathcal{E}^{in}_{i\sigma}$ is takes the form

$$\mathcal{E}^{in}_{i\sigma} = w_i \nu_i a_{\sigma i} r_i(c_i) \ , \tag{A.3}$$

where $r_i(c_i)$ is a non-decreasing function of the concentration of resource i, and ν_i is the maximal intake rate of resource i. The elements $a_{\sigma i} \in [0, 1]$ measures the intake rate of metabolite i species σ relative to the maximum ν_i.

We introduce a metabolic tradeoff by considering

$$\frac{\delta_\sigma}{\eta_\sigma} = \frac{1}{\tau}\left(1 + \sum_{j\in R} \chi_j a_{\sigma j}\right)(1+\epsilon_\sigma) \ . \tag{A.4}$$

In the simple setting of $a_{\sigma j} \in \{0,1\}$, the parameter χ_j measure the cost of being able to metabolize metabolite j. The parameter ϵ_σ contributes to the fitness differences between species. In the following we will consider it randomly drawn from a disribution with mean zero and variance σ_ϵ^2.

A.2 Functional attractor

In the eco-evolutionary simulations, we always consider resources and populations changing over a similar timescale. To make analytical progress we approximate the full dynamics with the effective one obtained by assuming timescale separation — i.e. resource concentrations equilibrate faster than the changes in population abundances. We underline that we assume the separation of timescales only as an approximation, for the purpose of predicting analytically the outcomes of the numerical simulations, which are always obtained with explicit resource dynamics.

In this case, one can effectively describe the dynamics of populations as

$$\frac{dn_\sigma}{dt} = n_\sigma \left(\eta_\sigma \sum_{i\in R} a_{\sigma i}\nu_i \frac{h_i^{eff}}{\sum_{\mu\in\mathcal{S}} n_\mu a_{\mu i}\nu_i} - \delta_\sigma\right) \ , \tag{A.5}$$

where $h_i^{eff} = (1-\ell)\sum_{j\in R} B_{ij}h_j w_j$ and the matrix $B = (I-\ell D)^{-1}$. It is useful to notice that, in the limit $\chi \to \infty$ and $h_i^{eff} \to \infty$ (such that the ration h_i^{eff}/χ is finite in the limit) reduces to the model with constant total energy budget [58, 48]. It is known [58] that

$$L(\{n\}) = \sum_\sigma \frac{\delta_\sigma}{\eta_\sigma} n_\sigma - \sum_i h_i^{eff} \log\left(\sum_\sigma \nu_i a_{\sigma i} n_\sigma\right) \ , \tag{A.6}$$

is a Lyapunov function. With our choice for the metabolic trade-off (**??**), such functional can be conveniently rewritten as

$$L(\{n\}) = \sum_\sigma n_\alpha \left(1 + \chi \sum_{j\in R} a_{\sigma j}\right)(1-\epsilon_\sigma) - \sum_i h_i^{eff} \log\left(\sum_\sigma \nu_i a_{\sigma i} n_\sigma\right) \ . \tag{A.7}$$

We then introduce the total population size $N = \sum_{\sigma\in\mathcal{S}} n_\sigma$ and define the functional abundances F_i as

$$F_i = \sum_{\sigma\in\mathcal{S}} a_{\sigma i}\frac{n_\sigma}{N} \ , \tag{A.8}$$

which correspond to the fraction of individuals that are able to metabolize resource i. Interestingly, and surprisingly, when $\epsilon_\sigma = 0$, the Lyapunov function can then be written as function of N and $\{F\}$ alone:

$$L(N, \{F\}) = N \left(1 + \chi \sum_{j \in R} F_j \right) - \sum_{j \in R} h_j^{eff} \log \left(N \nu_j F_j \right) \ . \tag{A.9}$$

The fact that the Lyapunov function depends only on the total biomass and the functional profile already suggest, even if it does not imply, that functional abundances are the relevant variable for the study of community composition.

By minimizing the function over F_i in $[0, 1]$ one obtains

$$F_i^* = \min\{1, \frac{1}{N^*} \frac{h_i^{eff}}{\chi}\} \ , \tag{A.10}$$

where the total biomass is the solution of

$$N^* = \frac{\sum_{j \in R} h_j^{eff}}{\left(1 + \chi \sum_{j \in R} F_j^* \right)} \ . \tag{A.11}$$

These equation can be solve iteratively, starting from $F_i = 1 \ \forall i$ and $N = \sum_{j \in R} h_j^{eff} / (1 + \chi R)$.

In the case with no intrinsic fitness differences ($\epsilon_\sigma = 0$), the equilibrium solutions are identified by equations A.10 and A.11. For a given system, a fraction of resources will be core resources, i.e. shared by everyone $F_i^* = 1$. These core resources are the ones for which $h_i^{eff} \geq \chi N^*$.

A.3 Eco-Evolutionary dynamics

The mutation probability of a preference of resource i in strain μ depends on whether μ consumes or not i. The rate $U_{-,i}$ at which a mutant $\tilde{\mu}$ stops consuming resource i (the parent has $a_i = 1$ and the mutant $a_i = 0$) is constant, independent of i, and equal to U_-. The rate at which a mutant starts consuming a resource i (the parent has $a_i = 0$ and the mutant $a_i = 1$) equals to $U_{+,i} = U_+(P_h F_i + P_{dn})$. The quantity P_h is the probability that an addition happens because of horizontal gene transfer, while $P_{dn} = 1 - P_h$ the probability of "de novo" mutations. The rate of horizontal transfer is proportional to the frequency F_i of that allele in the population, while the rate of a de-novo mutation is independent of i.

The rate at which the resource preference i mutates in strain μ is then equal to

$$W_{\mu,i}^{mut} = b_\mu n_\mu \left(a_{\mu i} U_{-,i} + (1 - a_{\mu i}) U_{+,i} \right) \ , \tag{A.12}$$

where b_μ is the per-capita birth rate on strain μ, which is equal to

$$b_\mu = \eta_\mu \sum_j a_{\mu j} r_j(c_j) \ . \tag{A.13}$$

In theory one could expect a new mutant to have abundance 1. The initial phase of its dynamics is then dominated by demographic stochasticity, with many mutants going to extinction despite having a positive (average) growth rate. In our framework, we do not consider this effect of demographic stochasticity explicitly, but we include it effectively. Since the initial abundance of the mutant $\tilde{\mu}$ is a small fraction of the total population, its stochastic dynamics can be approximated by a stochastic exponential growth. In this regime, the per-capita birth rate of the mutant is given by $(1-\ell)\sum_i a_{\tilde{\mu}i} r_i(c_i^*)$, where c_i^* is the concentration of resource i prior to the mutant arrival. The per-capita death rate of the mutant reads $(1+\chi\sum_i a_{\tilde{\mu}i})(1-\epsilon_{\tilde{\mu}})$. Under the assumption of a stochastic exponential growth the survival probability is given by

$$p_{\tilde{\mu}}^{surv} = 1 - \min\left(1, \frac{(1+\chi\sum_i a_{\tilde{\mu}i})(1-\epsilon_{\tilde{\mu}})}{(1-l)\sum_i a_{\tilde{\mu}i} r_i(c_i)}\right) \ . \tag{A.14}$$

The strain intrinsic fitness values ϵ_σ are independently drawn from a Gaussian distribution with mean 0 and standard deviation ϵ.

By calculating all these quantities for all possible mutations of all existing strains, one obtain the rate of invasion $W_{i\mu}^{inv}$ of a mutant $\tilde{\mu}$ which is obtained by changing the resource preference of strain μ for resource i. The rate of invasion $W_{i\mu}^{inv}$ reads

$$W_{\mu,i}^{inv} = W_{\mu,i}^{mut} p_{\tilde{\mu}}^{surv} \ , \tag{A.15}$$

where the mutant $\tilde{\mu}$ differ from μ in the resource preference i.

We simulate the eco-evolutionary dynamics as a sequence of discrete small time steps Δt. After a step of integration of equations A.1 and A.2 we update the values of $W_{i\mu}^{inv}$, as they depend on strain abundances, and checked whether a mutant appeared. Each mutant, identified by the parent strain μ and a resource i, has probability $W_{i\mu}^{inv}\Delta t$ to invade. If such an event occurs, the new mutant is introduced with an initial relative density equal to 10^{-5}.

If no mutations appear for a long enough time, the ecological dynamics (obtained by integrating equations A.1 and A.2) reach an equilibrium point, identified numerically when the absolute value of the population growth rate is lower than 10^{-4}. If the strain abundances are not changing, also the rates of invasions $W_{i\mu}^{inv}$ are constant in time (until the next successful invasion), and one can use a Gillespie algorithm. The time of the next successful invasion is drawn from an exponential distribution with average $T = 1/\sum_{i\mu} W_{i\mu}^{inv}$. The probability that the new mutant will replace strain μ differing in resource preference i is simply $TW_{i\mu}$.

A.4 Choice of parameters and sensitivity analysis

The results presented in this paper were achieved using generic parameters, whose details can affect the distribution of taxa or relaxation time but not the macroscopic observables that characterize the functional attractor. In order to quantify the convergence to the functional attractor, we measure the discrepancy between the functional composition of the community during its eco-evolutionary trajectory and the functional composition predicted by equations A.10 and A.11. As a measure of the discrepancy, we consider the Kullback-Leibler divergence between the normalized functional profiles

$$D = \sum_i \frac{F_i^*}{\sum_j F_j^*} \log \left(\frac{F_i^*}{\sum_j F_j^*} \frac{\sum_j F_j}{F_i} \right) . \tag{A.16}$$

The divergence D is equal to zero if and only if the functional composition of the community (quantified by the F_i) matches the analytical expectation, i.e. if $F_i = F_i^*$ for all the i.

We considered $\eta_\sigma = \nu_i = w_i = 1$ in eq. A.1 and A.2. These choices do not affect the results, as they do not affect the ecological fixed point and its stability property (up to a rescaling of the abundances and concentrations). The timescale τ was also set to 1, without loss of generality.

In the main text we considered $r_i(c_i) = c_i$. In the Supplementary Materialswe explore the effect of non-linear intake functions $r_i(c)$ by considering a Monod-like form $r_i(c_i) = \mu_{max} c_i/(c_i + K_s)$ with different values of K_i. Figure A.4 shows that the value of K_i has no effect on the functional composition of evolved communities.

The intrinsic fitness on any new mutant was drawn from a Gaussian distribution with mean zero and variance ϵ^2, independently of the fitness of the parent. In the main text we considered $\epsilon = 0.001$. Fig A.1 explores the sensitivity of the results to the magnitude of the noise. Much larger values of noise (of the order 0.1) often disrupt the properties of the manifold. For instance, strains not consuming core resources are still able to survive because of high intrinsic fitness. For intrinsic fitness differences with a width of the order of 10^{-2}, the functional composition converges to the analytical prediction, which becomes more and more accurate as fitness differences decrease.

The strength of cross-feeding ℓ has no effect on convergence to the functional attractor (see Fig. A.2). The cross-feeding matrix D has been chosen following [37]. Entries were extracted according to a Dirichlet distribution, where resources are in three classes. We considered an effective sparsity of $s = 0.1$. The fraction of resources remaining in the same class was $f_s = 0.7$ while the ones going to the waste class is $f_w = 0.28$. The structure of the cross-feeding matrix D does not affect the stationary functional composition of the community. Fig. A.2 compares a fully random D with the ones proposed in ref. [37] and used in the text, observing no difference in the results.

Figure A.3 shows that the outcomes of the evolutionary trajectories are independent of frequencies of the different mutation steps. We varied the (average) total mutation rate $U_{tot} = (U_+ + U_-)/2$, the ratio U_-/U_+ between mutation leading to deletions of resource preferences (with rate U_-) and the ones leading to additions (U_+), and the ratio P_h/P_{dn} between horizontal gene transfer and de-novo mutations. While the total mutation rate, and partially the ratio U_-/U_+, affected the evolutionary trajectories and speed of adaptation, none of these parameters affected the convergence of the functional composition to the predicted attractor.

Both the eco-evolutionary simulations and the analytical approximation are based on the assumption that the metabolic cost is linear in the consumed resources, as expressed in equation **??**. In general, one could assume a non-linear tradeoff [12] that takes the form

$$\frac{\delta_\sigma}{\eta_\sigma} = \frac{1}{\tau} g \left(\chi \sum_{j \in R} a_{\sigma j} \right) (1 - \epsilon_\sigma) \ , \tag{A.17}$$

where $g(z)$ is an arbitrary non-linear, monotonically increasing, function. We considered the outcomes of the evolutionary trajectories in the case of a super-linear cost ($g(z) = (1+z)^2$, in Fig. A.5) and a sub-linear cost ($g(z) = \log(1+z)$, in Fig.A.6). In both scenarios, the functional composition converges to reproducible values, minimally affected by fitness differences. On the other hand, the taxonomic composition is much largely affected by fitness differences. Similarly to the linear metabolic cost functions, some resources correspond to core-functions ($F_i^* = 1$) while the functional occurrences F_i^* of non-core resources are linearly related to the effective influx rates h_i^{eff}. These evolutionary outcomes, obtained under non-linear metabolic costs, confirm the generality of our results beyond the linear metabolic cost case.

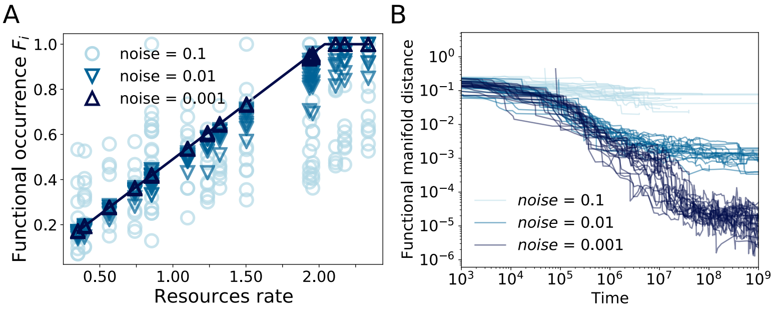


Figure A.1: Noise amplitude on fitness affects the convergence to the functional manifold. 20 realizations for three different amplitudes, $\epsilon = 10^{-3}$ (dark blue), $\epsilon = 10^{-2}$ (blue) and $\epsilon = 10^{-1}$ (light blue). **A** final functional occurrences of the samples. In the case of $\epsilon = 0.1$ the results falls very far from the noiseless theoretical predictions. **B** distance from the manifold as a function of time. The distance is calculated as $d = -\sum_i \tilde{F}_i^* \ln \left(\frac{\tilde{F}_i}{\tilde{F}_i^*} \right)$, where $\tilde{F}_i := \frac{F_i}{\sum_i F_i}$. In all simulations all the other parameters were set to the same values used in the main text.

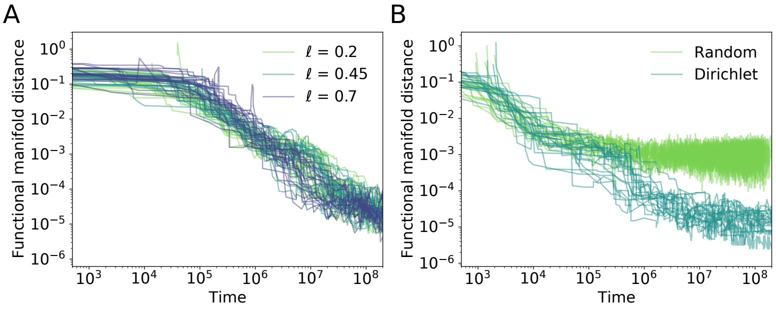


Figure A.2: Cross feeding effect on the convergence to the functional manifold. The shape and intensity of cross-feeding affect neither the final distance from the manifold nor the path used to reach it. 20 realizations for every choice of the parameters are shown. **A** shows the effects of the amplitude of cross-feeding ℓ. Three values are here considered, ($\ell = 0.2$) in light-green, $\ell = 0.45$ in green and $\ell = 0.7$ in blue. **B** Difference of convergence behavior in presence of a random cross-feeding matrix and a Dirichlet distributed matrix. The distance is calculated as in Fig. A.1 In all simulations all the other parameters were set to the same values used in the main text.

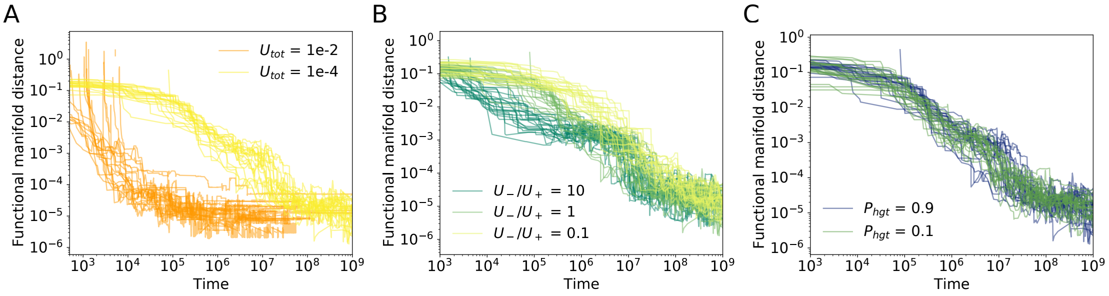


Figure A.3: The choice of evolutionary parameters does not affect the distance from the functional manifold but can modify the path walked to reach it. 20 realizations for every choice of the parameters are shown. **A** shows the effects of the mutation rate U_{tot}. Two values are here considered, a fast mutation rate ($U_{tot} = 10^{-2}$) in orange and a slow one $U_{tot} = 10^{-4}$ in yellow. **B**: effects of the ratio between function loss and function gain rates. Dark green for the case where losing a gene is more probable than gaining it. Light green stands for the even case and yellow for the samples where losing a function is less likely than gaining it. **C**: influence of the probability of gaining new genes via horizontal gene transfer (P_{hgt}) versus spontaneous mutation. The distance is calculated as in Fig. A.1 In all simulations all the other parameters were set to the same values used in the main text.

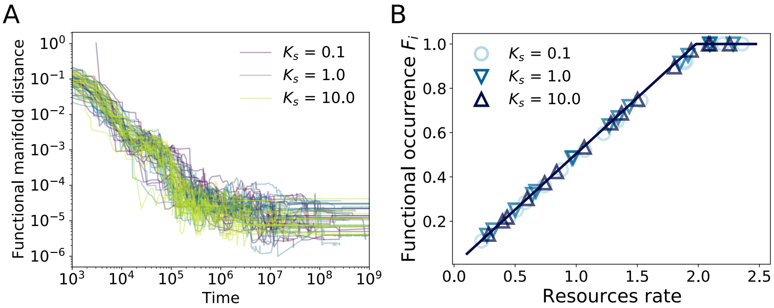


Figure A.4: The properties of the manifold are insensitive on the choice of the function $r_i(c_i)$ of Eq. (1.1), i.e. which species consume the resources does not affect the convergence to the functional manifold. In particular both the linear response ($r_i(c_i) = c_i$) and the Monod response ($r_i(c_i) = \mu_{max}\frac{c_i}{K_s+c_i}$) bring to the manifold with the same behavior. In **A** we show the time evolution of the distance from the theoretical manifold for 20 trajectories for every choice of K_s and in **B** the functional occurrence for one realisation for each K_s. Such a parameter is not determinant in the behaviour of the convergence to the manifold. The constant $\mu_{max} = (2 + K_s)(2 + \chi)$ is chosen to ensure that the growth rate is higher than the death rate at least for some species at the beginning of the dynamics. Such a choice also ensures that all the resources are properly consumed and none of them is growing indefinitely. In all simulations all the other parameters were set to the same values used in the main text.

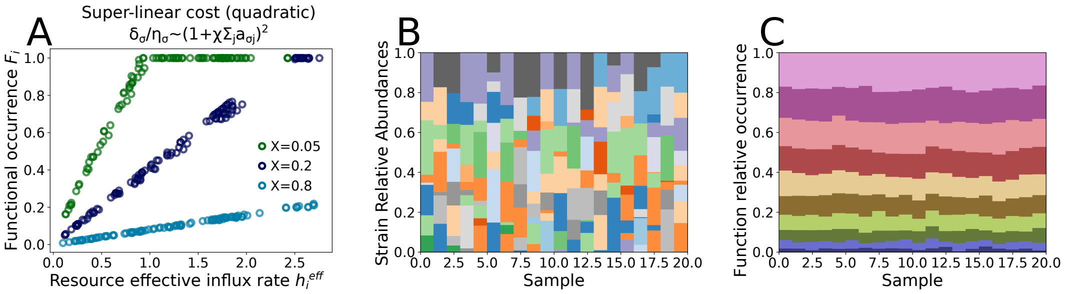


Figure A.5: Evolutionary outcomes under super-linear metabolic cost. All the panels consider a quadratic cost ($g(z) = (1+z)^2$ in eq. A.17). Panel A shows that the functional occurrences F_i^* depend on the effective resource influx rates h_i^{eff} in a similar fashion to what is observed for the linear metabolic cost (see Fig. 1.2). Different points correspond both to different resources and different realizations of the intrinsic fitness values. Similar to the linear cost, increasing the value of the cost per resource χ decreases the number of core resources. The other panels show the taxonomic (panel B) and functional (panel C) composition of different communities evolved in independent environments, characterized by the same effective resources influx rate h_i^{eff} but different intrinsic fitness values ϵ_σ. Panel B shows that the taxonomic composition varies widely across realizations, while the functional composition is much more stable and minimally affected by intrinsic fitness variation (Panel C). A color in panel B represents a strain, fully characterized by a given functional preference $a_{\sigma\cdot}$. Colors in panel C represent different functions. The overall qualitative picture that emerges confirms the results obtained in the main text for linear metabolic costs. In all simulations, all the other parameters were set to the same values used in the main text. Panel B and C were obtained with $\chi = 0.5$.

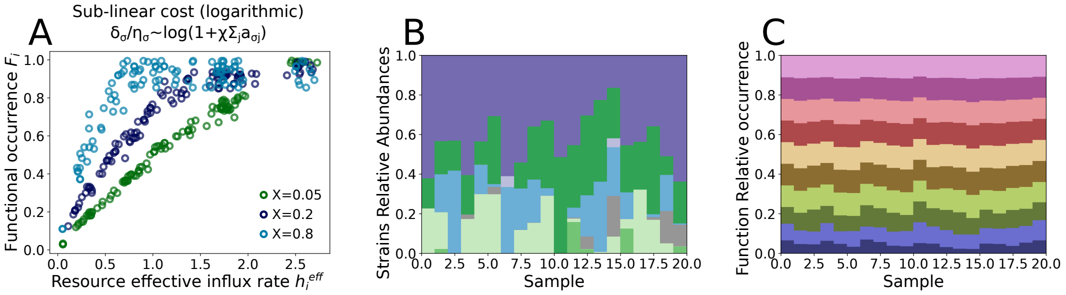


Figure A.6: Same as figure A.5 but with sub-linear metabolic cost. All the panel considers the case of a logarithmic cost ($g(z) = \log(1 + z)$ in eq. A.17).

Appendix B

Supplementary Informations to Chapter 2

B.1 Community structure function

Under the choices explained in the main text, the dynamics of population abundances is defined by

$$\frac{dn_\sigma}{dt} = \eta_\sigma n_\sigma \left(\sum_{i \in R} (1-\ell_i) w_i \nu_i a_{\sigma i} r_i(c_i) - \frac{1}{\tau} \left(1 + \sum_{j \in R} \chi_j a_{\sigma j} \right) (1+\epsilon_\sigma) \right) \ . \tag{B.1}$$

The values of η_σ only contribute to determine the time-scales of the process but do not affect the fixed point or its stability. We therefore limit the calculation to the case $\eta = 1$.

The dynamics, in the limit $\epsilon_\sigma = 0$, can be described in terms of the relative frequences x_σ

$$\frac{dx_\sigma}{dt} = x_\sigma \left(\sum_{i \in R} (1-\ell_i) w_i \nu_i \left(a_{\sigma i} - F_i \right) r_i(c_i) - \frac{1}{\tau} \left(\sum_{j \in R} \chi_j (a_{\sigma j} - F_j) \right) \right) \ , \tag{B.2}$$

where $F_j = \sum_\sigma x_\sigma a_{\sigma j}$. The resources evolve accordingly to

$$\frac{dc_i}{dt} = h_i(c_i) - \frac{1}{w_i} \sum_j N \left(\delta_{ij} - \ell_j D_{ij} \right) w_j \nu_j F_j r(c_j) \ , \tag{B.3}$$

where the total biomass N is the solution of the equation

$$\frac{dN}{dt} = N \left(\sum_{i \in R} (1-\ell_i) w_i \nu_i F_i r_i(c_i) - \frac{1}{\tau} \left(1 + \sum_{j \in R} \chi_j F_j \right) \right) \ . \tag{B.4}$$

Given that in both eq B.3 and B.4 the effect of population abundances only enter thorugh F_i, one could be tempted to write an equation for F_i, which in turns reads

$$\frac{dF_i}{dt} = \sum_{j\in R}\sum_{i\in R}(1-\ell_i)w_i\nu_i\left(F_{ij}-F_iF_j\right)r_i(c_i) - \frac{1}{\tau}\left(\sum_{j\in R}\chi_j(F_{ij}-F_iF_j)\right) , \tag{B.5}$$

that also depends on $F_{ij} := \sum_\sigma x_\sigma a_{\sigma i}a_{\sigma j}$, which quantifies the probability that two pathways occur together in the same individual. The equations therefore do not close in F_i, depending on the full distribution of pathways into individuals.

A similar argument can be more completely formulated by writing the dynamics for the community structure function G. It can be obtained by deriving the dynamics for $e^G = \sum_\sigma x_\sigma \exp(\sum_i k_i a_{\sigma i})$ from eq. B.2 and using $\partial G/\partial t = e^{-G}\partial e^G/\partial t$, from which one obtains

$$\frac{\partial G(\{k\},t)}{\partial t} = \sum_i \left(\frac{\partial G(\{k\},t)}{\partial k_i} - F_i\right)\left((1-\ell_i)w_i\nu_i r_i(c_i) - \frac{\chi_i}{\tau}\right) . \tag{B.6}$$

which corresponds to eq. 2.9.

The term $\partial G(\{k\},t)/\partial k_i - F_i$ depends, in full generality, on k, implying that the other terms $(1-\ell_i)w_i\nu_i r_i(c_i) - \frac{\chi_i}{\tau}$ are equal to zero. This is always true, unless $\partial G(\{k\},t)/\partial k_i$ is independent of k. This is possible if and only if all the individuals have the same value of $a_{\sigma i}$, which in turns imply $\partial G(\{k\},t)\partial k_i = F_i \in \{0,1\}$, therefore determing the first term is zero.

B.2 Stationary solutions

There are two types of resources therefore: the ones that all the individuals are able or not able to metabolize (for which $F_i \in \{0,1\}$) and those for which the stationary concentration reads

$$r_i(c_i^*) = \frac{\chi_i}{(1-\ell_i)w_i\nu_i\tau} . \tag{B.7}$$

We will neglect the resources with $F_i = 0$, and we will call core resources the ones with $F_i = 1$ and non-core resources the others. If $i \in R_c$ is a core resource, if $i \in R_{nc}$ is a non-core one.

Using equation B.3 it is easy to see that the stationary value $F_i^* r_i(c_i^*)$ reads

$$F_i^* r_i(c_i^*) = \frac{\sum_j B_{ij}^{-1} w_j h_j(c_j^*)}{N^*\nu_i w_i} =: \frac{q_i\chi_i}{N^*\nu_i w_i(1-\ell_i)\tau} , \tag{B.8}$$

where $B_{ij} = \delta_{ij} - \ell_i D_{ij}$ and

$$q_i = \frac{\tau(1-\ell_i)\sum_j B_{ij}^{-1} w_j h_j(c_j^*)}{\chi_i} . \tag{B.9}$$

By imposing stationarity in equation B.4 we obtain in turn

$$0 = \tau \sum_i (1-\ell_i)\nu_i w_i F_i^* r_i(c_i^*) - 1 - \sum_j \chi_j F_j^* = \frac{\sum_i \chi_i q_i}{N^*} - 1 - \sum_j \chi_j F_j^* \ , \tag{B.10}$$

from which we obtain

$$N^* = \frac{\sum_i \chi_i q_i}{1 + \sum_j \chi_j F_j^*} \tag{B.11}$$

For non-core resources we can use equation B.7 together with equation B.8 to obtain

$$F_i^* = \frac{q_i}{N^*} \text{ if } i \in R_{nc}, \tag{B.12}$$

Motivated by the fact that the functional abundances are bounded from above to 1, we finally obtain

$$F_i^* = \min\{1, \frac{q_i}{N^*}\} \ , \tag{B.13}$$

The value of q_i is determined by solving eq B.8 for both core and non-core resources. An efficient algorithm to solve these coupled equations computationally is presented in section B.3.

Note that if resources are externally supplied without dilution (or if dilution of resources is negligible compared to their consumption), i.e., if $h_i(c_i^*) = h_i$ equation B.9 becomes a definition and there is no need to find the stationary concentration of resources to determine the resource quality q_i.

B.3 Iterative algorithm to determine the manifold

The algorithm is iterative. It starts by considering all the resources, but one, as non-core. It further assumes that for the core resource i^c the two terms in the $\min(\cdot)$ are equal, i.e. $q_{i_c} = N^*$.

Starting from the 0-th step (evaluated in sequence)

$$\begin{aligned}
c_i^{(0)} &= r_i^{-1}\left(\frac{\chi_i}{(1-\ell_i)w_i\nu_i\tau}\right) \\
q_i^{(0)} &= \frac{\tau(1-\ell_i)\sum_j B_{ij}^{-1} w_j h_j(c_j^{(0)})}{\chi_i} \\
N^{(0)} &= \max_i q_i^{(0)} \\
F_i^{(0)} &= \min\{1, \frac{q_i^{(0)}}{N^{(0)}}\}
\end{aligned} \tag{B.14}$$

Starting from this condition we then evaluate

$$\begin{aligned}
N^{(d+1)} &= \frac{\sum_i \chi_i q_i^{(d)}}{1+\sum_j \chi_j F_j^{(d)}} \\
F_i^{(d+1)} &= \min\{1, \frac{q_i^{(d)}}{N^{(d+1)}}\} \\
R_c^{(d+1)} &= \{i : F_i = 1\} \quad , R_{nc}^{(d+1)} := \{i : F_i < 1\} \\
c_i^{(d+1)} &= r_i^{-1}\left(\frac{\chi_i q_i^{(0)}}{N^{(d+1)}\nu_i w_i (1-\ell_i)\tau}\right) \quad \text{if } i \in R_c \\
c_i^{(d+1)} &= r_i^{-1}\left(\frac{\chi_i}{(1-\ell_i) w_i \nu_i \tau}\right) \quad \text{if } i \in R_{nc} \\
q_i^{(d+1)} &= \frac{\tau(1-\ell_i)\sum_j B_{ij}^{-1} w_j h_j(c_j^{(d+1)})}{\chi_i}
\end{aligned} \tag{B.15}$$

B.4 Solution to the equation defining the functional manifold

We have to solve

$$F_i^* = \min\{1, \frac{q_i}{N^*}\} \ , \tag{B.16}$$

together with

$$N^* = \frac{\sum_i \chi_i q_i}{1+\sum_{j\in R} \chi_j F_j^*} \ . \tag{B.17}$$

It is convenient

Without loss of generality, we can order the resources so that $q_{i-1} \leq q_i \leq q_{i+1}$. We define i_c as the indices corresponding to the last resource with $F_{i_c}^* < 1$ (i.e. $F_{i_c}^* < 1$ and $F_{i_c+1}^* = 1$). We obtain therefore

$$F_i^* = \frac{q_i}{N^*} \quad \text{if } i \leq i_c \text{ and } F_i^* = 1 \text{ otherwise} \ , \tag{B.18}$$

and

$$N^* = \frac{\sum_i \chi_i q_i}{1+\frac{1}{N^*}\sum_{j\leq i_c} q_j \chi_j + \sum_{j>i_c} \chi_j} \ , \tag{B.19}$$

grom which one obtains

$$N^* = \frac{\sum_{j>i_c} \chi_j q_j}{1+\sum_{j>i_c} \chi_j} \ , \tag{B.20}$$

Introducing this expression in eq. B.18 we obtain that i_c is defined by

$$\frac{q_{i_c}}{\sum_{j>i_c} q_j \chi_j}\left(1+\sum_{j>i_c} \chi_j\right) < 1 \quad \text{and} \quad \frac{q_{i_c+1}}{\sum_{j>i_c+1} q_j \chi_j}\left(1+\sum_{j>i_c+1} \chi_j\right) < 1 \ . \tag{B.21}$$

By defining

$$\Xi(i) = q_{i_c}\left(1 + \sum_{j>i_c} \chi_j\right) - \sum_{j>i_c} q_j \chi_j \ , \tag{B.22}$$

the condition of equation B.21 can be rewritten as

$$\Xi(i_c) < 0 \ \text{ and } \ \Xi(i_c + 1) \geq 0 \ . \tag{B.23}$$

B.5 Infinite resource pool $R \to \infty$

It is convenient to define $\eta_i = q_i/\bar{q}$ and $\gamma_i = \chi_i/\bar{\chi}$, where $\bar{q} = \sum_i q_i/R$ and $\bar{\chi} = \sum_i \chi_i/R$. In the limit of large R their joint distribution function is $p(\eta, y)$. We can write the continuous version of eq. B.22 by considering that, in the continuum limit $\eta_i \to \eta$, $\eta_{i_c} \to \eta_c$

$$i \to R \int_0^{\eta_i} dz \ \int_0^{\infty} dy \ p(z, y) =: R\phi(\eta_i) \ , \tag{B.24}$$

and

$$\sum_{j>i} q_j \chi_j = \bar{q}\bar{\chi} \sum_{j>i} \eta_j \gamma_j \to R\bar{q}\bar{\chi} \int_{\eta_i}^{\infty} dz \ \int_0^{\infty} dy \ p(z, y) z y = R\bar{q}\bar{\chi} \left(\int_{\eta_i}^{\infty} dz \ \ p(z) z \langle\gamma\rangle_z \right) \ , \tag{B.25}$$

where $p(z) = \int_0^{\infty} dy \ p(z, y)$ and $\langle\gamma\rangle_z = \int_0^{\infty} dy \ p(z, y) y$. Similarly,

$$\sum_{j>i} \chi_j = \bar{q}\bar{\chi} \sum_{j>i} \gamma_j \to R\bar{q}\bar{\chi} \int_{\eta_i}^{\infty} dz \ \int_0^{\infty} dy \ p(z, y) y = R\bar{\chi} \left(\int_{\eta_i}^{\infty} dz \ \ p(z) \langle\gamma\rangle_z \right) \ , \tag{B.26}$$

Introducing this expressions in eq. B.22 one obtains

$$0 = \eta_c \int_{\eta_c}^{\infty} dz \ \ p(z) \langle\gamma\rangle_z - \bar{q} \int_{\eta_c}^{\infty} dz \ \ p(z) z \langle\gamma\rangle_z \ . \tag{B.27}$$

The equation for the total biomass N^* reads

$$N^* = \frac{R\bar{q}\bar{\chi} \int_{\eta_c}^{\infty} dz \ \ p(z) z \langle\gamma\rangle_z}{1 + R\bar{\chi} \int_{\eta_i}^{\infty} dz \ \ p(z) \langle\gamma\rangle_z} = \frac{\eta_c R\bar{\chi} \int_{\eta_c}^{\infty} dz \ \ p(z) \langle\gamma\rangle_z}{1 + R\bar{\chi} \int_{\eta_i}^{\infty} dz \ \ p(z) \langle\gamma\rangle_z} \ . \tag{B.28}$$

The number of species will read

$$S^* = R\left(1 - \int_{\eta_c}^{\infty} dz \ p(z)\right) \ . \tag{B.29}$$

B.5.1 Linear relation between q and χ

Let us consider $\langle\gamma\rangle_z = 1 - \lambda + \lambda z$. The parameter λ characterizes the correlation between the quality q_i and the cost χ_i.

B.5.2 Exact solution in the uniform case

Let's cosider the case when h are uniformly distributed, i.e. $p(\eta)$ if a uniform distribution between $1 - \sigma_h/2$ and $1 + \sigma_h/2$. Eq. B.27 reads therefore

$$0 = -1 + \eta_c \frac{1+X}{X} + \int_{1-\sigma_h/2}^{\eta_c} dz \, \frac{z}{\sigma_h} \, , \tag{B.30}$$

which is solved by

$$\eta_c = \sqrt{5/4 + \sigma_h + \left(\sigma_h \frac{1+X}{X}\right)^2} - \sigma_h \frac{1+X}{X} \, . \tag{B.31}$$

By definition η_c has to be in the interval $[1 - \sigma_h/2, 1 + \sigma_h/2]$, which occurs if $\sigma_h > 2/(1+X)$. In this regime, the total biomass can be simply obtained as $N^* = \eta_c H/X$, while the richness equals $S = 1 + R\frac{X}{\sigma}\frac{N^*}{H} = 1 + R\eta_c/\sigma_h$. If $\sigma_h < 2/(1+X)$, then $N^* = H/(1+X)$ and $S = 1$.

Appendix C

Supplementary Informations to Chapter 3

C.1 A Mathematical description of Community Functional Composition

We can describe each genome is a vector in the discrete space of functions $g_i = (f_1, f_2, f_3..)$ with $f_i \in \mathbb{N}$. Not all the vectors exist as evolution puts constraints on the shape of these vectors. Since communities are composed by integer number of individuals, the functional composition can assume just values in the lattice generated by the existing vectors $C_f = \sum_i x_i g_{if}$ with $x_i \in \mathbb{N}$. The number of nodes of this lattice is way smaller than the dimension of the abstract space of all possible functional compositions.

C.2 Generating a Virtual Community

To create the virtual community we begin from a genome length distribution $P_v(\ell)$ and attribute to every strain S in the database a probability to be selected proportional to $P(\ell_s)$. We then use such probabilities to select K strains.

We add to the community s_k copies of every selected strain, with $s_k = \frac{N}{n} P(\ell_k)\xi$, where N is the average total number of bacteria in the virtual community and ξ is a random number uniformly distributed in $[0.5, 1.5]$. Defining a vector $\tilde{s}_s$ equal to s_k for $i \in \{k\}$ and 0 otherwise, the family abundances are obtained as $R'_f = \sum_i D'_{is} \tilde{s}_s$. We eventually apply a sampling error to the family abundances by extracting R_f from a normal distribution with mean R'_f and standard deviation $\sqrt{R'_f}$

C.3 Uniqueness of the Solution

This is a speculation on the mathematical reasons supporting the convergence of the algorithm.

Consider the very simplified case of one strain and two genes. Each family scales with the genome length ℓ with a different exponent. Without loss of generality consider the abundance of two genes g and $1-g$ with power laws $a_1\ell^{e_1}$ and $a_2\ell^{e_2}$. Given a we have

$$\frac{a_1\ell^{e_1}}{a_1\ell^{e_1}+a_2\ell^{e_2}}=g \tag{C.1}$$

that univoquely determines the required length of the genome as

$$\ell^*=\left(\frac{ga_2}{a_1(1-g)}\right)^{e_2-e_1}$$

Consider now two strains in the community, with lengths ℓ_1 and ℓ_2 and relative abundances x and $1-x$. We can write Eq. C.1 as

$$\frac{xa_1\ell_1^{e_1}+(1-x)a_1\ell_2^{e_1}}{xa_1\ell_1^{e_1}+xa_2\ell_1^{e_2}+(1-x)a_1\ell_2^{e_1}+(1-x)a_2\ell_2^{e_2}}=g \tag{C.2}$$

This can be written as

$$x=\frac{g(a_1\ell_2^{e_1}+a_2\ell_2^{e_2})-a_1\ell_2^{e_1}}{(1-g)(a_1\ell_1^{e_1}-a_1\ell_2^{e_1})-g(a_2\ell_1^{e_2}-a_2\ell_2^{e_2})} \tag{C.3}$$

The right hand side is positive, leading to a possible solution, for many walues of ℓ_1, ℓ_2. It has two nodes, one in 0 and one in $\ell_1=\ell_2=\ell^*$.

Considering now more than one couple genes we obtain many equations like C.3. The values of the r.h.s. depend on the parameters of the genes, making it unprobable to find a value of ℓ_1 and ℓ_2 that provides the same value to all the r.h.s. of the equations. The only stable points are the nodes, granting the convergence of the method. The node in 0 does create problems as we can see in Fig.3.5, where a peak at low length is detected by the algorithm even though there are no bacteria of that length.

Appendix D

Supplementary Informations to Chapter 4

D.1 Mathematics

Beginning from two SDE with multiplicative noises

$$\begin{aligned}\dot{x}_1 &= \mu x_1 + x_1\sigma\epsilon_1 + \frac{\alpha_2 x_2 - \alpha_1 x_1}{2} \\ \dot{x}_2 &= \mu x_2 + x_2\sigma\epsilon_2 + \frac{\alpha_1 x_1 - \alpha_2 x_2}{2}\end{aligned} \tag{D.1}$$

Where the noise can be correlated both among individuals and in time as described by the following equation

$$< \epsilon_i(t)\epsilon_j(t') >= (\rho(1-\delta_{ij}) + \delta_{ij})\frac{e^{-\frac{|t-t'|}{\tau}}}{2\tau} \tag{D.2}$$

We consider the logarithm of the variables $q := log(x)$ and the difference between the tho logarithmic variables $d := q_2 - q_1$

$$\begin{aligned}\dot{q}_1 &= \mu - \frac{\sigma^2}{2} + \sigma\epsilon_1 + \frac{\alpha_2 e^{q_2-q_1} - \alpha_1}{2} \\ \dot{q}_2 &= \mu - \frac{\sigma^2}{2} + \sigma\epsilon_2 + \frac{\alpha_1 e^{q_1-q_2} - \alpha_2}{2}\end{aligned} \tag{D.3}$$

The equation for $d(t)$ is easily obtained as

$$\begin{aligned}\dot{d} &= \sigma\sqrt{2(1-\rho)}\zeta + \frac{\alpha_1 e^{-d} - \alpha_2 e^{d}}{2} + \frac{\alpha_1 - \alpha_2}{2} \\ < \zeta(t)\zeta(t') > &= \frac{e^{\frac{|t-t'|}{\tau}}}{2\tau}\end{aligned} \tag{D.4}$$

Where we used

$$\begin{aligned} \sigma(\epsilon_1 - \epsilon_2) &= \sqrt{2\sigma^2(1+\rho)}\zeta \\ \sigma(\epsilon_1 + \epsilon_2) &= \sqrt{2\sigma^2(1-\rho)}\zeta =: \sqrt{2}\tilde{\sigma}\xi \end{aligned} \tag{D.5}$$

D.1.1 The $\tau = 0$ case

If no time correlation is present in the multiplicative noises we can exactly find the stationary distribution of d. Defining $f(x) := \frac{\alpha_1 e^{-x} - \alpha_2 e^x}{2} + \frac{\alpha_1 - \alpha_2}{2}$ we have

$$P^*(d) = \frac{1}{z} exp\left(\frac{1}{\sigma^2(1-\rho)} \int_0^d dx f(x)\right) \tag{D.6}$$

This, if $y(t) := e^d(t)$, becomes

$$P^*(y) = \frac{1}{z} y^{\frac{\alpha_1 - \alpha_2}{2\sigma^2(1-\rho)}} exp\left(-\frac{\alpha_1 \frac{1}{y} + \alpha_2 y}{2\sigma^2(1-\rho)}\right) \tag{D.7}$$

We want to find $\langle \dot{q} \rangle \propto \langle e^d \rangle = \langle . \rangle_y = \frac{1}{z}\int_0^\infty dy P^*(y)$. Recalling that $dd = \frac{dy}{y}$ we have that $z = \int_0^\infty \frac{dy}{y} P^*(y)$ and so

$$\langle e^d \rangle(\alpha_1, \alpha_2) = \langle . \rangle_y = \frac{\int_0^\infty dy P^*(y)}{\int_0^\infty \frac{dy}{y} P^*(y)} = \frac{\sqrt{\frac{\alpha_1}{\alpha_2}} BesselK\left(-1 + \frac{\alpha_1 - \alpha_2}{2(\rho-1)\sigma^2}, -\frac{\sqrt{\alpha_1 \alpha_2}}{(\rho-1)\sigma^2}\right)}{BesselK\left(\frac{\alpha_1 - \alpha_2}{2(\rho-1)\sigma^2}, -\frac{\sqrt{\alpha_1 \alpha_2}}{(\rho-1)\sigma^2}\right)} \tag{D.8}$$

By inserting this in Eq.D.3 we have an exact analytical expression for $\langle q_i \rangle(t)$.

D.1.2 The $\tau \neq 0$ case

If the noise is time correlated we can use the Coloured Noise Approximation [46] to find the stationary probability distribution of $d(t)$ reading

$$P^*(d) = \frac{1}{z} exp\left(\frac{1}{\sigma^2(1-\rho)} \int_0^d dx f(x) - \frac{\tau}{2} \frac{f^2(d)}{\sigma^2(1-\rho)}\right)(1 - \tau f'(d)) \tag{D.9}$$

with $\int_0^d dx f(x) = -\frac{\alpha_1 e^{-d} + \alpha_2 e^d}{2} + d\frac{\alpha_1 - \alpha_2}{2} + C$ and $f'(d) = -\frac{\alpha_1 e^{-d} + \alpha_2 e^d}{2}$
if $y(t) := e^d(t)$ we have

$$\begin{aligned} f(y) &= \frac{\alpha_1 \frac{1}{y} - \alpha_2 y}{2} + \frac{\alpha_1 - \alpha_2}{2} \\ \int_0^d dx f(x) &= -\frac{\alpha_1 \frac{1}{y} + \alpha_2 y}{2} + log(y)\frac{\alpha_1 - \alpha_2}{2} + C \\ f'(y) &= -\frac{\alpha_1 \frac{1}{y} + \alpha_2 y}{2} \\ f^2(y) &= \frac{1}{4}\left[\alpha_1\left(1+\frac{1}{y}\right) - \alpha_2(1+y)\right]^2 \end{aligned} \tag{D.10}$$

and this allows us to rewrite the stationary probability distribution as

$$\begin{aligned} P^*(y) = \frac{1}{z} y^{\frac{\alpha_1-\alpha_2}{2\sigma^2(1-\rho)}} exp\left(-\frac{\alpha_1 \frac{1}{y} + \alpha_2 y}{2\sigma^2(1-\rho)} - \frac{\tau}{8}\frac{\left[\alpha_1\left(1+\frac{1}{y}\right) - \alpha_2(1+y)\right]^2}{\sigma^2(1-\rho)}\right) * \\ * \, (1 + \tau\frac{\alpha_1 \frac{1}{y} + \alpha_2 y}{2}) \end{aligned} \tag{D.11}$$

As we can see from Fig.D.1, such a result is in agreement with the numerical simulations for all the different parameter combination examined.

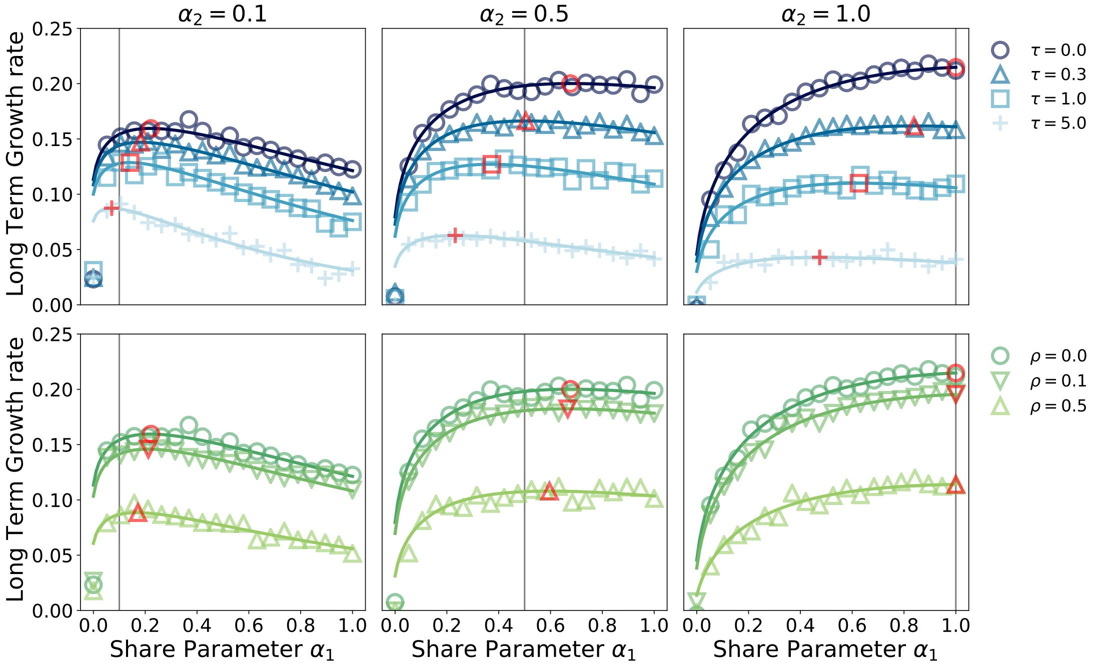


Figure D.1: The typical long term growth rate of individual 1 as a function of how much he shares (α_1). Simulation (markers) are in accordance with the analytical calculations (solid lines). Each column of panels is calculated for a different value of the partner's share α_2. **A**,**B**,**C**: show the behaviour at different values of τ, the decorrelation time of the multiplicative noise (the lighter the higher τ). In red a marker representing the maximum of each curve. We can see that, as τ increases the best choice of α_1 passes from being grater than α_2 to the opposite. **D**,**E**,**F**: show the behaviour at different values of ρ, the correlation between the multiplicative noise of the two individuals (the lighter the higher). In red a marker representing the maximum of each curve. We can see that, as ρ increases the best choice of α_1 always remains greater than α_2.

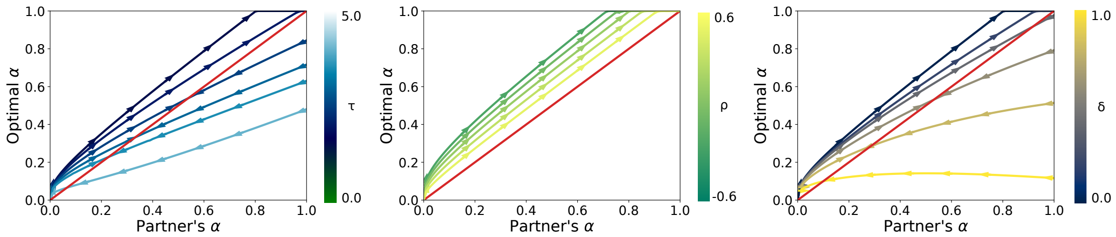


Figure D.2: The behaviour of the optimal value α^* depends on the noise and the co-operation cost. **A** Shows that the auto-correlation time τ affects the stability of the system. By increasing τ (lighter blue) we observe a transition towards a phase where the equilibrium α lies in the interval $(0,1)$ and goes towards 0 as $\tau \to \infty$. In **B** the behaviour with ρ the correlation among the noise of the two individuals. The qualitative behaviour does not change with ρ as the optimal α is always higher than the partner's α. The fully cooperative phase is thus stable for any value of ρ. **C** shows, for the cost of cooperation δ, a similar behaviour to the one observed with τ. As expected, increasing the cost of cooperation (towards yellow) makes the fully cooperative phase unstable, returning an α of equilibrium in $(0,1)$.

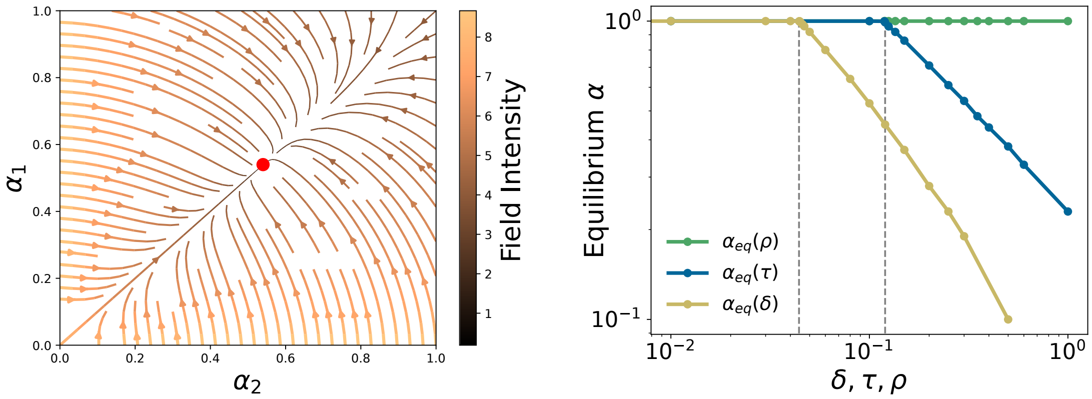


Figure D.3: The equilibrium α of the system presents a phase transition as both τ and δ increase. **A** shows the evolutionary dynamics of the αs at an intermediate value of τ. The vector field converges to a finite value of $\alpha \in (0,1)$. **B** The values of α_{eq} as functions of ρ (green), τ (blue) and δ (yellow)

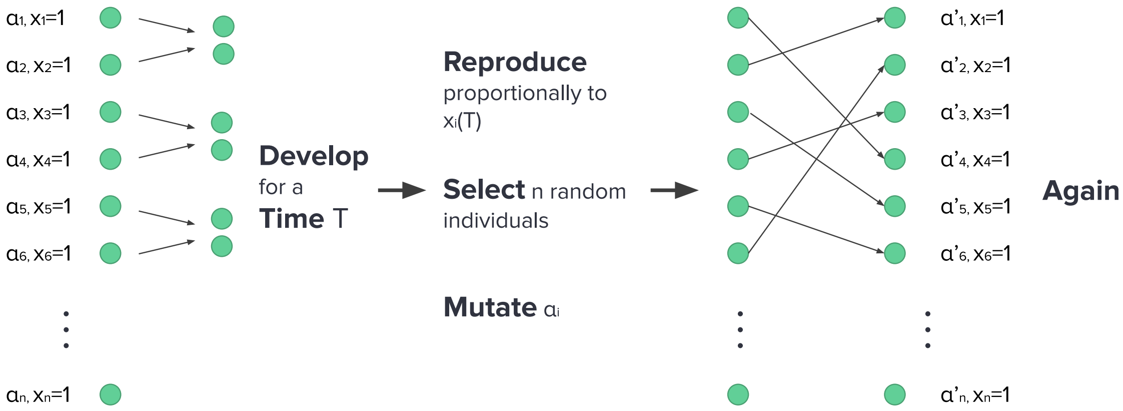


Figure D.4: A sketch representation of the Wright-Fisher model we used to simulate the evolution of cooperation among individuals. A set of n individuals, each characterised by a value of the cooperation parameter α_i and by an initial fitness $x_i = 1$. Individuals are randomly coupled. Each couple is let develop independently in a stochastic multiplicative environment for a time T, cooperating accordingly to α_i,α_j. Individuals reproduce proportionally to the final value of the fitness $x_i(T)$ and n individuals are randomly selected among the offsprings. A mutation is applied to the alpha parameters by extracting a new value of α from a gaussian distribution centered on the parent's α and with width 0.05. Individuals are shuffled and fittnesses are put back to 1. The process is repeated for m times.

Bibliography

[1] Gilbert J. A. and Neufeld J. D. “Life in a World without Microbes”. In: *PLOS* (Dec. 2014).

[2] Melbinger A. and Vergassola M. “The Impact of Environmental Fluctuations on Evolutionary Fitness Functions”. In: *Scientific Repoprts* (2015).

[3] GJ Ackland and ID Gallagher. “Stabilization of large generalized Lotka-Volterra foodwebs by evolutionary feedback”. In: *Physical review letters* 93.15 (2004).

[4] Madhu Advani, Guy Bunin, and Pankaj Mehta. “Statistical physics of community ecology: a cavity solution to MacArthur’s consumer resource model”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2018.3 (2018), p. 033406.

[5] Stefano Allesina. “Going Big”. In: *Unsolved Problems in Ecology.* Ed. by Andrew Dobson, Robert D. Holt, and David Tilman. Princeton Univeristy Press, 2020. ISBN: 9780691199825.

[6] Stefano Allesina and Si Tang. “Stability criteria for complex ecosystems”. In: *Nature* 483.7388 (Feb. 2012), pp. 205–208. ISSN: 0028-0836.

[7] Stefano Allesina et al. “Predicting the stability of large structured food webs.” en. In: *Nature communications* 6 (Jan. 2015), p. 7842. ISSN: 2041-1723.

[8] Britannica. *Bacteria.* URL: `https://www.britannica.com/science/bacteria`.

[9] Guy Bunin. “Ecological communities with Lotka-Volterra dynamics”. In: *Physical Review E* 95.4 (2017), p. 042414.

[10] Catherine Burke et al. “Bacterial community assembly based on functional genes rather than species”. In: *Proceedings of the National Academy of Sciences* 108.34 (2011), pp. 14288–14293.

[11] Stacey Butler and James P O’Dwyer. “Stability criteria for complex microbial communities”. In: *Nature communications* 9.1 (2018), pp. 1–10.

[12] RODRIGO A Caetano, Yaroslav Ispolatov, and Michael Doebeli. “Evolution of diversity in metabolic strategies”. In: *bioRxiv* (2021), pp. 2020–10.

[13] Ted J. Case and Richard G. RG Casten. “Global stability and multiple domains of attraction in ecological systems”. In: *The American Naturalist* 113.5 (May 1979), pp. 705–714. ISSN: 0003-0147. DOI: `10.1086/283427`.

[14] Peter Chesson. “MacArthur’s consumer-resource model”. In: *Theoretical Population Biology* 37.1 (1990), pp. 26–38.

[15] The Human Microbiome Consortium. “Structure, function and diversity of the healthy human microbiome”. In: *Nature* 486 (June 2012), pp. 207–214.

[16] Wenping Cui, Robert Marsland, and Pankaj Mehta. “Effect of Resource Dynamics on Species Packing in Diverse Ecosystems”. In: *Physical Review Letters* 125.4 (July 2020), p. 048101. ISSN: 10797114. DOI: 10.1103/PhysRevLett.125.048101. arXiv: 1911.02595.

[17] Martina Dal Bello et al. “A simple linear relationship between resource availability and microbial community diversity”. In: *bioRxiv* (Sept. 2020), p. 2020.09.12.294660. DOI: 10.1101/2020.09.12.294660.

[18] John P DeLong and Jean P Gibert. “Gillespie eco-evolutionary models (GEM s) reveal the role of heritable trait variation in eco-evolutionary dynamics”. In: *Ecology and evolution* 6.4 (2016), pp. 935–945.

[19] Fant L. Mazzarisi O. Panizon E. and Grilli J. “Cooperation in Stocastic Mutiplicative Environments is Evolutionary Stable”. In: *ArXiv* (2022).

[20] Yaari G. and Solomon S. “Cooperation evolution in random multiplicative environments”. In: *Eur. Phys. J.* (2010).

[21] G. Gause. “The Struggle for Existence”. In: ().

[22] Stefan AH Geritz, Géza Mesze, Johan AJ Metz, et al. “Evolutionarily singular strategies and the adaptive growth and branching of the evolutionary tree”. In: *Evolutionary ecology* 12.1 (1998), pp. 35–57.

[23] Theo Gibbs et al. “Effect of population abundances on the stability of large random ecosystems”. In: *Physical Review E* 98.2 (Aug. 2018), p. 022410. ISSN: 2470-0045. DOI: 10.1103/PhysRevE.98.022410.

[24] Joshua E Goldford et al. “Emergent simplicity in microbial community assembly”. In: *Science* 361.6401 (2018), pp. 469–474.

[25] Benjamin H Good, Stephen Martis, and Oskar Hallatschek. “Adaptation limits ecological diversification and promotes ecological tinkering during the competition for substitutable resources”. In: *Proceedings of the National Academy of Sciences* 115.44 (2018), E10407–E10416.

[26] S.J. Gould. “Kropotkin Was No Crackpot”. In: (1988).

[27] Christopher J. Graves and Daniel M. Weinreich. “Variability in Fitness Effects Can Preclude Selection of the Fittest”. In: *Annual Review of Ecology, Evolution, and Systematics* 48.1 (2017), pp. 399–417. DOI: 10.1146/annurev-ecolsys-110316-022722.

[28] Jacopo Grilli, Tim Rogers, and Stefano Allesina. “Modularity and stability in ecological communities”. In: *Nature Communications* 7 (2016).

[29] Gary W Harrison. “Global stability of predator-prey interactions”. In: *Journal of Mathematical Biology* 8.2 (1979), pp. 159–171.

[30] *HMP Mock Community samples.* URL: `https://www.ebi.ac.uk/metagenomics/studies/MGYS00000300`.

[31] Fant L. Macocco I. and Grilli J. “Eco-evolutionary dynamics lead to functionally robust and redundant communities”. In: *ArXiv* (2021).

[32] J.L. Kelly. “A New Interpretation of Information Rate”. In: *Bell Labs Theoretical Journal* 35 (July 1956).

[33] Thomas Liebmann, Stefan Kassberger, and Martin Hellmich. “Sharing and growth in general random multiplicative environments”. In: *European Journal of Operational Research* 258.1 (2017), pp. 193–206. ISSN: 0377-2217.

[34] Stilianos Louca, Laura Wegener Parfrey, and Michael Doebeli. “Decoupling function and taxonomy in the global ocean microbiome”. In: *Science* 353.6305 (2016), pp. 1272–1277.

[35] Stilianos Louca et al. *Function and functional redundancy in microbial systems.* June 2018. DOI: `10.1038/s41559-018-0519-1`.

[36] Robert Macarthur and Richard Levins. “The Limiting Similarity, Convergence, and Divergence of Coexisting Species”. In: *The American Naturalist* 101.921 (Sept. 1967), pp. 377–385. ISSN: 0003-0147. DOI: `10.1086/282505`.

[37] Robert Marsland, Wenping Cui, and Pankaj Mehta. “A minimal model for microbial biodiversity can reproduce experimentally observed ecological patterns”. In: *Scientific Reports* 10.1 (Dec. 2020), p. 3308. ISSN: 2045-2322. DOI: `10.1038/s41598-020-60130-2`.

[38] Robert Marsland et al. “Available energy fluxes drive a transition in the diversity, stability, and functional structure of microbial communities”. In: *PLoS Computational Biology* 15.2 (Feb. 2019). ISSN: 15537358. arXiv: `1805.12516`.

[39] Robert M May. “Will a Large Complex System be Stable?” In: *Nature* 238.5364 (Aug. 1972), pp. 413–414. ISSN: 0028-0836. DOI: `10.1038/238413a0`.

[40] Andreas Mayer et al. “Transitions in optimal adaptive strategies for populations in fluctuating environments”. In: *Phys. Rev. E* 96 (3 Sept. 2017), p. 032412. DOI: `10.1103/PhysRevE.96.032412`. URL: `https://link.aps.org/doi/10.1103/PhysRevE.96.032412`.

[41] Jaina Mistry et al. “Pfam: The protein families database in 2021”. In: *Nucleic Acids Research* 49.D1 (Oct. 2020), pp. D412–D419. ISSN: 0305-1048. DOI: `10.1093/nar/gkaa913`. eprint: `https://academic.oup.com/nar/article-pdf/49/D1/D412/35363969/gkaa913.pdf`. URL: `https://doi.org/10.1093/nar/gkaa913`.

[42] Erik van Nimwegen. “Scaling laws in the functional content of genomes”. In: *Genome Analysis* 19 (Sept. 2003), p. 479.

[43] Martin A. Nowak. “Five rules for the evolution of cooperation”. In: *Science* (2006).

[44] Peters O. “The ergodicity problem in economics”. In: *Nature Physics* 15.1 (Dec. 2019), p. 1216. ISSN: 2041-1723. DOI: 10.1038/s41567-019-0732-0.

[45] S. Ohno. “The reason for as well as the consequence of the cambrian explosion in animal evolution”. In: *Journal of Molecular Evolution* 44 (1997). DOI: https://doi.org/10.1007/PL00000055.

[46] Jung P. and Hanggi P. “Dynamical Systems: a Unified Colored-Noise Approximation”. In: *Physical Review A* (1987).

[47] Ole Peters and Alexander Adamou. “An evolutionary advantage of cooperation”. In: (2018). arXiv: 1506.03414 [nlin.AO].

[48] Anna Posfai, Thibaud Taillefumier, and Ned S. Wingreen. “Metabolic Trade-Offs Promote Diversity in a Model Ecosystem”. In: *Physical Review Letters* 118.2 (Jan. 2017), p. 028103. ISSN: 10797114. DOI: 10.1103/PhysRevLett.118.028103.

[49] Daniel J. Rankin, Katja Bargum, and Hanna Kokko. “The tragedy of the commons in evolutionary biology”. In: *Trends in Ecology and Evolution* 22.12 (2007), pp. 643–651. ISSN: 0169-5347. DOI: https://doi.org/10.1016/j.tree.2007.07.009. URL: https://www.sciencedirect.com/science/article/pii/S0169534707002741.

[50] O. Rivoire and S. Leibler. “The Value of Information for Populations in Varying Environments”. In: *J Stat Phys* (2011).

[51] Nayfach S. and Pollard K.S. “Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome”. In: *Genome Biology* (2015).

[52] *SAMEG324767*. URL: https://www.ebi.ac.uk/biosamples/samples/SAMEG324767.

[53] Carlos A. Serván et al. “Coexistence of many species in random ecosystems”. In: *Nature Ecology & Evolution* 2.8 (Aug. 2018), pp. 1237–1242. ISSN: 2397-334X. DOI: 10.1038/s41559-018-0603-6.

[54] Microbiology Society. *Bacteria*. URL: https://microbiologysociety.org/why-microbiology-matters/what-is-microbiology/bacteria.html.

[55] Reed M. Stubbendieck, Carol Vargas-Bautista, and Paul D. Straight. “Bacterial Communities: Interactions to Scale”. In: *Frontiers in Microbiology* 7 (2016), p. 1234. ISSN: 1664-302X. DOI: 10.3389/fmicb.2016.01234.

[56] Samir Suweis, Jacopo Grilli, and Amos Maritan. “Disentangling the effect of hybrid interactions and of the constant effort hypothesis on ecological community stability”. In: *Oikos* 123.5 (May 2014), pp. 525–532. ISSN: 00301299. DOI: 10.1111/j.1600-0706.2013.00822.x.

[57] Edward O. Thorp. “The Kelly Criterion in Blackjack Sports Betting, and the Stock Market”. In: 2008.

[58] Mikhail Tikhonov. “Community-level cohesion without cooperation”. In: *Elife* 5 (2016), e15747.

[59] David Tilman. “A consumer-resource approach to community structure”. In: *American Zoologist* 26.1 (1986), pp. 5–22.

[60] Vox. *All life on Earth, in one staggering chart.* URL: `https://www.vox.com/science-and-health/2018/5/29/17386112/all-life-on-earth-chart-weight-plants-animals-pnas`.

[61] Wade W. “Unculturable bacteria–the uncharacterized organisms that cause oral infections”. In: *J R Soc Med* (2002). DOI: `doi:10.1258/jrsm.95.2.81`.

[62] Marcu A et al. Wishart DS Feunang YD. “HMDB 4.0 — The Human Metabolome Database for 2018”. In: *Nucleic Acids Res* (Jan. 2018).