



**HAL**  
open science

# Evolution of Collective-level Darwinian Properties

Guilhem Doucier

► **To cite this version:**

Guilhem Doucier. Evolution of Collective-level Darwinian Properties. Ecology, environment. Université Paris sciences et lettres, 2019. English. NNT : 2019PSLET047 . tel-02917058

**HAL Id: tel-02917058**

**<https://pastel.hal.science/tel-02917058>**

Submitted on 18 Aug 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



**THÈSE DE DOCTORAT**

**DE L'UNIVERSITÉ PSL**

Préparée à l'École supérieure de physique  
et de chimie industrielles de la ville de Paris

**De l'Évolution des Propriétés Darwiniennes Collectives**

Soutenue par

**Guilhem Doucier**

Le 2 Décembre 2019

École doctorale n°474

**Frontières de l'Innovation  
en Recherche et Educa-  
tion**

Spécialité

**Ecologie, évolution et bi-  
ologie environnementale**

Composition du jury :

Hélène Morlon Directrice de Recherche, ENS-CNRS	<i>Présidente</i>
Sylvain Gandon Directeur de Recherche, CNRS- Université de Montpellier	<i>Rapporteur</i>
Joachim Krug Professeur, University of Cologne	<i>Rapporteur</i>
Guillaume Martin Chargé de Recherche, CNRS- Université de Montpellier	<i>Examineur</i>
Paul B. Rainey Professeur, ESPCI Paris	<i>Directeur de thèse</i>
Silvia De Monte Chargée de Recherche, ENS-CNRS	<i>Co-Directrice de thèse</i>



Thèse de Doctorat de l'Université Paris Sciences et Lettres  
Laboratoire Génétique de L'Évolution, École Supérieure de Physique et de Chimie Industrielles  
Laboratoire Écologie-Évolution Mathématique, École normale supérieure

---

# On the Evolution of Collective-Level Darwinian Properties

---

Soutenue par Guilhem DOULCIER

*Composition du Jury:*

Pr. Paul B. RAINEY	<i>Directeur de Thèse</i>	Dr. Hélène MORLON	<i>Présidente</i>
Dr. Silvia DE MONTE	<i>Directrice de Thèse</i>	Dr. Sylvain GANDON	<i>Rapporteur</i>
Pr. Amaury LAMBERT	<i>Encadrant</i>	Pr. Joachim KRUG	<i>Rapporteur</i>
		Dr. Guillaume MARTIN	<i>Examineur</i>
		Dr. Philippe NGHE	<i>Examineur</i>

December 2nd, 2019



## Acknowledgements

When I first knocked on Silvia's door six year ago during my bachelor, it was really on the tantalising promise to know more about evolution, complex systems and game theory. Little did I know that it would be the start of a wonderful collaboration that would lead us from marvelling at *Dictyostelium* aggregation, to the mathematical wonders of adaptive dynamics and up to the most conceptual consideration about the nature of life.

Thus, when I later heard Paul at a seminar, I was deep into game theory. At the time, my mind was (rightfully) blown by the cheat as first propagule hypothesis that he presented based on his experiments on *Pseudomonas*. He was kind enough to welcome me in his brand-new Parisian lab where I hoped experimental evolution could use the help of a theoretician like me.

I could not have asked for better PhD thesis directors than the two of them. I spent three extraordinary years, constantly surrounded by fascinating people and ideas. Their mentorship, high standards, and willingness to discuss and work with me at length and in depth is the reason why I was able to produce the document you are holding in your hands.

Amaury gave the best math lectures I ever had the pleasure to attend, to the point I switched my major from biology to mathematics during my master. I owe a lot to his mentorship, support and his patience with fledgling mathematicians. It was a great pleasure, and an honour to be able to teach the practical course supporting his famous *Ce qu'un biologiste ne peut ignorer* lecture during my time as a graduate student.

This work has been funded by the *Origines et conditions d'apparition de la vie* programme in *Paris Sciences et Lettres* University (ANR-10-IDEX-001-02). OCAV is like no other place I ever saw in academia: it offers a forum on the origins of life that gather all specialities from social sciences and philosophy to chemistry and astrophysics. . . and it works really well. The long discussions I had both during and after seminars with Stéphane, Ludovic, Philippe, Alex, Heng and Cyrille were highlights of my time as a graduate student. I also thank the École Doctorale 474 for additional training and funding and my thesis committee members Ludovic and Ivan.

I would like to thank all the members of the Laboratoire de Génétique de l'Évolution in ESPCI, Maxime, Steven and Clara as well as the members of the Écologie-Évolution Mathématiques team in ENS, Régis, Bernard, Minus, Frantz, Sandrine, Stéphane, Éric, Jacques, and my fellow grad students and postdocs: Boris, Clara, Elsa, Phuong, Benjamin, Mathieu, Antonin and Philippe. A dream team of like-minded individuals with whom time stops and I could discuss seemingly for ever without noticing it. Thank you.

I am grateful to the people in ESPCI working on inspiring millifluidic technology: Jérôme, Jean, Denis, Laurent, Jean-Baptiste, Ankur and Wilfried, as well as the ever available administrative staff Hélène and Isa. I am also grateful to my collaborators on *Pseudomonas*, Philippe and Daniel who did the experiments from Chapter 3 and Katrin and Caroline who pioneered them. I also sincerely thank the visiting scholars Ben, Will and Pierrick, the teaching staff at ENS, Andréa, Morgane, Anne, and of course, Pierre who mentored me as a student, and as a teaching assistant.

This thesis owes a lot to the people who welcomed me in science: Yann in Cadarache, Claude, Sonia and Vincent in Montpellier, Matt and Simon in Tucson, Daniel in New Zealand. Thanks. I will strive to be as welcoming as you with my own interns, as I started with Maël and Éléonore.

I am grateful, of course, to my people of the long-hair phenotype: François and Benjamin who pioneered the shift to mathematics and my dear friend Maxime with whom I shared the many highs and lows of student life. I am also thankful to my grad student fellows and flatmates Guillaume, Adam, Lambert, Anca, and François as well as the game night regulars, my hosts in Montpellier, Denis and Jacqueline, my hosts in Paris, Gérard and Agnès, Anne for the Oulipian rendition of my work and all the other people I am forgetting right now.

In a sense, this thesis might not have existed without Fabrice, teacher extraordinaire from the Collège-Lycée André Chamson (Le Vigan) who stoke the love of life sciences in generations of students and without whom I might not even have thought to come to Paris.

Last, but not least, I would like to thank my dearest parents Éric and Monique, as well as my sister Mayane. Through joy and sorrow, they taught me all that really matters in the end.

— Paris, 1st December 2019.

## Abstract

LIFE HAS A NESTED STRUCTURE where lower level entities are embedded in higher level collectives (genes in chromosomes, organelles in cells, cells in organisms, organisms in eusocial groups). All levels are subject to evolution by natural selection. This arises from the fact that at each level the focal entities are Darwinian, that is, they are discrete and vary one to another, they replicate and give rise to offspring that resemble parental types. The emergence of a new level of organisation is a relatively rare event in the history of life, and requires the *de novo* evolution of level-specific properties that allow the new level of organisation to participate directly in the process of evolution by natural selection. In this manuscript I explore, using mathematical models, the idea that Darwinian properties can be exogenously imposed (scaffolded) by the environment. I show how natural selection can build upon those scaffolded properties to promote the emergence of endogenous traits underpinning collective-level reproduction and heredity.

# Contents

<b>Contents</b>	<b>iii</b>
<b>Introduction</b>	<b>1</b>
<b>1 Historical Perspectives</b>	<b>7</b>
<b>2 Darwinian Properties</b>	<b>15</b>
<b>3 Neutral Diversity in Experimentally Nested Populations</b>	<b>31</b>
3.1 Modelling Nested Population Dynamics . . . . .	32
3.2 Modelling Neutral Diversity . . . . .	37
3.3 Artificial selection of droplets . . . . .	47
3.4 Discussion . . . . .	49
Appendix . . . . .	52
<b>4 Locating Mutations in Collective-level Genealogies</b>	<b>67</b>
4.1 The cheat as first propagule hypothesis . . . . .	68
4.2 Establishing collective level genealogies . . . . .	70
4.3 Survival probability estimation . . . . .	73
4.4 Propagating sequencing information to the full genealogy . . . . .	79
4.5 Discussion . . . . .	81
<b>5 From Particle Traits to Collective-level Demography</b>	<b>85</b>
5.1 Evolution of collective survival and reproduction . . . . .	86
5.2 Trade-off between survival and reproduction . . . . .	91
5.3 Particle dynamics . . . . .	95
5.4 Collective Population Dynamics . . . . .	100
5.5 Migration between collectives . . . . .	104
5.6 Discussion . . . . .	106
Appendix . . . . .	111
<b>6 An Ecological Recipe for the Evolution of Collective-level Heredity</b>	<b>117</b>
6.1 Results . . . . .	120
6.2 Discussion . . . . .	131
6.3 Stochastic Model . . . . .	135
6.4 Lotka-Volterra deterministic particle ecology . . . . .	141
<b>Conclusion</b>	<b>151</b>
<b>Bibliography</b>	<b>153</b>

# Introduction

“Il est question ici d’hérédité et de reproduction. Il est question des transformations qui ont progressivement modifié la manière de considérer la nature des êtres vivants, leur structure, leur permanence au fil des générations.”

— FRANÇOIS JACOB, *La Logique du Vivant: Une histoire de l’hérédité* (1970)

THE PRESENT MANUSCRIPT CONCERNS the mechanisms by which collectives of entities become the subject of evolution by natural selection in their own right. Before getting to this precise question though, it is necessary to carefully lay out a series of ideas and concepts about the structure of the “living”, and the nature of evolutionary processes.

That everything changes and nothing stands still is a commonplace observation at least as old as ancient antiquity. In this context, living systems differ from inert systems in the *way* they change, and by the *organisation* that results from those changes. In two words: *mechanisms* and *patterns*. In the same way, evolutionary biology is the science that strives to, on the one hand, reconstruct the history of life and, on the other hand, describe the rules that govern the long-term dynamics of living systems.

I start by defining life from this point of view. Is considered living any system that partakes in a specific kind of change — a specific kind of evolutionary dynamic. Preventing this statement becoming circular requires a clear definition of what constitutes “biological” evolutionary dynamics: namely, evolution by the process of natural selection. Establishing “a recipe for evolution by natural selection” — or, more formally, a set of minimal conditions that must be verified by a collection of objects to partake in evolution by natural selection — has been at the centre of evolutionary biology from the onset (Godfrey-Smith, 2009). In the closing paragraph of the *Origin of species*, Darwin (1872) advocates that the diversity of the living world is a consequence of a “struggle for life”, summarized by three key phenomena: “*Growth with reproduction*”, “*Inheritance*” and “*Variability*”. A modern version of these properties comes from a seminal article by Lewontin (1970), one century later. A population of entities is said to be *Darwinian*, and thus constitutes a unit of evolution (or unit of selection, depending on the authors), if it exhibits the following *Darwinian properties*:

**Phenotypic Variation** The entities within the populations are different from one another.

**Differential Reproduction** The entities are able to produce offspring, the number of offspring is dependent to some extent on the phenotype.

**Phenotypic Heritability** The phenotype of the offspring resembles the parental phenotype.

A critical feature of this paradigm is its *abstract* nature (Okasha, 2006): it refers to a collection of objects and properties without explicit requirement on their material support, either in terms of chemical composition or energetic balance. Notably, this definition does not rely on genetics: its first formulation famously predates the discovery of the law of heredity in diploids (Mendel, 1866) or the DNA molecule as a support for heritable information (Watson and Crick, 1953a). However, those simple concepts alone lead to a very rich range of falsifiable predictions about the fate of living systems. Above all, they offer a causal mechanism that explains the apparent purposefulness observed in the organisation living beings, why they seem adapted to their environment, to survive and to reproduce: their characteristic of *teleonomy* (Monod, 1970). As a consequence, they bring an invaluable contribution to the study of the emergence of life and the apparition of biological complexity in the solar system and beyond. If it is near to impossible to predict the diversity of shapes and metabolisms that life could display somewhere else in the universe, chances are that it would have to present Darwinian properties for us to recognise it as such.

Biological life on earth is organised in various, roughly nested scales: genes, chromosomes, cells, organism, populations, communities... All these entities can manifest to a greater or lesser extent Darwinian properties and thus participate in the process of evolution by natural selection. Mammals, for instance, present heritable phenotypic variation in their reproductive output: they undeniably form a Darwinian population. However, this is also the case for their cells. A major difficulty in studying evolution lies in deciding (explicitly or implicitly) the relevant units of evolution (i.e., the relevant Darwinian population) that explain the observed behaviours, in particular when conflicts arise between levels (Maynard Smith and Brookfield, 1983). Cell apoptosis cannot be predicted when considering cells as Darwinian entities, and requires treating the individual as a whole to explain its evolutionary origin. Conversely, cancerous cell lineages cannot be conceived when treating the individual as the Darwinian entity, but are a natural consequence of the Darwinian nature of the cells. Natural selection acts simultaneously on all levels of organisation: untangling their effects is complex. However, in most situations, the gene-centred view that treats genes as the ultimate cause (and all others levels as by-products) yields robust predictions (Dawkins, 1976).

Naturally following the observation that life is organised in many levels, is the question of their origin. Evolutionary biologists have named the emergence of these new levels *Major Evolutionary Transitions* (Szathmáry and Maynard Smith, 1995) or evolutionary transitions in individuality (Buss, 1987; Michod, 2000). Staple examples include the emergence of chromosomes from genes, of eukaryotic cells from prokaryotic ancestors and multicellular organisms from individual cells.

It must be stressed that the problems of the emergence of life and of the origins of Darwinian properties are essentially the same. It is expected

that a transition in individuality might more likely arise from an existing Darwinian population than through abiogenesis: there are more than twenty-five documented examples of transition to multicellularity, but only one origin of life. Nonetheless, each major evolutionary transition consists in the rise of a new “living” population. To illustrate this, consider an entity that presents some qualities of living systems such as metabolism or growth. If this entity is incapable of reproduction it cannot be classified as alive (Jacob, 1970). A bacterial lawn could expand to recover most of the Earth’s surface; it would still be non-living at the level of the collective of cells. Even if its components are alive, it lacks the ability to beget collectives in its likeness —a *teleonomic apparatus* in the sense of Monod (1970). Overall, without major transitions, life would not have changed beyond early self-replicating chemistry.

However, recognising the need for the emergence of collective-level Darwinian properties does not constitute a mechanism to explain their origin. Natural selection at the level of collectives cannot be invoked as the initial cause: it requires the very Darwinian properties that need explaining (Black et al., 2019). Moreover, it is impossible to assume that Darwinian properties are simply transferred from the lower to the higher level (Griesemer, 2001). Indeed, collective-level Darwinian processes rely on mechanisms that are qualitatively different from their counterparts at the lower level: for instance, the reproduction of multicellular organisms involves a developmental process that cannot be reduced to the simple reproduction of cells. Overall, what requires explanation is “how Darwinian properties might emerge from non-Darwinian entities, and therefore by non-Darwinian means” (Black et al., 2019).

To address this issue, this manuscript extensively uses a framework for the description of Darwinian populations due to Godfrey-Smith (2009) and centred around the idea that Darwinian properties can be quantified. A Darwinian population in a *minimal* sense is “a collection of causally connected individual things in which there is variation of character, which leads to differences in reproductive output, and which is inherited to some extent” (Godfrey-Smith, 2009, p. 39). This minimal level is shared by all Darwinian populations. A subset of those minimal populations define the *paradigmatic* populations, corresponding where the Darwinian dynamics is at its clearest, and giving rise to complex and adapted structures (Godfrey-Smith, 2009, p. 41). Conversely, at the edge of the minimal concept are the *marginal* populations. Marginal populations do not possess exactly the minimal properties, but approximate them to some extent. As a consequence, they can exhibit behaviours resembling the one of minimal Darwinian populations (Godfrey-Smith, 2009, p. 42). Overall populations can be ordered on a “Darwinian spectrum” that goes from non-Darwinian, to marginal, minimal and paradigm.

This distinction alleviates part of the difficulties. If there is a mechanism that can promote the emergence of *marginal* Darwinian properties at the collective-level, then natural selection can be invoked as a process for their further evolution, and potentially their refinement toward *paradigm*-level. While, as stated before, simple transfer of Darwinian properties from lower to higher level is not possible, a first way collectives can gain marginal Darwinian properties is by co-opting lower level traits. Examples include the co-option of ancestral cell-cycle regulation mechanisms in volvocine green algae for the formation of groups via cell–cell adhesion (Hanschen et al., 2016),

and the co-option of ancestral apoptosis mechanisms in the experimentally derived snowflake yeast for the fragmentation of cell clusters (Ratcliff et al., 2012). However, co-option might not always be achievable from the onset, considerably increasing the challenge of a transition.

The main hypothesis explored in this manuscript is a second way collectives can gain *marginal* Darwinian properties, by having them exogenously imposed by the environment, a phenomenon called *ecological scaffolding* (Black et al., 2019). This process does not *a priori* require co-opting of existing traits, but relies first on population structure. Indeed, under specific conditions, that will be detailed, the existence of resource patches and limited dispersal between patches can bestow marginal Darwinian properties upon collectives of unwitting individuals. As a consequence, natural selection can act at the level of the collectives leading to refinement of collective-level Darwinian properties, pushing collectives toward the *paradigm* end of the spectrum.

In this work, I contribute to the search for general mechanisms underlying the emergence of new levels of biological organisation in three specific settings: neutral variation patterns in nested populations, the emergence of reproduction in early multicellular organisms, and the evolution of heredity at the level of communities. This is a theoretical work that uses a diversity of modelling approaches, mainly stochastic processes, Bayesian statistics and dynamical systems. Nonetheless, I have tried as much as possible to avoid a purely formal exercise: all models were defined in relation to ongoing developments in experimental evolution. Experiments conducted by my colleagues both sparked my questioning, and inspired my answers.

## Outline

This manuscript is structured in six chapters, the first two give historical (Chapter 1) and conceptual (Chapter 2) context. The four remaining chapters describe new results and progress through the Darwinian properties in order: Variation (Chapter 3), Reproduction (Chapters 4 and 5) and Heredity (Chapter 6). I end with a brief concluding discussion. The content of the chapters is detailed below.

- **Chapter 1 - Historical Perspectives** (p. 7).

This chapter is dedicated to briefly put Darwinian properties in a more general context. Sketching the genealogy of the ideas of variation, reproduction and heredity through time, from the Enlightenment's concepts of natural history to the separation of life from non-life and the transformists ideas of the 19th century, and finally the new synthesis of the 20th century and the realisation of the abstract nature of Darwinian processes.

- **Chapter 2 - A Primer on Darwinian properties** (p. 15).

This chapter presents in details the conceptual framework used in the manuscript. Namely, a mechanistic view of evolution, the concept of Darwinian Properties and their arrangement in a Darwinian space, and finally, the concept of ecological scaffolding. A toy-model is presented

to build intuition about a simple three dimensional Darwinian space (heredity, selection, competition) and link it to well-known biological patterns (neutral biodiversity, adaptation, adaptive diversification).

- **Chapter 3 - Neutral diversity in experimentally nested populations** (p. 31).

This chapter explores neutral diversity patterns in a metapopulation structure inspired by ongoing developments in digital millifluidics —a technology that can be used to implement a nested population structure for microbial populations and communities in an automated and high-throughput fashion. The model is composed of particles following a neutral birth-death-mutation process, embedded into collectives that themselves undergo discrete non-overlapping generations in the form of a growth-selection-dilution cycle. Diversity measures are derived for both single (serial transfer) and splitting collective lineages, and parameters maximising diversity are explored, using coalescent point processes. The advantage of collective-level selection for artificial selection of neutral mutations is discussed.

- **Chapter 4 - Locating Mutations in Collective Genealogies** (p. 67).

This chapter proposes a general method to visualise and interpret collective-level genealogies obtained experimentally by my collaborators working on *Pseudomonas fluorescens*. A thought experiment on the ecological scaffolding of collective-level reproduction is presented, along with the description of its experimental implementation. A Bayesian-Network model is used with a belief-propagation algorithm to assign an adaptive-value (survival probability) to each vertex of the genealogy, and pinpoint key biological innovations. This model is also used to propagate partial DNA sequence information to the whole genealogy. A case is made for this approach that can be used for interpreting past results and planning new experiments.

- **Chapter 5 - From Particles Traits to Collective Demographics** (p. 85).

This chapter uses an adaptive dynamics framework to model the long term evolutionary dynamics of a nested population under ecological scaffolding. This approach focuses on trade-offs between survival and reproduction of collectives. In particular, it explores how the evolutionary dynamics of individual traits (with a focus on evolutionary singular strategies) is constrained by collective-level selection, and in turn how they shape the evolution of collective-level life history parameters. The models are inspired by the system presented in Chapter 4 and range from close to the experimental setup, to close to a thought experiment. The models allow derivation of simple necessary conditions for ecological scaffolding.

- **Chapter 6 - An ecological recipe for the emergence of collective-level heredity** (p. 117).

This chapter presents a model of the evolution of collective-level heredity in a simple two-species communities with density-dependent competition. It starts with the observation that collectives reaching an evolutionary successful state (i.e., a composition and organisation of particles that gives rise to numerous of collective-offsprings) can only subsist for a meaningful amount of time if offspring are able to reliably reach a fertile state similar to that of the parent. If collectives produce many low-fidelity offspring that through noise explore the phenotypic space, some are statistically ensured to reach a fertile state and continue the lineage (stochastic corrector). Ecological scaffolding is proposed as a possible scenario for the refinement of this mechanism leading to the evolution of a proto-developmental programme based on density dependant interactions (developmental corrector), which marks a level of innovation that delivers a means of ensuring that offspring resemble parental types.

Finally, the manuscript ends with some brief and general concluding remarks (p. 151).

## Chapter 1

# Historical Perspectives

“La science d’hier était riche du manque de la science d’aujourd’hui. La connaissance d’aujourd’hui est riche du manque de la recherche à venir. Et tout espoir que la science fera demain de grande découvertes est d’abord l’affirmation de la certitude de ce qui manque.”

— JEAN-CLAUDE AMEISEN, *Dans la lumière et les Ombres, Darwin et le Boulversement du Monde* (2008)

**D**ELINEATION OF PROPERTIES NECESSARY AND SUFFICIENT for entities to participate in the process of evolution by natural selection — the so-named Darwinian properties — is a relatively recent development. Regardless on whether the origin is considered to be Darwin (1872) or Lewontin (1970), the delineation of Darwinian properties constitutes the last entry in a wide and tangled web of concepts that have been used for centuries to answer the fundamental question: “What is life?”. Laying down the characteristics that set living entities apart from the rest of the physical world, is a slow and, to some extent, still ongoing process.

The aim of this chapter is not to offer a summary of the history of biology, but to broadly sketch the history of how the idea of abstract characteristics of living individuals emerged. Given the current focus on Darwinian properties, this chapter addresses more generally concepts of variation, reproduction, and heredity in western biological thought.

### A web of similarities

Up to the Enlightenment, an important part of natural history consisted in the exegesis of ancient texts such as *On the generation of Animals* by Aristotle (1887). The general view at the time was that natural objects have two components: form (or movement) and matter.

The matter is eternal: when an entity is “created”, it is the form that emerges, whereas it disappears whenever the object ceases to be (Fresnel 1637 cited in Jacob (1970)). This is true for living organisms as well as inert objects, and reflects a central feature of the paradigm of this era: the lack of separation between the living and the non-living. In fact, natural history recognises three domains: animals, plants and minerals. There is no



**Figure d’un monstre à demi homme et demi pourceau,**  
(Paré (1641),  
Gallica/Bibliothèque Nationale  
de France)

category regrouping animals and plants together, setting them apart from minerals. Animals are *generated* and not *reproduced*. The word reproduction is restricted to the regrowth of severed organs that can be observed in plants or some taxa of animals (Diderot and d’Alembert, 1751) and was not widely used in its current meaning until the 19th century.

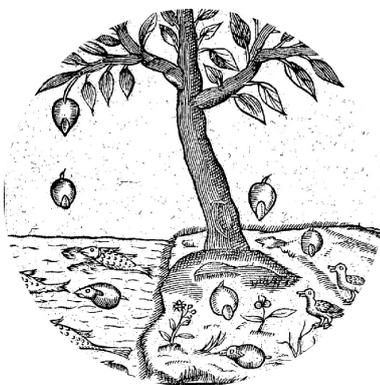
Thus, species are not defined by genealogy but by shape. The fact that vastly different organisms cannot produce each other is not clearly established. Duret (1605) describes at length, with accompanying etchings, the rotting trees that produce “worms, then living and flying ducks”, goose barnacles that produces geese and even a specific tree whose leaves turn into fishes (if they fall in the water) or ducks (if they fall on land). Spontaneous generation for macro-organisms was refuted as early as the 17th century (Francesco Redi, 1686) but continued to be widespread for smaller organisms until the 19th century.

Heredity is observed, particularly in cattle and humans, but is restricted to the immaterial “form”: it has no physical support. In a sense, heredity is akin to the “artist’s touch” (Jacob, 1970) that makes one’s work identifiable but is not characteristic of the underlying material. Heredity of traits bridging over multiple generations (e.g., traits shared by grandparent and their grandchildren, but not the parents) contradicts this, and must be interpreted as a failure of the parental form to settle, accompanied by a supposed—and vague—natural tendency of things to “fall back” on the form of their ancestors when that of their parents is unavailable (Aristotle (1887), chapter 3).

Additionally, heredity is limited to sexually reproducing organisms: animals that mate generate offspring that looks like them. Conversely, animals that are spontaneously generated do not mate, and do not generate individuals that look like them (Aristotle (1887), chapter 1). This concerns most pest “generated from putrid matter” such as “snakes, grasshopper, worms, flies, mouses, bats [and] moles” (Fernel 1637, cited in Jacob (1970)). This widespread belief in spontaneous generation is a telltale sign of the lack of the separation between life and non-life.

The extreme diversity of living systems is recognised, but not systematically studied. Similarities between living objects are observed, but they are studied for their own sake as similarities and not as the result of similar or converging processes. Still, diversity and similarities are thought to reveal the hidden structure of Nature: Lungwort (genus *Pulmonaria*) was used to treat pulmonary infection due to its resemblance to ulcerated lungs. This vision of nature as a wide tapestry of analogies to untangle constitutes the *doctrine of signatures* (Foucault, 1966).

“Monsters” are studied with interest as deviations from the usual forms of organisms. Their status is very different from the study of mutants in current biology: the concurrent lack of a theory of heredity as well as systematic experiments prevents those observations to fit in a wider understanding of life. Texts are replete with description of hybrids between species (Paré, 1641), even though some authors points out, as early as classical antiquity, that extraordinary hybrids like animal-headed humans are not the product of cross-breeding, preferring to explain them as a failure of the form against the matter (Aristotle, 1887).



Portrait de l’arbre qui porte des feuilles lesquelles tombées sur terre se tournent en oiseaux volants et celles qui tombent dans les eaux se muent en poisson., (Duret (1605), Gallica/BnF)

## A fixed collection of species

Slowly, Science distanced itself from the goal of mapping similarities in nature in order to unveil its secrets. The new reductionist paradigm, illustrated by the *obviousness criterion* of Descartes, is rooted in the idea that natural systems must be separated in parts that are small enough to be understood simply (Legay, 1997).

The emergence of the concept of the species at the end of the 17th century constitutes a true paradigm shift, in which heredity plays a central role. For the first time, living beings are not categorised by their shape but their descent, regardless of differences between stages of life (for instance larvae or imago) or individual variations. This classification method stems from the observations of naturalists such as Ray (1686). It gained universal adoption as part of the Linnean system of classification (Linné, 1735).

The Linnean system extends beyond life as it still distinguishes between minerals, plants and animals. Indeed, for Buffon (1829), the frontier between the living and non-living is fuzzy: it is possible to “climb down by seamless steps from the most perfect of creatures to the most shapeless matter, from the most well-organised animals to the most crude of minerals”. However, common characteristics of life start to emerge: “A species is nothing else than a constant succession of similar individuals, that reproduce themselves, and it is clear that this term should only apply to vegetals and animals [and it is by mistake that it was used for minerals]” (Buffon, 1882). Nonetheless, mechanisms of heredity are still out of reach. Buffon imagines an “internal mould” that constrains not only the external form —as the moulds commonly used in casting— but that would also determine internal structure of individuals (Buffon, 1882) but he cannot provide any insight in its mechanism of action.

On the subject of variation, individual differences from the ideal mould of the species are mostly considered as punctual defects without consequence (Buffon, 1882). It is recognised that some variations are heritable, and possibly stabilised by artificial selection (Maupertuis, 1754). However, generation of new species by this mechanism is rejected (Buffon, 1882). Since transformations of species are not yet deployed on the long term to explain the origin of life’s diversity, biology in the 18th century can be qualified as resolutely fixist (Jacob, 1970).

## A set of evasive living particles

In the 18th century, advances in physics and chemistry open the path for an in-depth exploration of natural phenomena such as respiration and digestion (Lavoisier, 1862). Vaucanson (1738) presents to the *Académie des Sciences*, an automaton of a duck that “eats, drinks, digests and empty its bowels”. Even though it illustrates how the phenomenon of life could be imitated by machines, the artificial duck is first and foremost a feat of mechanical engineering. The author himself admits that the artificial digestion does not “produce new blood” or participate in the upkeep of the organism. If the engineering metaphors are successful in addressing physiology problems, they are vastly under-equipped to tackle the problems of reproduction and



**Allégorie de la Science,**  
(Buffon (1829), Gallica/BnF)



**Le canard artificiel,**  
(Vaucanson (1738),  
Gallica/BnF)



intellectual life to incorporate the evolutionary world view”, after cosmology (with the nebular hypothesis for the formation of planets, formulated by Kant in 1786 and formalised by Laplace in 1796), geology (with the principle of uniformitarianism which implies that general laws apply over long periods of times, attributed to Hutton in 1785 and popularised by Lyell in 1830), thermodynamics (with Carnot in 1824 and Thomson in 1851) and linguistics (with the idea that languages have been developed rather than created, which was the dominant view in 1857 according to Spencer). In natural sciences, the influence of geological studies cannot be understated, even more so that they were coherently fitting in with the early transformist theories, as for instance proposed by Lamarck (Mayr, 1991). Darwin himself wrote that Lyell’s Principles “altered the whole tone of one’s mind” (Ameisen, 2008).

Darwin’s magnum opus about the common descent with modification (Darwin, 1872) closes with the outline of the characteristics of all living organisms that will become the Darwinian properties: “growth with reproduction”, “inheritance” and “variability”.

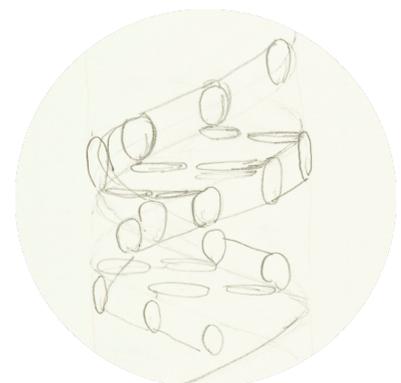
Darwin’s own vision of heredity changes dramatically during his lifetime (Mayr, 1991). In 1868, he proposes the provisional hypothesis of *pangenesis*, which accounts for both heredity and development (Darwin, 1968). In a manner reminiscent of the living-particles theories from the previous century, Darwin postulates the existence of *gemmules*. Gemmules are self-dividing particles, produced by cells and that can, under the right conditions, develop into a cell of the same kind as its producer. Gemmules are collected from all parts of the body to form the sexual elements. Once again, development is explained by the affinity of particular gemmules to particular types of cells. Heredity of acquired traits is possible. However, the theory of pangenesis is quickly disproved by blood transfusion experiments (Galton, 1871). Later, the first statistical theory of heredity is established by Galton (1894).

Even the scale of the unit of life changes during the 19th century: cell theory gives to cells the status of the smallest unit presenting all characteristics of life. This is in contrast with the purely structural vision of cells that can be found earlier (for instance in Lamarck (1809)). From this moment on, multicellular organisms can be considered as collectives of interacting cells: every “so-called individual represents a kind of social arrangement of parts”, writes Virchow (1863) in his lectures about cell pathologies, “an arrangement of a social kind, in which a number of individual existences are mutually dependant.”

By the end of the 19th century, neo-Darwinist thought came to question the inheritance of acquired traits (Mayr, 1991). At the center of this rebuttal, a growing body of evidence shows the separation of germinal and somatic lines. This theory leaves no room for cytological mechanisms explaining the transmission of characters from soma to germ (Weismann, 1892).

## A complex system

The 20th century opens with the rediscovery of Mendel’s laws of inheritance and constitute the final nail in the coffin of the theory of blending inheritance. These organism-level observations are rapidly correlated with the movements of chromosomes during meiosis and fecundation. Observations of



**Sketch of DNA structure,**  
(Francis Crick, Wellcome  
Collection)

the segregation of both traits and chromosomes in *Drosophila* firmly establish chromosomes as the physical support for heredity (Morgan, 1915).

The first quarter of 20th century offers also statistical refinements for the analysis of heredity, with the spreading of the Analysis of Variance. More than a technical improvement, those methods pioneered by Fisher constitute a true epidemiological shift. For the first time, it is indicated clearly that there may exist many factors influencing a character, even though the scientist can affect only a handful. The other factors are not ignored in the analysis but considered as “uncontrolled”. This is the beginning of complex systems thinking, in clear contrast with the previous criterion of evidence (Legay, 1997). In addition, Fisher is at the centre of the reductionist approach of evolution, that defines evolution purely in terms of the dynamics of genes frequency (Mayr, 1991).

Before the middle of the century, the idea that the gene is a molecule is commonplace. At the meantime, the veil is slowly lifted on the physical mechanisms of variation: the spontaneous nature of mutations is demonstrated in bacteria by the fluctuation test (Luria and Delbrück, 1943). This constitutes the proof that the source of biological variation is largely independent of the selective process. Additionally, decisive experiments with X-ray-induced mutagenesis place the gene-molecule size at about a thousand atoms, and hint at it being a kind of “aperiodic crystal” (Schrödinger, 1944), or, rather, an aperiodic polymer. Finally, the discovery of the molecular structure of DNA (Watson and Crick, 1953b) and, shortly after, the discovery of the genetic code mark the beginning of the molecular genetics era.

By the middle of the 20th century, the modern evolutionary synthesis unifies the fields of genetics, systematics and palaeontology that used to be separated. Heredity and variation play a central role in this unified view. The synthesis cements and diffuses within the scientific community the concepts of genotype and phenotype, the idea that the source of variation is spontaneous mutations, that heredity is not blending and does not concern acquired traits (Mayr, 1991).

## A vehicle for information

The years following the synthesis (from 50s to 70s) see the reaffirmation of the individual as the object of selection (Mayr, 1991). However, naïve individual-centred approaches have the shortcoming of not being able to predict, or explain, neither cooperative behaviours nor control of inter-individual conflict. In a seminal article, Hamilton (1964) puts forward the concept of *inclusive fitness*, a quantity that aims to correct for social context and take into account the harm and benefits caused by the focal individual on its neighbours, proportionally to their relatedness. Concurrently, Maynard Smith and Price (1973) introduce evolutionary game theory as a method to solve the problem of restraint in animal conflict. These developments participate in the effort of distinguishing what kind of entities might be subject to evolution by natural selection, and operate a much-needed clarification of the possible mechanisms of evolution: the “good-of-the-species” cannot (as seen in Wynne-Edwards (1962)) be invoked as the evolutionary cause for a behaviour or a trait. This does not disqualify the plurality of levels of organisation to play

a role in evolution, but their Darwinian nature has to be justified (Maynard Smith et al., 1993). In 1966, Williams (2018) states that “only by a theory of between-group selection could we achieve a scientific explanation of group-related adaptation.” He classifies mechanisms by their effect on the individual level (named organic) or beyond (named biotic). Finally, he concludes that biotic selection, if possible in theory, is not potent enough to be significant in most natural situations and famously states “Group-related adaptation do not, in fact, exist.”

The 70s mark a turning point in the study of what constitutes the unit of evolution. Lewontin (1970) puts forward the modern formalisation of the set of properties that are necessary for evolution by natural selection: phenotypic variation, differential fitness and heritable fitness. He decouples those necessary properties from the mechanisms of heredity. Natural selection is expected if these properties are verified, had they arisen from “Mendelian, cytoplasmic or cultural inheritance”. Pioneered by Williams (2018), and popularised by Dawkins’ widely influential *Selfish Gene* (Dawkins, 1976), the gene-centered view of evolution postulates that the gene is the only entity that really fits the definition of a unit of evolution. All in all, if the 18th century saw living beings as machines made of pumps and furnaces, the metaphor favoured by the 20th century is the one of information and programme, encoded in genes. As a prime example, Monod (1970) distinguishes two properties of living systems: reproductive invariance, (i.e., information transmitted across generations) and structural teleonomy (i.e., the apparent purposefulness of living beings to fulfil the programme of invariant reproduction).

## A nested structure

The nested organisation of living systems becomes a focus of research in the last part of the 20th century. The endosymbiosis theory for the origin of the eukaryotic cell advocated by Margulis (1970) makes manifest the fact that, even at the cellular level, organisms are the result of the integration of different components that used to have their own individuality. On the other side of the spectrum, the field of sociobiology funded by Wilson (1975b) uses Hamilton’s inclusive fitness to propose an evolutionary theory for social structures.

Jacob (1970) claims that the advances of molecular biology mark, at last, the merging of physiology and natural history: organisation and evolution, molecular mechanisms of heredity and teleonomy can finally be treated in a unified biological framework. As a consequence, the nested organisation of living beings, composed of layers of integrated sub-units (called integrons), becomes a consequence of evolutionary processes. In the following years, the origin of new levels of organisation is treated a general question on its own right, with for instance the work of Buss (1987) on the evolutionary origins of individuality. The concept of major transitions in evolution, formalised in the seminal book of Maynard Smith and Szathmary (1995), applies such question across scales, from the origin of chromosomes and cells, to multicellular organisms and societies.

Despite its apparent dismissal in the 60s, group selection was the subject of continuous of research until today (Wilson and Wilson, 2007). On the side of theory, multi-level selection extends Darwinian principles to a hierarchi-

cal organisation: particles populations nested in a population of collectives (Okasha, 2006). Two kinds of approaches are distinguished: Multi-level selection 1 (MLS-1) models focusing on particles, where the group structure of the population is modelled as environment (such as the classic trait-group model of Wilson (1975a)), and Multi-level selection 2 (MLS-2) models where collectives are treated as unit of evolution on their own right. On the side of experiments, a wealth of results confirmed that group-level selection is a potent force in controlled conditions: successful selection at the level of the group was performed on hens to select for non-aggressive behaviour (Hesters et al., 1996). Artificial selection is also possible on whole communities as shown in floor beetles (Goodnight, 1990b,a) with the selection of population size and immigration rates. In microbial communities, experiments selecting for pollutant degradation (Swenson et al., 2000a), selecting plant microbiome for increased biomass (Swenson et al., 2000b) or flowering time (Panke-Buisse et al., 2015) opens new perspectives in microbiome “breeding” (Arias-Sánchez et al., 2019; Xie et al., 2019). In evolutionary microbiology, the influence of population structure on viral restraint (Kerr et al., 2006) and early multicellularity (Hammerschmidt et al., 2014; Ratcliff et al., 2012) illustrates how some key phenomena of the history of life can be explained by a nested Darwinian population structure.

Both Darwinian properties and the concept of evolutionary transition in individuality are the product of the long process of formalisation of biological thought. They reflect the current version of a constantly changing framework for the study of the principles organising the living world. This history offers an example on the long term dynamics of knowledge itself, with paradigm shifts (Kuhn, 1970) when an explanation emerges and replaces another. Keeping this context in mind, the next chapter focuses on a very contemporary view of evolutionary processes, based on mechanistic explanations as well as a quantifiable formalisation of Darwinian properties.

## Chapter 2

# Darwinian Properties

“There is grandeur in this view of life, with its several powers, having been originally breathed into a few forms or into one; and that, whilst this planet has gone cycling on according to the fixed law of gravity, from so simple beginning endless forms most beautiful and most wonderful have been, and are being, evolved.”

— CHARLES DARWIN, *On the Origin of Species* (1859)

**D**ARWINIAN PROPERTIES are characteristics of populations that, to some extent, constrain their ecological and evolutionary dynamics.

This chapter introduces the core concepts used in the rest of this manuscript. It starts by building a mechanistic view of eco-evolutionary processes, and re-introduces the concept of Darwinian properties in this context. Then, the ordering of Darwinian populations from marginal to paradigm is discussed in detail, using a simple toy model to build intuition. This toy model allows to qualitatively link Darwinian properties and common patterns of eco-evolutionary dynamics: neutral diversity, adaptation and branching. Finally, the emergence of collective-level Darwinian properties by ecological scaffolding is described.

### Eco-evolutionary Dynamics Are Composed of Population and Trait Dynamics

Consider a population of individuals in a given environment. Each individual is a distinct entity, characterised by a set of physical properties: its traits. In general, a trait can be any characteristic of the individual, from very coarse-grained properties like biomass or volume, to very information-rich ones like the nature and position of all the molecules composing the individual. Traits can also be behavioural characteristics like feeding habits or propensity for social interactions. Enumerating all traits of individuals is already a near-impossible task. Practical models of evolution always limit themselves to a handful of focal traits.

Eco-evolutionary dynamics can be decomposed into two classes of phenomena.

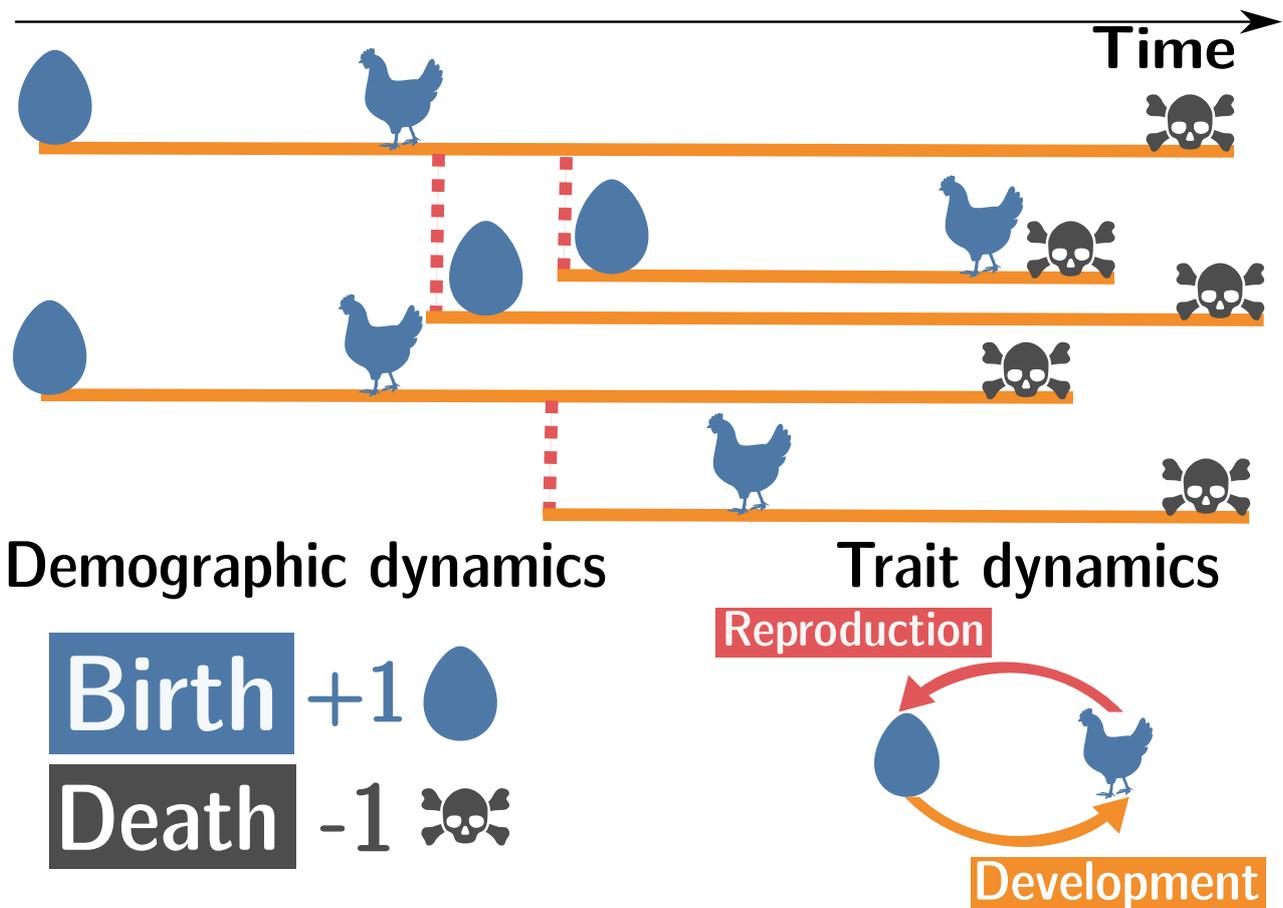


Figure 2.1: **Population processes are composed of two kinds of dynamics.** Demography, related to the addition and removal of individuals, and traits dynamics that encodes the change in the internal state of individuals.

The first one is *demography*: the number of individuals fluctuates through time via four fundamental events: birth, death, emigration and immigration. The rate at which these events occur is influenced by a variety of biotic and abiotic factors. For instance birth and death rates may depend on resource availability (e.g., phytoplankton blooms resulting from nutrient concentration fluctuations), and migration rates may be modified by the location and accessibility of a population (e.g., populations on remote or small islands experience less immigration). Additionally, interactions between individuals can have a large effect. For instance, a higher population size may increase the death rate by crowding competition, or, depending on context, increase the birth rate by alleviating mate limitation (Allee effect). All those factors might be modulated by the traits carried by individuals. Overall, there are two categories of demographic events: the ones that add individuals, generally called *births*, (but also including immigration) and the ones that remove individuals, generally called *deaths* (but also including emigration).

The second key phenomenon is *traits dynamics*: the fluctuation through time of individual traits values. Individuals do not stay identical during their lifetime: their chemical composition change, they grow, move, transform and age. Traits dynamics can also be subdivided into two categories: changes in traits that occurs during the lifetime of an individual, generally called *development* (but also including growth and ageing), and changes that occurs when new individuals are generated from existing ones: generally called *reproduction*.

Birth, Death, Development and Reproduction (Figure 2.1) are the four essential phenomena necessary to model the dynamics of a given population (with the possible addition of feedback loops between the abiotic environment and the population). There is a wide diversity of modelling approaches in evolutionary biology: stochastic or deterministic, discrete or continuous, mechanistic or phenomenological. However, all models make implicit or explicit assumptions about those four processes. The next section details how, with a specific set of assumptions, those dynamics are governed by natural selection.

## Darwinian Properties

The most distinctive feature of the eco-evolutionary dynamics of a Darwinian population is adaptation by means of natural selection. To restate what was presented in the introduction of this manuscript, Darwinian properties are *abstract*, in the sense that their definition does not refer to the underlying material composition of the objects. Moreover, since the phenomenon of evolution by natural selection follows as a consequence of these properties, they constitute the only known causal explanation for the *teleonomic* nature of living entities (i.e., their apparent goal-directedness). Finally, Darwinian properties might be expressed by different kinds of units throughout nature, and potentially units at different, *nested levels of organisation*, which poses the problem of the origin of those levels as well as their relative relevance in both historical and current life's patterns.

The minimal conditions for change by natural selection, spelled by Lewontin (1970), can be transposed to the framework of demography and trait dynamics presented in the previous section (Figure 2.2). First, the condition of phenotypic variation requires that traits can take several values. More formally, it means that the state-space of individual traits must not be reduced to a point. Second, the condition of differential fitness translates to a dependency between the rate of birth and death events and the trait values of individuals. Third, the condition of heritable fitness implies that, between reproduction and development, “like begets like”. More formally, it means that the distribution of offspring traits (as the product of reproduction and developmental processes) should not be uniform on trait state-space, but be somehow centred on the parental traits.

The problem of “summary” approaches to describe Darwinian properties, such as the one of Lewontin (1970) is that they are the product of a trade-off between two distinct goals, as aptly pointed out by Godfrey-Smith (2009). On the one hand, they strive to give “a laterally true description of all the important features of all cases of [natural selection]” and, on the other hand,

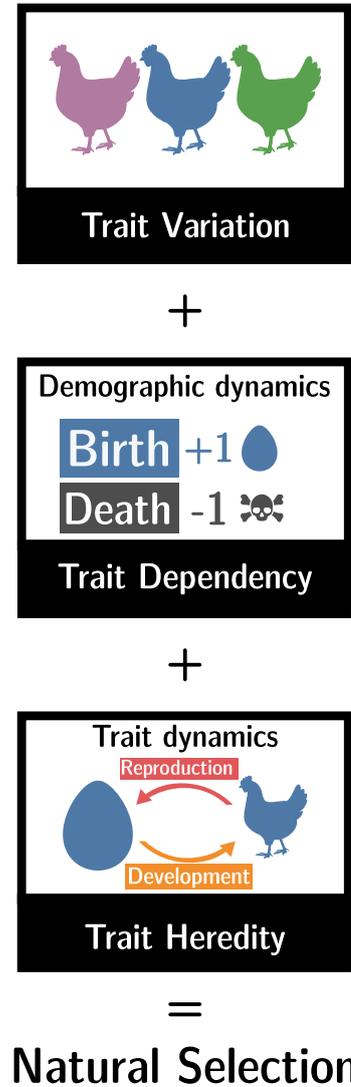


Figure 2.2: **Darwinian Properties are conditions that population processes must verify to undergo evolution by natural selection.** The classical triptych “variation, reproduction, heredity”, can be reformulated as constraints on trait dynamics and demography.

to describe a simple case of natural selection that can be used as the basis for understanding the phenomenon in its generality (p. 26). To some extent, this fails on both accounts. For instance, a population can satisfy all three conditions, but be at an evolutionary equilibrium, thus not exhibiting any change in trait distribution. The conditions are not sufficient for change, they are not a complete “recipe”. Conversely, Lewontin-like summaries are not precise enough to constitute a model, but any additional details, for instance on the mechanism of heredity, tie them to a specific system and limit their generality. Overall, one cannot hope to achieve simultaneously both goals of expressing the general nature of natural selection and a mechanistic description of change with a compact, one sentence summary. This is where the distinction between *minimal*, *marginal* and *paradigm* Darwinian populations comes into play (Godfrey-Smith, 2009, p. 41).

A population is *minimally* Darwinian if it exhibits the properties of variation, differential reproduction and heredity as laid-out by the classical three-parts summaries such as Lewontin (1970). This minimal level is a starting point shared by all Darwinian populations (Godfrey-Smith, 2009, p. 41). A subset of those minimal populations are the *paradigm* populations. Paradigm populations display the minimal properties plus others. For instance, Mendelian inheritance is a characteristic of sexually reproducing population, but not a minimal property. They correspond to populations where the Darwinian dynamics is at its clearest, and that give rise to complex and adapted structures. Conversely, at the edge of the minimal concept are the *marginal* populations. Marginal populations do not possess exactly the minimal properties, but approximate them to some extent: as a consequence they can display some behaviours expected of minimally Darwinian populations. These categories are useful verbal arguments, but their limits are actually blurry. Consequently, *paradigm* and *marginal* serve best as labels for the two ends of a continuous spectrum of properties, on which one may order models or real life example.

Once again, Godfrey-Smith (2009) proposes a strategy to deal with the comparison of level of “Darwinianess” in different populations. Here the term “level” refers to the intensity of the Darwinian properties, and not the level of organisation. Nonetheless, it is entirely possible to imagine comparing human cells as a population with the human population. He establishes a partial list of features that relates to the distinction between marginal, paradigm and minimal cases (Table 2.1). Each biological example or model can in theory be assigned a coordinate in each of those dimensions. The general tendency being that high values are considered paradigm, whereas lower values are marginal. He proposes to call the resulting vector space a *Darwinian space*, and argues that the associated spacial representation is a powerful conceptual tool (p. 64).

As a matter of fact, the spatial representation of the Darwinian space is a powerful tool. However, there is no systematic method to decide which dimensions are relevant, nor how to place models or biological systems within them. In the next section, I sketch a simple toy model that can be used to build intuition about Darwinian properties and their classification from marginal to paradigm.

H	Fidelity of Heredity
V	Abundance of Variation
A	Competitive interactions with respect to reproduction
C	Continuity, or smoothness of fitness landscape
S	Dependence of reproductive differences on intrinsic character

Table 2.1: **Some Dimensions of the Darwinian Space** (from p. 63 in Godfrey-Smith (2009))

## A toy model to explore the Darwinian space

Let us define a simple toy model of eco-evolutionary dynamics, parametrised by the dimension of the Darwinian space. The model is described as a stochastic process, meaning that a given initial condition yields a set of different possible trajectories, associated with a probability. More precisely, it is a continuous time Markov process, meaning that the population changes via point events whose propensity is a function of the current state of the population. This reflects the assumption that the description of the population within the model at any given time is enough to completely define the future distribution of trajectories. There is no memory of past states, outside what is encoded in the population state at a given moment in time.

This model will be limited to three evolutionary patterns: neutral diversity distribution, in relation with the  $H$  dimension of the Darwinian space, adaptation, in relation with the  $S$  dimension, and adaptive diversification or branching in relation with the  $A$  dimension. Once again, the goal is to build intuition about those phenomena and properties, the analysis will be kept short and qualitative.

### Simple parameters

The following formalism is used: Each individual  $i$  has traits  $\theta_i \in \Theta$ . The population is described at any point in time by the trait value of all its individuals. For each individual, births events are given a rate  $b_i$ , and the death event rate  $d_i$ . Rates may depend on  $\theta_i$  and on the state of the population. When a birth occurs, a new individual is added with a trait drawn from a distribution  $r$ , that may depend on the parental trait. Development is ignored in this toy-model: the trait value stays constant until the individual is removed from the population by a death event. This analysis is limited to parameters that influence  $b$ ,  $d$  and  $r$ , the role of initial conditions will not be systematically explored.

Consider a discrete trait space  $\Theta$ . There is only a finite number  $V$  of possible trait values. Thus, let  $\Theta = \{1, 2, 3, \dots, V\}$ . Let  $\Gamma$  be a distribution on  $\Theta$  parametrised by  $\mu$  and  $p$ :

$$\Gamma(x, \mu, p) = \sum_{i \in \mathbb{Z}} \int_{i \lfloor \frac{x}{V} \rfloor}^{i \lfloor \frac{x}{V} \rfloor + 1} \phi \left( s, \mu, \frac{1}{p^{0.1}} - 1 \right) ds$$

Where  $x \mapsto \phi(x, \mu, \sigma)$  is the Gaussian probability distribution function with mean  $\mu$  and standard deviation  $\sigma$ . The parameter  $\mu \in \Theta$  controls the

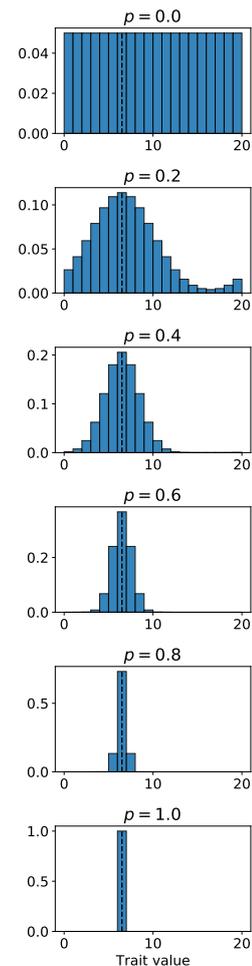


Figure 2.3: **Simple parametrisation of a local to global effect.** Distribution  $\Gamma$  over  $\Theta$  with  $V = 20$ ,  $\mu = 8$  for different values of  $p$ . When  $p = 0$ , the distribution is uniform on  $\Theta$ . When  $p = 1$  it is an atom in  $\mu$ .

expected value of the  $\Gamma$  distribution (if  $p \neq 0$ ), while  $p \in [0, 1]$  controls its spread. If  $p = 0$  the distribution is uniform on  $\Theta$ , whereas if  $p = 1$  the distribution is a single atom in  $\mu$  (Figure 2.3). Note that  $\Gamma$  “wraps on the edges” of the trait space, giving it a torus-like structure.

### Neutral diversity pattern

Documenting diversity is a task as old as Natural History. However, biodiversity starts to become the subject of mechanistic and quantitative inquiries with the seminal work on island biogeography by MacArthur and Wilson (1967). The unified theory of neutral biodiversity (Hubbell, 2001), inspired by the neutral theory of molecular evolution of Kimura (1983), builds on Island biogeography and describes in more details mechanisms and patterns of non-adaptive diversity. One of the main results of neutral theory is that the distribution of variants belong to a family of distribution parametrised by a single adimensional biodiversity parameter.

Neutral diversity implies that the dimension  $S$  (intrinsic effects of traits on demographic rates) is low, but it still assumes some fidelity of heredity. Let  $H \in [0, 1]$  be the one-dimensional parameter representing the fidelity of heredity in the Darwinian space.  $H$  parametrises the trait distribution of offspring  $r$ . A higher  $H$  means that the distribution of offspring trait  $\theta_{off}$  is more localised around the parental phenotype  $\theta_{par}$ :

$$r(\theta_{off}|\theta_{par}) = \Gamma(\theta_{off}, \theta_{par}, 1 - H)$$

Additionally consider that  $b = 1$  for all trait values, and  $d = N/K$ , where  $N$  is the number of individuals and  $K$  is an arbitrary carrying capacity of the medium. The initial condition is a monomorphic population of  $K$  individuals.

If  $H = 1$ , no change occurs and the trait distribution stays constant. Figure 2.4 presents the pattern of neutral diversity for some other values of  $H$ . If  $H = 0$ , the distribution quickly becomes uniform. For  $H \notin \{0, 1\}$  the asymptotic distribution (after an infinite duration) also seems uniform. This can be confirmed by noticing that this model can be approximated by a diffusion in the trait-space, which has a uniform asymptotic distribution. However, at intermediate time points, the rank abundance distribution follows a power law with an increased slope for higher heredity values and lower time duration. When Heredity is high, the trait abundance distribution shows a heavy-tail of a few trait values that are shared by numerous individuals. In addition, the most abundant trait-class is the one that is carried by the lowest number of individuals. This pattern fits with both empirical observations and the prediction of neutral diversity theory. However, when the population sit lower on the Darwinian space, this pattern degrades and is unrecognisable.

Overall, neutral diversity patterns relies on a certain level of fidelity of trait heredity. Chapter 3 of this manuscript explores further neutral diversity patterns in a structured population. If neutral diversity mechanisms are pervasive in the living world, they cannot explain complex structures and functions: an eye cannot emerge from neutral evolution alone. The next section addresses the question of adaptation.

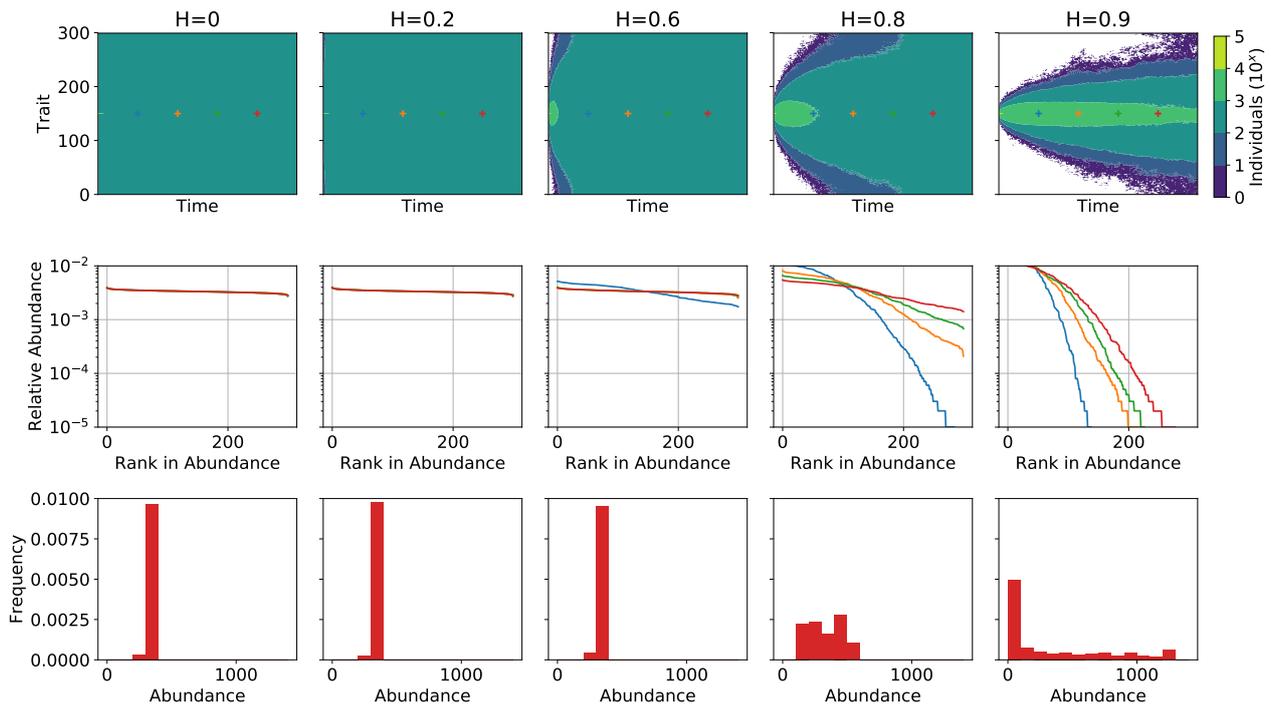


Figure 2.4: **Neutral diversity patterns along the Heredity axis.** *Top:* Distribution of traits value through time. *Middle:* Rank Abundance distribution at four different time points. *Bottom:* Frequency distribution for the last time point.  $V = 300$ .

## Adaptation

Adaptation, defined as the fit between an organism to its environment, a result of the process of evolution by natural selection. Adaptive values (or fitness measures) are the most common way to quantitatively assess adaptation. A fitness measure of an individual is proportional to its long term contribution to the future population. Usually it is measured as an average on a given sub-population. A common example is the exponential growth factor of the population (or Malthusian factor, Fisher (1930)). The process of natural selection is predicted to increase the adaptive value: evolution would act as an optimising process on this quantity.

While the teleonomic nature of organisms is a commonplace observation, quantitative, fitness measures are notoriously complex and contentious because of inherent dependencies on the ecological context of individuals as well as the measurement protocol. Nonetheless, in a constant, controlled environment, fitness measures can increase for tens of thousands generations as demonstrated in the famous Long Term Evolutionary Experiment (LTEE) on *Escherichia coli* (Wiser et al., 2013).

If some organisms can be more adapted than others, it implies that their intrinsic characteristics have an impact on their reproductive output. This

relates to the dimension  $S$  of the Darwinian space. As an example, let  $S$  be a one-dimensional parameter in  $[0, 1]$  that parametrises the per-capita birth rate  $b$ . A low  $S$  means that all trait values have the same birth rate, while a higher  $S$  means that trait values around an arbitrary optimum  $\theta_{opt} = V/3$  are favoured:

$$b(\theta) = \Gamma(\theta, \theta_{opt}, S)$$

The offspring trait distribution  $r$  is the same as previously, with  $H = 0.9$ . As before, the death rate is density dependant  $d = N/K$  and the initial conditions are a monomorphic population of  $K$  individuals.

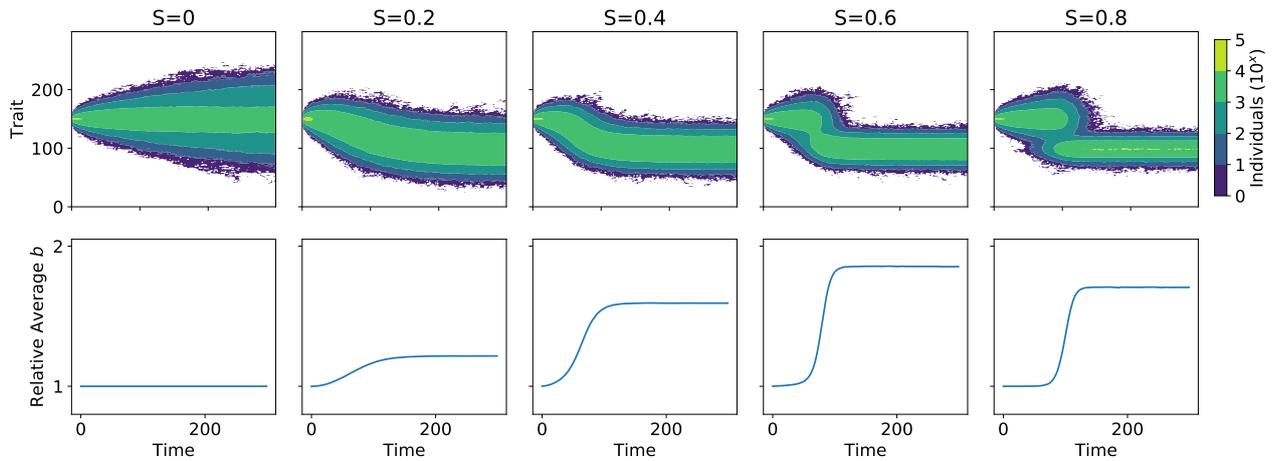


Figure 2.5: **Adaptation along the selection axis** *Top*: Distribution of traits value through time. *Bottom*: Average birth rate through time relative to the initial birth rate.  $V = 300$ ,  $H = 0.9$ ,  $\theta_{opt} = 100$ .

Figure 2.5 presents trajectories and adaptation patterns for different values of  $S$ . When  $S = 0$ , the birth rate is independent on the trait value, and the simulation is identical to the previous section (with  $H = 0.9$ ). The corresponding adaptive value is constant. When the birth rate depends on the trait value,  $S \neq 0$ , the trait distribution changes through time and become centred around the optimal value  $\theta_{opt} = 100$ . Concurrently, the adaptive value increases, up to a plateau. The height of the plateau depends on the maximal value of  $b$ , which increases with  $S$  (as seen in Figure 2.3). Note that the speed of the adaptation is higher when the adaptive landscape is steeper (higher values of  $S$ ).

Overall, adaptation requires a certain level of dependency between trait values and demographic rates. The outcome of the adaptive process in terms of trait value can be predicted using adaptive dynamics (this method will be used in Chapters 5 and 6). The adaptation pattern presented here is straightforward, and results in an apparent optimisation of the adaptive value. While this description is accurate in this specific example, it does not reflect the diversity of neutral patterns, as exemplified in the next section.

## Branching and Diversification

Treating adaptation as a naïve optimising process misses on a wide class of ecological and evolutionary feedbacks (Doebeli, 2011). Indeed, when demographic event rates depend on the trait distribution, individuals with different traits may have differential effects on their environment or directly interfere with each other. In this situation, named *frequency-dependent selection*, adaptation to a constant environment seldom occurs. This can lead to diversification in remarkably different ways than neutral evolution. A particularly relevant example for the rest of this manuscript is the adaptive diversification observed in cultures of *Pseudomonas fluorescens* (Rainey and Travisano, 1998). In a few days, an unshaken culture of this aerobic bacteria reproducibly diversifies into three phenotypes with complex ecological dynamics. Among those phenotypes are cells over-producing cellulose, that can form a mat at the air-liquid interface. Thus, they benefit from an increased oxygen supply, and increase in frequency with respect to the ancestral planktonic phenotype. However, when mat-forming cells are numerous, the ancestral type is favoured because it can reap benefits of the mat structure, without paying the cost of the cellulose over-production. It results in coexistence of the two phenotypes via time-lagged frequency dependant interactions. This adaptive diversification process is the starting point for the study of collective-level reproduction found in Chapter 4.

To illustrate this phenomenon in the toy model, demographic rates must somehow be made dependent on the trait distribution. This relates to the dimension  $A$  of the Darwinian space. As an example, let  $A$  be a one-dimensional parameter in  $[0, 1]$  that parametrises the per-capita death rate  $d$ . High values of  $A$  means that competition between individuals is limited to similar traits, while lower values means that the competition is more diffuse in the trait space:

$$d(\theta) = \sum_{x \in \Theta} N_x \Gamma(x, \theta, 1 - A) \quad (2.1)$$

Where  $N_\theta$  is the number of individuals carrying the trait  $\theta$ . The distribution of offspring is identical to the previous section (with  $H = 0.9$ ), as well as the birth rate (with  $S = 0.8$ ). The initial condition is a monomorphic population of  $K$  individuals.

Figure 2.6 presents the diversification pattern for increasing values of  $A$ . When  $A$  is low and competition is diffuse among the population, the pattern is not qualitatively different from the previous section, with a trait distribution centred around the optimal trait value. However, when  $A$  increases, the trait value distribution segregate into two stable sub-populations. This is a phenomenon known as branching. Adaptive dynamics can be used to find necessary conditions for branching (Geritz et al., 1998).

## Insights from the toy-model

This illustrative toy model, and the basic qualitative analysis conducted in the previous section is naturally limited in terms of new insights. However, it clarifies a few important points for the rest of the discussion.

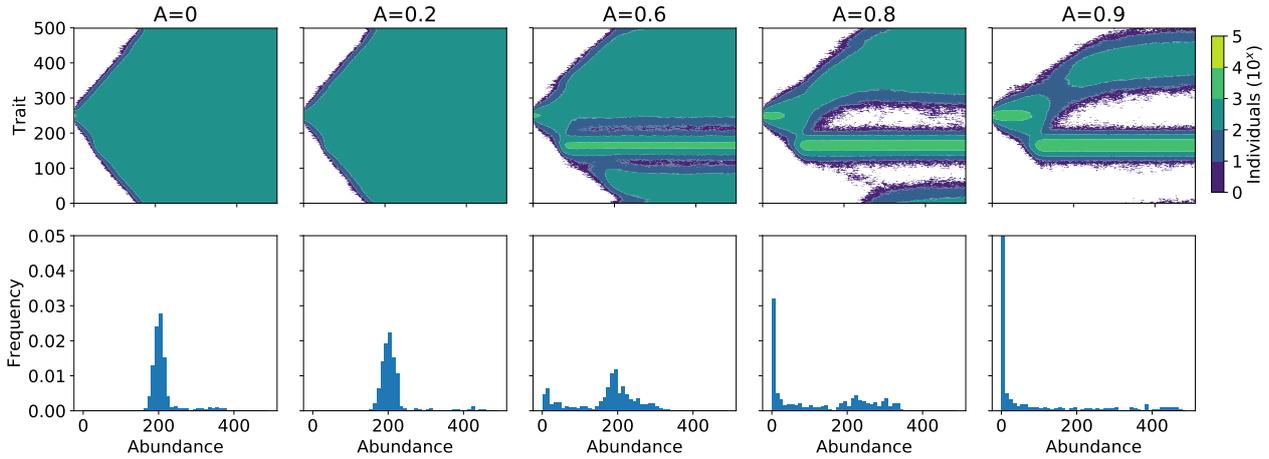


Figure 2.6: **Adaptive diversification along the competition axis.** *Top:* Distribution of traits value through time. *Bottom:* Frequency distribution at the end of the simulation.  $V = 500$ ,  $H = 0.9$ ,  $S = 0.8$ ,  $\theta_{opt} = 200$ .

First, it shows that population patterns relevant to evolutionary biology (such as neutral diversity, adaptation and adaptive diversification) are not just phenomenological, but result of simple population mechanisms. Those mechanisms can be formalised in two classes: demography and trait dynamics. Second, it illustrates the basic corollary of Darwinian properties: it is possible to imagine a population with both a demography and a trait dynamics that lacks Darwinian properties ( $H = 0$ ,  $S = 0$ ,  $A = 0$ ), and as a result, does not present the patterns mentioned above.

More precisely, this model is a way to build intuition about the quantification of Darwinian properties, and the concept of Darwinian space. Populations can be roughly ranked onto several dimensions with a progression from non-Darwinian, to marginal, minimal and paradigm. Linking this space to parameters of a mechanistic model is a non-trivial exercise and the avenue explored in this section should be challenged and improved. This model uses parameters that control a gradation from uniform to local effects in the trait space for each Darwinian dimension (illustrated by the function  $\Gamma$ , Figure 2.3). This is conceptually simple, but has shortcomings. For instance, high values of parameters routinely yields constant trait distribution ( $H = 1$ ) or non-viable population ( $S = 1$  if the initial condition does not include  $\theta_{opt}$  individuals). Nonetheless, this approach is fruitful because it points out which patterns are expected to be robust in less-than-perfectly Darwinian populations. Neutral diversity requires high  $H$ , adaptation additionally requires a high  $S$ , and adaptive diversification a high  $A$ .

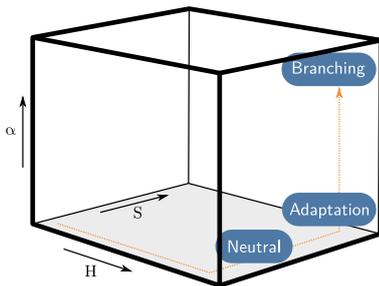


Figure 2.7: **Darwinian Space Revisited.** The orange line follow the path illustrated in this chapter.

Overall, this section can be summarised by a path in Darwinian space (Figure 2.7). Starting from a non-Darwinian or very marginal case ( $H = 0$ ,  $S = 0$ ,  $A = 0$ ), increasing  $H$  leads to neutral diversity pattern. From there, increasing  $S$  allows the existence of adaptation. Finally, increasing  $A$  from this point reveals adaptive diversification. This path is merely didactic though: nothing in this model implies that it reflects the trajectory of natural

populations in Darwinian space.

#### **A side note about fitness**

At this point of the manuscript, the reader might have noticed the relatively low importance accorded to the concept of *fitness* compared to the central role it has in mainstream evolutionary theory. This is a deliberate epistemological choice, that I would like to briefly justify here. I think, building on Doebeli et al. (2017), that the core ingredients of eco-evolutionary dynamics are demography (births and deaths), and trait dynamics (reproduction and development), not “fitness”. Of course, fitness measures are important quantities that arise from a variety of models. They are also derived, information-poor, quantities that obscure underlying phenomena, like the difference between birth and death, or context-dependency (such as frequency-dependant selection, except in specific case, like the *invasion fitness* of adaptive dynamics). As a result, I avoid using fitness as a fundamental quantity and rely only on fitness measures, justified by a stochastic process, where it brings clarity to an argument.

Now that all the necessary concepts of Darwinian properties have been laid out for a given level of organisation, the last section turns to multiple levels, and precisely defines the mechanism for the emergence of collective-level Darwinian properties that is the subject of this manuscript.

### **Ecological scaffolding: a mechanism for the emergence of complexity**

Biological life is organised in various, roughly hierarchical, scales: genes, chromosomes, cells, organisms, eusocial populations. . . all these entities manifest to an extent Darwinian properties and thus participate in the process of evolution by natural selection. These structures do not emerge from nothing. They are nested in a way that partly recapitulates their history: self replicating molecules predate cells, in the same way, cells necessarily predate multicellular organisms, and organisms predate eusocially organised populations. However, this linear view is a simplification. The emergence of a new level of organisation is more than a juxtaposition of lower-level entities, with causality flowing unidirectionally from individuals to collectives. To quote Levins and Lewontin (1985), “parts makes wholes, and whole makes parts” (p. 272), both levels are interlocked and the evolution of one partly drives the other. In other words, multicellular organisms are more than a collection of unicellular organisms. Untangling the process by which new level of organisation come to be, a rare event in the history of life, called Major Evolutionary Transition (Szathmáry and Maynard Smith, 1995) or Evolutionary Transition in Individuality (Michod, 2000; Godfrey-Smith, 2009) is a challenge for evolutionary biology.

The crux of the matter is the emergence of a new level of Darwinian population. To illustrate, consider again the transition to multicellularity.

Multicellular organisms such as animals are a prime example of paradigm Darwinian population. They have clear mechanisms ensuring their Darwinian properties: meiosis and fecundation are mechanisms that ensure the variation, reproduction and heredity at the level of the multicellular organism. They are distinct from the mechanism of mitosis that is found at the level of the cell: they do not come for free as a simple by-product of lower level processes (Griesemer, 2001).

In evolutionary biology, the causal explanation for complex, adapted teleonomic structures is usually natural selection. The problem is that this mechanism necessitates Darwinian properties at the considered level. Thus, it cannot be invoked from the onset of an evolutionary transition. In order to avoid circular reasoning, Darwinian properties cannot be used to explain their own origins. The initiation of a major evolutionary transition is necessarily the emergence of Darwinian properties in non-Darwinian entities, therefore it is a non-Darwinian process (Black et al., 2019).

However, these initial Darwinian properties do not need to be paradigmatic. If there is a mechanism that can promote the emergence of *marginal* Darwinian properties at the collective-level, then natural selection can be invoked as a process for their further evolution, and potentially their refinement toward *paradigm*-level (see Figure 2.8). A first way collectives can gain marginal Darwinian properties is by co-opting lower level traits. This means that lower-level mechanisms are used at the higher-level for a different purpose. This constitutes a prime example of the “tinkering” nature of evolutionary processes (Jacob, 1977). Examples include the co-option of ancestral cell-cycle regulation mechanisms in volvocine green algae for the formation of groups via cell–cell adhesion (Hanschen et al., 2016), and the co-option of ancestral apoptosis mechanisms in the experimentally derived snowflake yeast for the fragmentation of cell clusters (Ratcliff et al., 2012). However, co-option might not always be achievable from the onset, considerably increasing the challenge of a transition.

There is however a mechanism that does not necessitate the *a priori* co-option of lower-level traits to give collective marginal Darwinian properties: *ecological scaffolding* (Black et al., 2019). Rather, it relies on their exogenous imposition by the population structure. As a first example, consider a case where the habitat is structured in patches because of rare resources, the existence of physical support, or physical boundaries: spatial structure define collectives. Additionally, suppose that particles composing collectives can, relatively rarely, disperse to other patches. Finally, consider that either new patches of resources are introduced in the environment, or that catastrophic events can lead to the demise of whole collectives, freeing resource patches.

As a consequence, it becomes possible to define a birth-death process at the level of the collectives. Birth events happen when a new patch is colonised by particles, and death events when the population of a patch goes extinct. If migration between patches are limited, it is possible to also define a genealogy at the level of collectives, by tracking the flow of particle: a collective is the offspring of the collective from which its founding particles are originally from. The notion of collective-level genealogy is explored in details within Chapter 4. In essence, and to use the terms defined at the beginning of this chapter, this defines a demography at the level of collectives. Similarly,

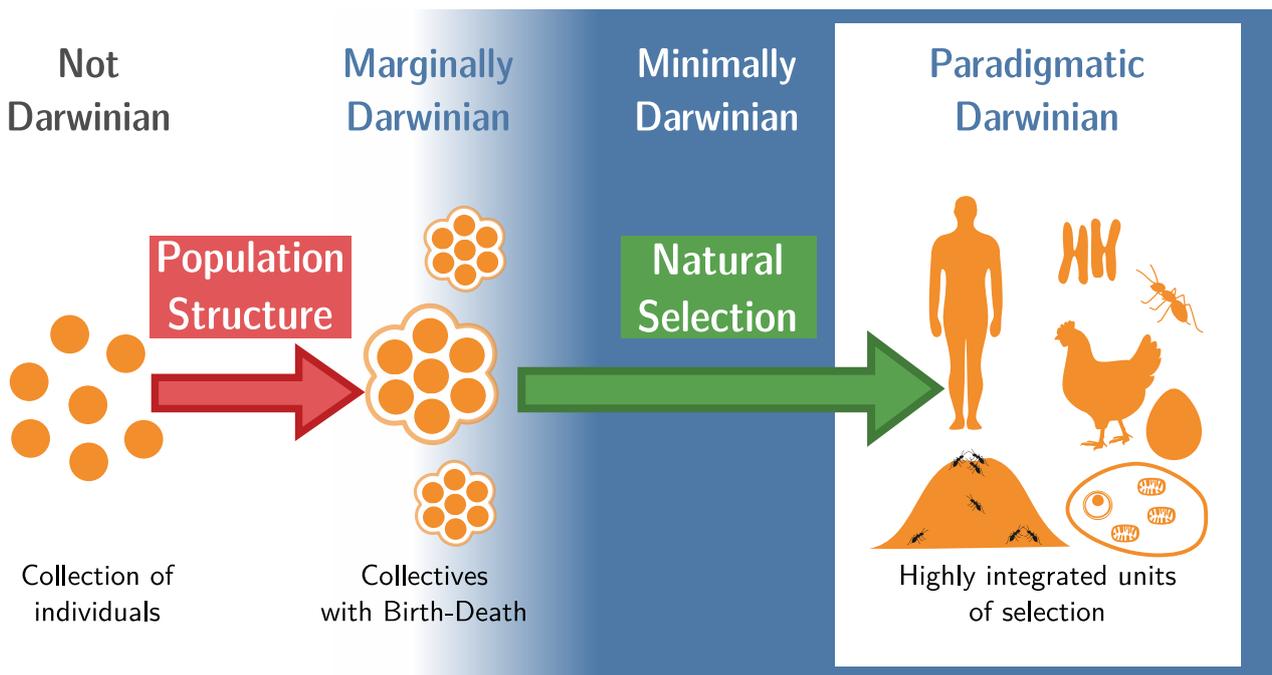


Figure 2.8: **Ecological recipe for a major evolutionary transition.** This two-step scenario for a major evolutionary transition consist in (1). A population structure phase during which collectives of particles are formed. Collectives are subject to a birth-death process, bestowing them marginal Darwinian properties. (2). Natural selection can build up on those marginal Darwinian properties to refine Darwinian properties up to paradigmatic Darwinian collectives.

particle composition is a collective-level trait. The ecology of particles within the patch of resource defines its development dynamics (in the sense given at the beginning of this chapter), and conversely, the way of dispersal of particles defines the collective reproduction mechanism.

This population structure ensures a marginal level of Darwinian properties at the collective level: collectives may vary in their particle composition, their demographic dynamics can depend on this composition, and since new collectives are founded by sampling, particles from the parent collective, the composition can recapitulate the parental phenotype. This is one of the core themes of this manuscript: Chapter 3 study how collectives of Darwinian particles diverge in their composition by neutral evolution. Chapter 5 explores how collective-level demography can depend on particle composition. Finally, Chapter 6 explores how particle composition may become more hereditary by the effect of natural selection.

Figure 2.9 sketches three examples of population structures that lead to ecological scaffolding. The first consists in meta-population of bacterial mat at the air-liquid interface structured by ponctual support (e.g., reeds). It is inspired by extensive studies of such a system in *Pseudomonas fluorescens* in relation with the origins of multicellular life (Rainey and Travisano, 1998; Rainey and Kerr, 2010; Hammerschmidt et al., 2014) that is studied in more

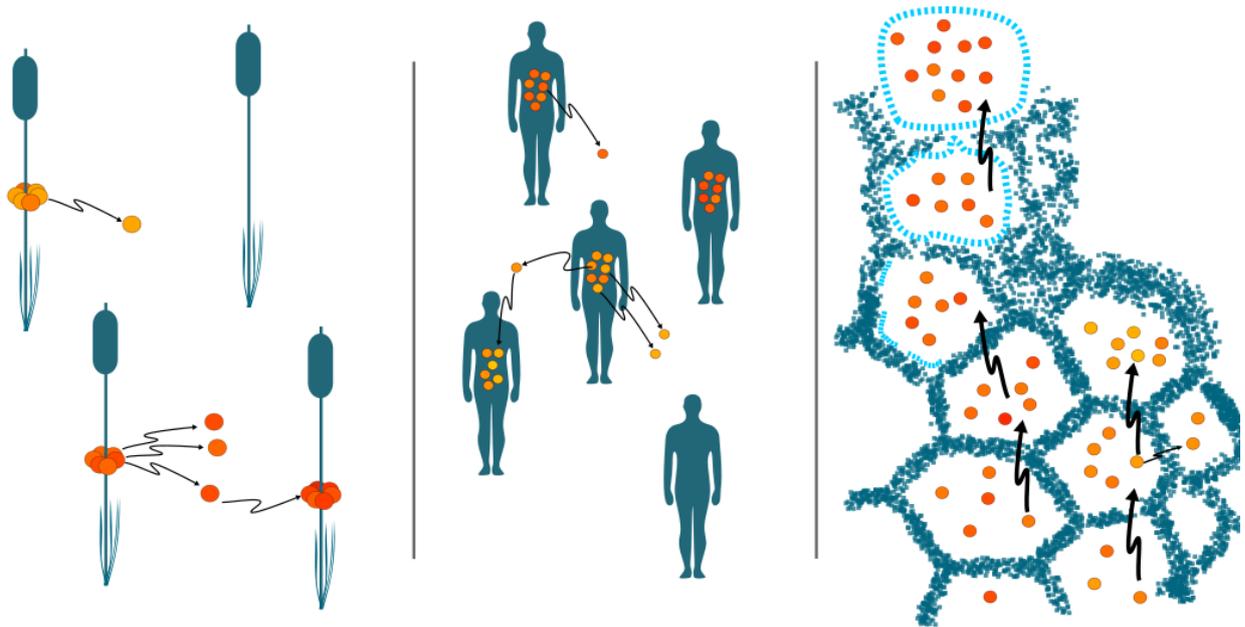


Figure 2.9: **Examples of Ecological Scaffolding.** The environment provides boundaries (blue) that highly constraint the structure of an underlying Darwinian population (red-orange) bestowing Darwinian properties to the collectives themselves. **Left:** Pond scum such as *Pseudomonas fluorescens* forming mats on reeds at the air-liquid interface. **Middle:** Pathogens or parasites colonising human hosts. **Right:** (Redrawn from Martin and Russell (2003)) Early biological molecules in an iron sulphide precipitate in a submarine hydrothermal vent.

details in Chapter 4. The second is inspired by the population structure of infectious bacteria where the rate of infection are low enough and the inoculum small enough so that the population consist in an array of isolated lineages or clonal clusters (Levin, 1981; Maynard Smith et al., 2000; Feil, 2010). The last illustrates a well-supported theory for the origins of the cell: three-dimensional chambers formed by iron sulfide precipitates in hydrothermal vents offer the literal scaffold for early cell content (Martin and Russell, 2003).

Once marginal Darwinian properties are displayed by the collective, the second part of the ecological scaffolding mechanism is the refinement of those properties. Consider a variant collective that would present enhanced (more paradigm) Darwinian properties such as a more faithful reproduction or an independence on the scaffold (See the apparition of the plasma membrane, light blue in figure 2.9). This variant would have an advantage in the collective-level population, and increase in proportion by the mechanism of natural selection. Traits improving Darwinian properties at the collective level include a reproductive *division of labor* (See Chapter 5) and the emergence of a reliable developmental programme (See Chapter 6). This process can involve the co-option of lower-level mechanisms: in the example of bacterial mats of *Pseudomonas fluorescens*, the proto-reproductive division of labour co-opt the existing ability of cells to over-produce cellulose (See Chapter 4).

Another long-term consequence of major evolutionary transition is the “De-Darwinisation” of the lower level entities. While cells within multicellular organisms reproduce, and vary in their reproductive rate, they are arguably “less paradigm Darwinian”, as cells, than unicellular eukaryotes (Godfrey-Smith, 2009, p. 124). The single cell bottleneck that constitute the zygote means that the population of cells within a multicellular organism stands low on the variation dimension  $V$  of the Darwinian space. Additionally, the reproductive fate of cells is mainly controlled by signalling pathways that depend more on the context than their intrinsic characteristics, such as their position in relation with other cells. Thus, populations of cells within multicellular organisms are also relatively low in the  $S$  dimension (Godfrey-Smith, 2009, p. 56). Of course, cells within multicellular organisms have not lost all their Darwinian properties, they are still at least minimally Darwinian. As a result, natural selection can still act at the level of the cell, for instance in cancers or in the adaptive immune system (Godfrey-Smith, 2009, p. 103).

Successful ecological scaffolding in nature is likely rare, given the relatively small number of major evolutionary transitions documented today. This might come from the relative exceptional nature of the required population structure, but also from the fact that transition in individuality are not harmonious phenomenon from the onset and many evolutionary conflicts must be overcome to see it to completion (Buss, 1987; De Monte and Rainey, 2014). For instance, division of labour is not an unconditional consequence of a transition: over 23 monophyletic protist groups, 17 of them display multi-cellularity (thus endogenous collective Darwinian properties) but only three (plants, animals and fungi) have a meaningful cell differentiation as pointed out by Buss (1987). Conversely, some strains of pathogens (*Salmonella*, Stecher et al. (2007)) might display a division of labour without formal boundaries at the collective level.

To summarise, one can imagine an *archetypal ecological scaffolding scenario* and use this simplified version as the basis to explain the origins of Darwinian properties at higher levels of organisation that fuel Major Evolutionary Transitions (Figure 2.10). This scenario is characterised by three externally imposed constraints: (1) *boundaries* that insure a complete insulation of the populations (no inter-demic migrations), (2) a *collective birth-death process* with non overlapping collective generations, and (3) a collective reproduction mechanism relying on the *random sampling of individuals from the parental collective* to fund offspring collectives. One of the prime advantage of such a setup is that it can be readily implemented experimentally in microbiology. First, tubes or millifluidic compartments provide boundaries to cultures. Second, the experimenter can implement collective birth and death of cultures by transferring cells from parent cultures into fresh medium (serial transfer).

This population structure constitute the initial, non-Darwinian step of the evolutionary transition in individuality. Once it is effected, natural selection of more complex teleonomic mechanisms is possible. Overall, this scenario constitutes the mould from which all models within this manuscript are derived.

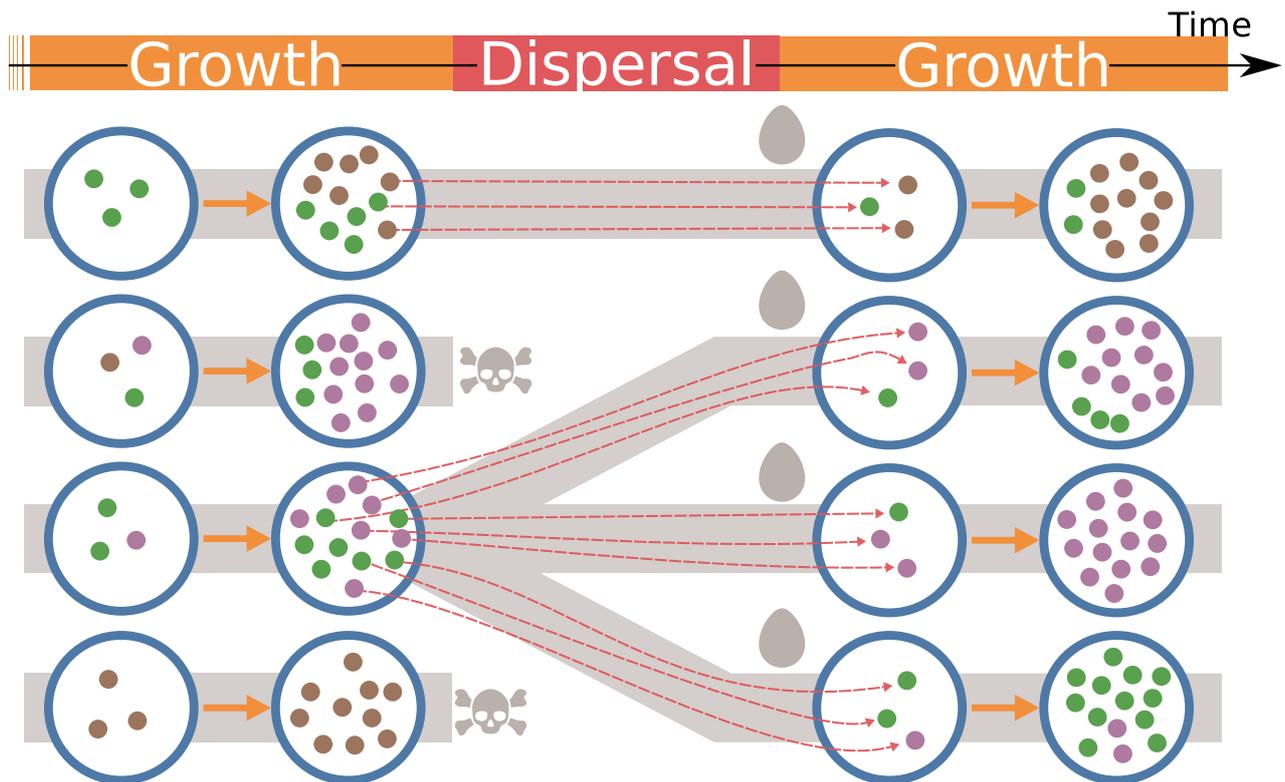


Figure 2.10: **The archetypal ecological scaffolding scenario.** *blue circles:* Collectives with well-defined artificial boundaries (e.g., tubes or millifluidic compartments), *small discs:* Particles (e.g., bacteria). The three phenomenon to model are: (1) The intra-collective dynamics or particle ecology (thick orange arrows), (2) the collective birth-death process (large background lines) and (3) the serial transfer of the particles (red dashed arrows).

## Chapter 3

# Neutral Diversity in Experimentally Nested Populations

“Mais une fois inscrit dans la structure de l’ADN, l’accident singulier et comme tel essentiellement imprévisible va être mécaniquement et fidèlement répliqué et traduit, c’est-à-dire à la fois multiplié et transposé à des millions ou milliards d’exemplaires. Tiré du règne du pur hasard, il entre dans celui de la nécessité, des certitudes les plus implacables.”

— JACQUES MONOD, *Le Hasard et la Nécessité* (1970)

EXPERIMENTAL EVOLUTION is the study of evolutionary dynamics happening in real time as a response to conditions imposed by the experimenter (Kawecki et al., 2012). Microbial populations are widely used because they offer numerous experimental advantages: large population sizes, easily manipulable environments, possibility to freeze and store whole populations indefinitely... Experimental evolution requires the set up of many parallel bacterial cultures that can take several forms from bottles ( $\approx 10^1 L$ ) to tubes ( $\approx 10^{-3} L$ ), to microtater plates ( $\approx 10^{-4} L$ ), to microfluidic compartments ( $\approx 10^{-9} L$ ).

Recently, new techniques for the high-throughput manipulation of bacterial populations have emerged. For instance digital microfluidics (Cottinet, 2013; Dupin, 2018; Doulcier, 2019) allows the possibility of producing and imaging thousands of droplets of culture broth within a carrying fluid. The droplets amount to around  $2 \times 10^{-6} L$  with a carrying capacity of  $10^5$  to  $10^6$  cells, in the current version of the technology (Cottinet, 2013). Droplets can be imaged and some quantitative measures can be performed (optical density, fluorescence signal...) during growth of the bacteria, allowing a high-throughput monitoring of ecological dynamics.

Nested populations in which both particles and collectives are individuals with their own birth and death events can be readily implemented in experimental microbiology, for instance, Chapter 4 presents an experiment performed in tubes. However, the ability of microfluidic devices to monitor in the order of a thousand of cultures and retrieve some of them for analy-

sis makes them particularly suitable for the experimental implementation of ecological scaffolding (as presented in Chapter 2).

Neutral diversity in nested populations is the focus of this chapter. The aim is to build a quantitative understanding of simple diversity patterns within the experimental setup. First, a model of the device is presented. It relies on the assumption that cells are in constant exponential growth. The optimal operating regime parameters of the machine (dilution ratio, duration of collective growth cycles, carrying capacity...) are derived from characteristics of the biological material: birth and death rates. Second, a coalescent model of the population across bottlenecks is proposed and coupled to a neutral mutation model with infinite alleles. This allows computation of the number of mutations, and the distribution of allele frequencies within droplets after several collective growth cycles. It shows that small bottlenecks are required to maximise diversity in one cycle, but larger bottlenecks are more favourable for diversification across many cycles. The speed at which diversity accumulates decreases with time. Then, the effect of splitting a droplet into several lineages is studied by computing the number of mutations accumulated in a single, or all the droplet lineages. Finally, a simple mutation accumulation model illustrates the interest of droplet-level selection for artificial selection.

### 3.1 Modelling Nested Population Dynamics

	Description
$D$	Population size
$T$	Cycle duration
$n$	Number of Cycles
$K$	Carrying capacity
$c$	Initial number of particles

Table 3.1: Collective parameters

	Description
$b$	Birth rate
$d$	Death rate
$r$	Malthusian parameter (b-d)
$\delta$	Survival probability (at a bottleneck)
$\theta$	Mutation rate

Table 3.2: Particles parameters

Consider a device that allows the manipulation of collectives of Darwinian particles via serial transfers (Figure 3.1). Cells (or particles) are distributed among a train of  $D$  droplets. The birth and death of cells are modelled by a linear branching process with constant rates  $b$  for birth and  $d$  for death. The net birth rate  $r := b - d$  is called the Malthusian parameter. After a duration  $T$ , a new train of  $D$  droplets is prepared by diluting them  $\frac{1}{\delta}$  fold. Hence, for each dilution event, a cell has a probability  $\delta$  of being sampled and thus being present in the new droplet. This procedure is repeated periodically, each new growth phase followed by a dilution constitutes a *cycle* of the experiment or a *collective generation*.

Birth  $b$  and death  $d$  rates depend on the biological material used (species, strain...) as well as the culture medium and are not easily controlled. However, the duration of the growth phase  $T$ , the dilution factor  $\delta$  and the number of collectives  $D$  can be changed by altering the experimental setup. A model can help predict the effect of those parameters and find the ones that should be the focus of engineering efforts.

#### 3.1.1 Optimal Operating Regime

When designing a serial transfer experiment, the operator has three main parameters that might be controlled: the size of the cultures (and by extension the carrying capacity of the particles  $K$ ), the duration of the growth phase separating two successive transfers  $T$ , and the dilution rate  $\delta$ . Two problems must be avoided: if populations sizes are too small and dilution too high, the resulting cultures might be empty. Conversely, if the population sizes are too large, and dilution too low, the population will spend most of its time in stationary phase.

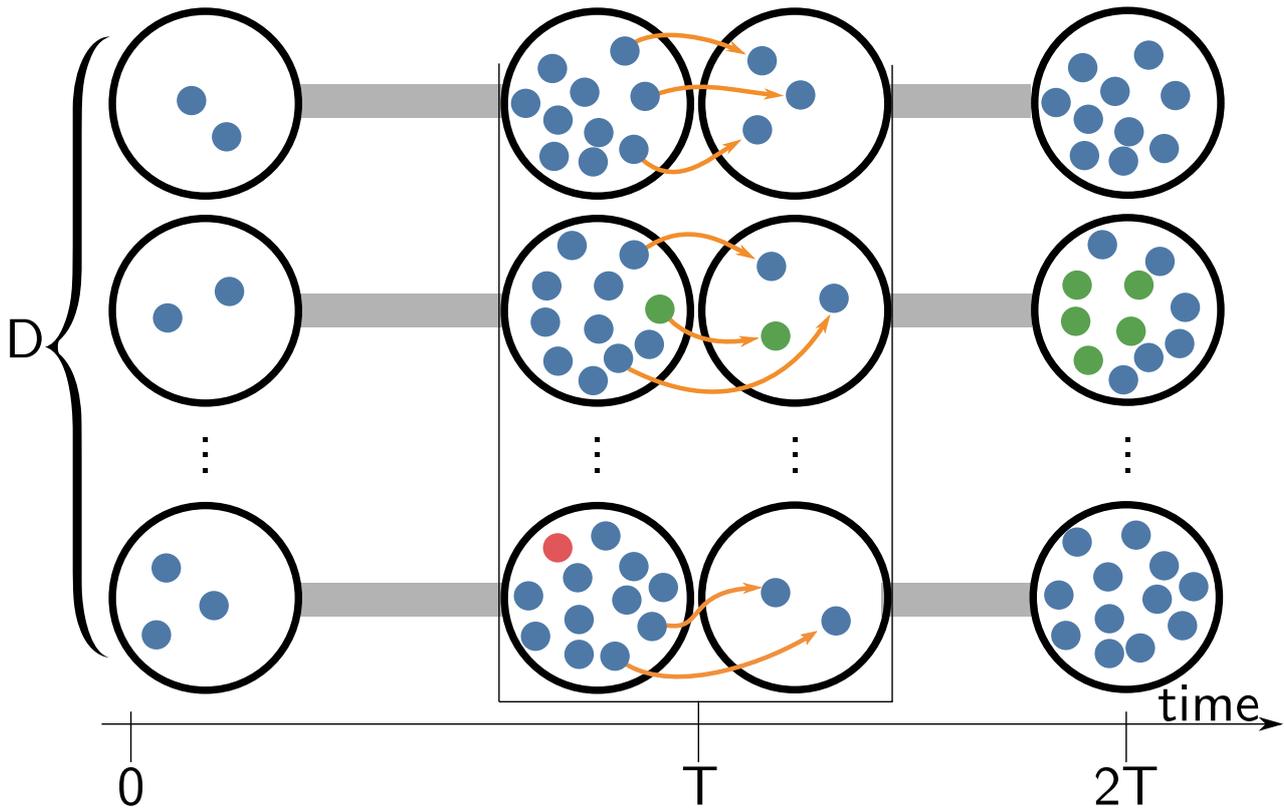


Figure 3.1: **Sketch of the experimental setup** Darwinian particles following a birth-death process with rates  $b, d$  are distributed in collectives within  $D$  droplets. After a growth phase duration  $T$ , the content of each droplet is diluted to form a new collective generation (or cycle). Each particle has a probability  $\delta$  to be transferred in the next cycle. Here the *serial transfer* regime is depicted: each droplet is diluted into exactly a single new droplet in the next cycle. In the full *nested population* design, a droplet can be split in several droplets in the cycle, or removed altogether.

Any dilution event presents the risk of extinguishing the population. When performing a serial transfer experiment, this must be avoided at all cost because an empty microcosm signs the end of the experiment (at least for the given independent lineage). In a nested population design, the presence of some empty microcosms can be tolerated because empty niches in the population can be filled by splitting a single parent droplet into several offspring droplets in the next generation.

Stationary phase is not desirable in general for several reasons. First, a population that reach saturation will go through fewer generations than if it was growing freely, reducing the potential evolutionary dynamics. Moreover, physiological changes in stationary phase might result in undesired phenotypic effects on the population. Finally, in the case of millifluidic experiments, saturating densities are known to increase the risk of cross-contamination between droplets.

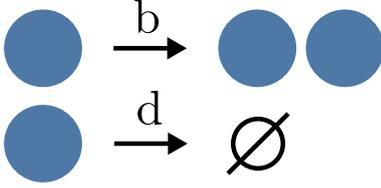


Figure 3.2: **Events in a Linear-Birth-Death Model.** Individuals give birth to new individuals at a constant rate  $b$ , and die at constant rate  $d$ . The process is supercritical if  $b > d$ .

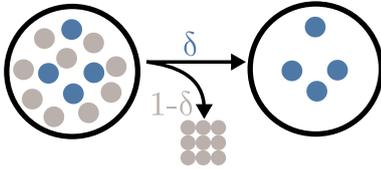


Figure 3.3: **Dilution process.** Individual particles are independently selected to be transferred to the next cycle (with probability  $\delta$ ) or sent to the waste (with probability  $1 - \delta$ ).

For all these reasons, there is an optimal dilution rate, that keep the population in exponential phase while maximising the population size, at which selection experiments should be conducted.

A model of population dynamics can provide a first guess at finding the optimal range of parameters to conduct a given experiment. In the following, a stochastic model of particles in exponential growth condition (i.e., supercritical) with periodic bottlenecks is used to derive the probability of losing a single particle lineage, or a single collective lineage due to the effect of dilution, as a function of experimentally accessible parameters. Saturation phenomena are not modelled explicitly as the birth and death rates are considered independent of population size, but the population dynamics are required to stay under a carrying capacity threshold.

### Survival of a single lineage

A first quantity that can be derived from the linear branching process with periodic dilution that models the population dynamics is the probability that a single initial particle has no descent in the population after  $n$  cycles.

#### Proposition 1 (Survival Probability):

Cells within droplets in serial transfers are modelled by a linear birth-death process with constant parameters  $b$  and  $d$ , that is subject to periodic bottlenecks every duration  $T$ .

Let  $s_n$  (respectively  $\bar{s}_n$ ) be the probability that a lineage spawned by a single cell is not extinct at time  $nT$ , just before the  $n$ -th dilution (respectively just after the  $n$ -th dilution). Then,

$$s_n = 1 - h_{Q^{n-1}R}(0)$$

$$\bar{s}_n = 1 - h_{Q^n}(0)$$

Where  $h_A$  is the linear fractional function with coefficient  $A$ :

$$h \begin{bmatrix} a & b \\ c & d \end{bmatrix} (s) = \frac{as + b}{cs + d} \quad (3.1)$$

And the matrices  $Q$  and  $R$  are:

$$Q = \begin{bmatrix} \delta(p - q) & p - \delta(p - q) \\ \delta(p - 1) & p - \delta(p - 1) \end{bmatrix}; \quad R = \begin{bmatrix} \delta(p - q) & q \\ \delta(p - 1) & 1 \end{bmatrix}$$

		$p(b, d, T)$	$q(b, d, T)$
Subcritical particles	$b < d$	$\frac{-r}{d - be^{rT}}$	$\frac{d(1 - e^{rT})}{d - be^{rT}}$
Critical particles	$b = d$	$\frac{1}{1 + bT}$	$\frac{bT}{1 + bT}$
Supercritical particles	$b > d$	$\frac{re^{-rT}}{b - de^{-rT}}$	$\frac{d(1 - e^{-rT})}{b - de^{-rT}}$

(Proof page 52.)

Proposition 1 shows that the survival probability of a lineage depends on the birth  $b$  and death  $d$  rates of the cells, but is also a function of the dilution rate  $\delta$ , duration of the growth phase  $T$  and the number of cycles  $n$ . When considering a single cycle ( $s_1$ , Figure 3.4), the survival probability increases with  $\delta$  and  $b$ .

The numerical computation of  $Q^n$  might be problematic because of repeated multiplication of the small numbers. However, since the final result only involves the ratio  $Q_{01}^n/Q_{11}^n$ , it is possible to normalise  $Q$  to have its smallest value being 1. Indeed, this ratio does not depend on a multiplicative scalar on the matrix:  $\forall \alpha > 0, h_{Q^n}(0) = h_{\alpha Q^n}(0)$ . Taking  $\alpha = \frac{1}{\delta}$  greatly improves the numerical stability of the computation.

The limit of this probability when the number of cycles increases gives a clearer understanding of the long term behaviour of the population:

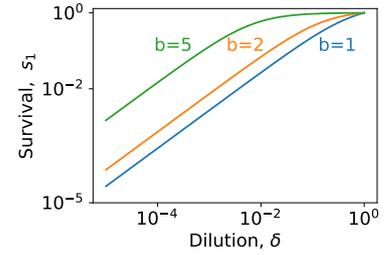


Figure 3.4: **Survival probability for one cycle  $s_1$**  The probability is presented as a function of the dilution rate  $\delta$  for pure birth processes  $b > 0, d = 0, T = 1$ .

**Proposition 2 (Long Term Survival Probability):**

Let  $\bar{s}_n$  be the survival probability after  $n$  cycles of the lineage spawned by a single cell.

$$\lim_{n \rightarrow +\infty} \bar{s}_n = \begin{cases} 0 & \text{if } b \leq d \text{ or } \delta < e^{-(b-d)T} \\ \frac{(b-d)(\delta - e^{-(b-d)T})}{b\delta(1 - e^{-(b-d)T})} & \text{otherwise.} \end{cases}$$

(Proof page 54.)

Proposition 2 shows that, in the long run, lineages go extinct with certainty ( $\bar{s}_\infty = 0$ ) if the death rate of the particle is higher than their birth rate (i.e., the associated branching process is not super-critical) or if the sampling-at-dilution probability  $\delta$  is lower than a threshold that corresponds to the inverse of the expected growth of the population during one cycle  $e^{rT}$  (Figure 3.5).

In the following, only viable super-critical populations with a birth rate higher than their death rate will be considered ( $b > d$ ).

**Optimal cycle duration and dilution**

Saturation of particle dynamics is not desirable, as mentioned earlier. Depending on the nature of the particles (species, strain), and of the medium (pH, nutrient availability, temperature), it is possible to define an experimental carrying capacity  $K$  that corresponds to the number of cells that can be sustained in a droplet without saturation. The simple linear-birth-death model cannot represent saturating population, thus for the model to be coherent, the duration of the growth phase must be short enough so the population size does not reach the carrying capacity:

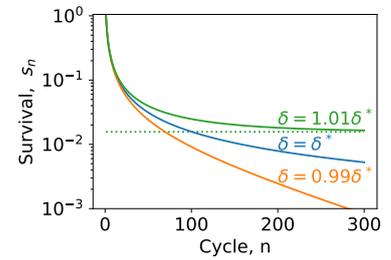


Figure 3.5: **Survival probability for several cycles  $\bar{s}_n$**  The probability is presented for pure birth processes  $b = 1, d = 0, T = 1$ . Above the critical threshold  $\delta^*$ , the survival probability does not tend toward 0. Dotted line corresponds to the limit  $\bar{s}_\infty$ .

**Proposition 3 (Maximal Cycle Duration):**

Let  $T^*$  be the maximal cycle duration before reaching saturation. Cells are following a supercritical birth-death process with rates  $(b-d) = r > 0$ . The carrying capacity is  $K$  and the initial number of cells is  $c$  :

$$T^* = -\frac{\ln(\frac{c}{K})}{r} \quad (3.2)$$

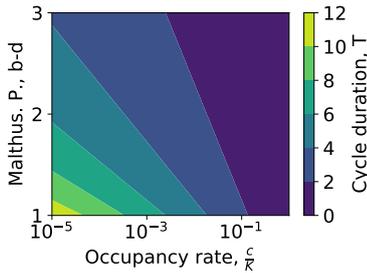


Figure 3.6: **Maximal Cycle Duration**  $T^*$  as a function of the Malthusian parameter  $b - d$  and the occupancy rate  $\frac{c}{K}$ .

Proposition 3 shows that the optimal duration of the growth phase is linear with the inverse of the Malthusian parameter  $r$  of the population (Figure 3.6), meaning that a population that grows (on average) twice as fast as another should be subject to cycles half as long as the other, for a given occupancy rate  $\frac{c}{K}$ .

Additionally, the optimal duration of the growth phase is proportional to the logarithm of the occupancy rate  $\frac{c}{K}$  of the droplet (with a minus sign since this logarithm is always negative or null as  $c \leq K$ ). As a consequence, multiplying the volume of the droplets by two, or dividing the inoculum size by two for a given strain will increase the optimal duration of the growth phase by  $\frac{\ln 2}{r}$ .

Normalising by the Malthusian parameter ( $r = 1$ ) shows clearly that doubling the volume of the droplets increases the duration of the experiment by around two-thirds ( $\approx \times 1.7$ ) which is equivalent to reducing the growth rate by a third ( $\approx \times 1.5$ ).

This result holds for a single cycle only. For a given dilution rate  $\delta$ , the population is shrunk by an expected ratio  $\delta$ , while for a given cycle duration  $T$ , the population is expanded by an expected ratio  $\delta e^{rT}$ . In order to prevent the population from saturating for all cycles, the occupancy ratio at the beginning of each cycle must be constant. This consideration allows discovery of the optimal dilution rate when the growth phase duration is fixed:

**Proposition 4 (Optimal Dilution Rate):**

Let  $\delta^*$  be the optimal dilution rate for which the expected number of cells is constant across generations. For cells following a optimal birth-death process with rates  $(b-d) = r > 0$  and a growth phase duration  $T$  :

$$\delta^* = e^{-rT} \quad (3.3)$$

If  $T = T^*$  (Proposition 3),

$$\delta^* = \frac{c}{K} \quad (3.4)$$

(Proof page 54.)

Proposition 4 shows that the dilution sampling probability should be equal to the initial occupancy rate when the duration of the cycle is maximal.

To summarise, the optimal operating regime of the experiment can be expressed from the Malthusian parameter of the population and the initial occupancy rate of the particles  $\frac{c}{K}$ . As a result, the dilution sampling probability is  $\delta^* = \frac{c}{K}$  and duration of a cycle is  $T^* = -\frac{\log(\delta^*)}{r}$ . Fixing any two of  $(c, K, T, \delta)$  values constrains the other two.

When the experiment is in the optimal regime, the expression of the survival of a lineage (not including sampling at the last cycle) is simpler:

**Proposition 5 (Optimal Regime Survival Rate):**

Let  $\bar{s}_n^*$  be the survival probability of a lineage after  $n$  cycles of duration  $T$  and with bottleneck  $\delta^* = e^{-rT}$ , where  $r = b - d > 0$  is the Malthusian parameter of the population.

Then,

$$\bar{s}_n^* = \frac{1}{\frac{b}{r}n(1 - \delta) + 1} \quad (3.5)$$

(Proof page 55.)

Proposition 5 shows that, in the optimal regime the survival probability of a lineage decrease in the inverse of the number of cycles and tends toward zero for a large number of cycles.

Additionally, when taking a finite number of cycles  $n$  the survival does not tend toward zero even for vanishingly small bottlenecks  $\delta$ . This derives from the fact that, in the optimal regime, a small bottleneck is compensated by a long cycle duration  $T^*$ , so vanishingly small bottlenecks correspond to infinitely long cycles.

Finally, in the case of pure-birth (i.e.,  $d = 0$ ), the survival of a lineage is independent from the birth rate  $b$ . It is certain if there is no bottleneck ( $\delta = 1$ ), and tends toward  $\frac{1}{2}$  for one cycle ( $n = 1$ ) and vanishingly small bottlenecks ( $\delta \rightarrow 0$ ).

Overall, once the size of the droplets (which constrains  $K$ ) and the initial occupancy rate (which constrains  $c$ ) have been chosen by the operator, other parameters of the machine (duration of the growth phase  $T$ , dilution rate  $\delta$ ) can be deduced—and conversely, fixing  $T$  and  $\delta$  constrain  $K$  and  $c$ . The next section explores how should one select these parameters when the aim is to maximise genetic diversity within and between the droplets.

## 3.2 Modelling Neutral Diversity

Neutral diversity concerns mutations arising in the population of cells that are assumed to not change their birth or death rates. According to the neutral theory of molecular evolution (Kimura, 1983), most point mutations in the

DNA follow this dynamics. As seen in Chapter 2, neutral diversity gives rise to recognisable *patterns* that can be predicted from a mechanistic model of birth-death in the population. In experimental evolution, and moreover artificial selection, it is desirable to increase the diversity within the population because it allows greater exploration phenotypic space. Indeed, mutations that are essentially neutral for cells might present an interest for the experimenter, or be intermediate states toward new phenotypes.

In the following, mutations follow a Poisson Point Process with constant rate  $\theta$  over the lifespan of the cells independently of their genealogy. As a consequence, the time between two mutations along a lineage (regardless of births and deaths) is exponentially distributed, and thus has no memory: the conditional expected time to the next mutation will be the same for all cells, irrespective of their age or the time of the last mutation in the lineage. This is a simplifying assumption that represents the spontaneous nature of mutations, while ignoring the existence of mutations that can change the mutation rate (Sniegowski et al., 1997, 2000).

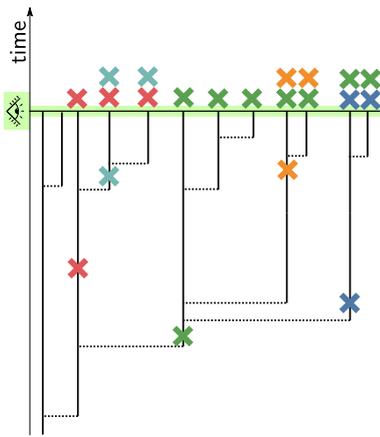


Figure 3.7: **Neutral mutations over the coalescent tree.** The neutral mutations (coloured crosses) are distributed following a Poisson Process over the (real) coalescent tree (black). Each individual may carry several mutation distinguishing it from the most recent common ancestor.

### Coalescent time and Coalescent Point Process

The linear branching process (with constant birth rate  $b$  and death rate  $d$ ) models the full genealogy of the population (Figure 3.8, left). However, the standing diversity at a given time in the population is not affected by the lineages that do not have extant individuals (because their mutations have been lost), nor ancestor to the most recent ancestor of the population (because their mutations are shared by all individuals in the population). Knowledge of the *coalescent tree* of the population (Figure 3.8, right), which is the genealogy of the extant individuals up to their most recent common ancestor is enough.

Coalescent Point Processes (Popovic, 2004; Lambert and Stadler, 2013) are stochastic processes whose realisations are real trees with the same probability as the coalescent tree of the corresponding branching process. A CPP is defined by a stopping time  $t$  and a branch length distribution  $f_H$ . The CPP is the sequence of independent and identically distributed variables  $(H_i)_{i=1\dots N}$  following  $f_H$  and stopped at the first element  $N$  such that  $H_N > t$ . Usually the branch length distribution is expressed in the form of the inverse tail distribution  $F$ :

$$F(t) := \frac{1}{P(H > t)} \quad (3.6)$$

### Measuring neutral diversity

Neutral mutations do not affect the genealogy nor the coalescent tree and can thus be added *a posteriori*. Consider that mutations appear following a Punctual Poisson process with constant rate  $\theta$  over the coalescent tree. Thus, a mutation is a point on the coalescent tree as illustrated in Figure 3.7. Additionally, suppose that reverse mutations are impossible (an assumption characterising so-called “infinite sites model”), it implies that all individual standing above the mutation in the coalescent tree (i.e., the descent of the mutation point) share the mutation (crosses at the top of Figure 3.7). Individuals may carry zero, one or several mutations.

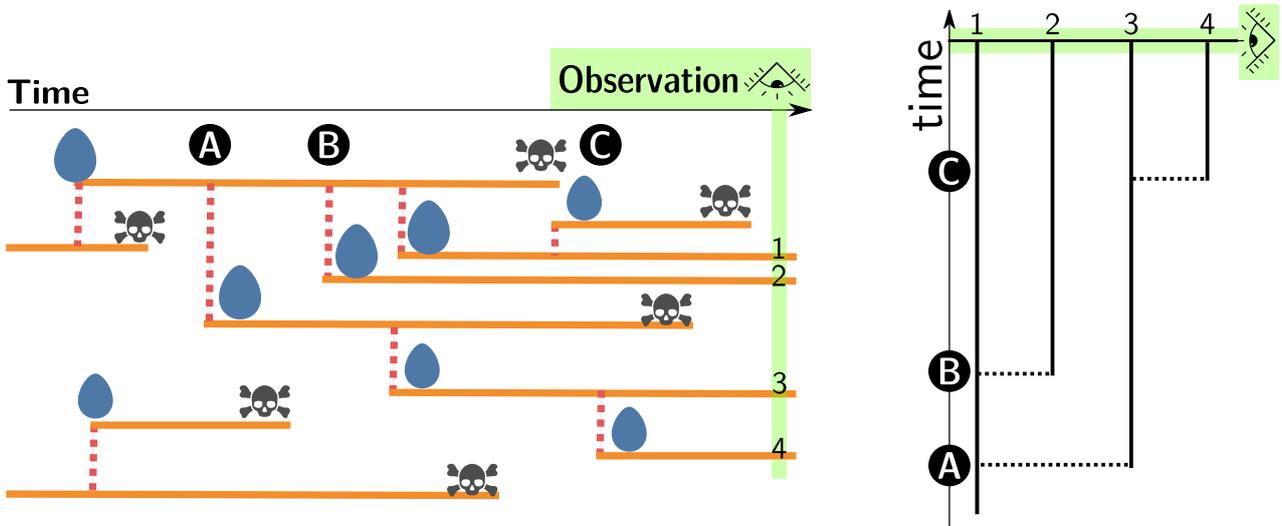


Figure 3.8: **From Birth-Death Process to Coalescent Point Process.** On the left side is a birth-death process where a number of individual give birth (eggs) and die (skulls) at different point in time. On the right side is the corresponding continuous coalescent tree. Note that at time  $C$  the lineage 3 coalesce with lineage 4 and that at time  $B$  lineage 1 and 2 coalesce. Finally, at time  $A$  the lineage (1, 2) coalesce with the lineage (3, 4).

The mutational richness of the population  $M$  (or total diversity) is the number of unique mutations found in the population. Its expected value is proportional to the length of the coalescent tree.

The mutation frequency spectrum  $(a_k)_{k \in \mathbb{N}}$  is another measure of diversity that counts how many mutations are represented by  $k$  individuals in the population.

All these measures requires some knowledge of the shape of the coalescent tree of the population. The next paragraph is dedicated to establishing this for the simple case of serial transfer, while the last paragraph is dedicated to the case of splitting droplets.

### 3.2.1 Diversity Within Droplets in Serial Transfer

Establishing the law of the Coalescent Point Process of a lineage within serial transfer requires identification of the law of the branch length. This law is well-known for simple branching process such as the Linear Birth-Death process with parameters  $(b, d)$  modelling the population dynamics (Lambert and Stadler (2013), Proposition 5).

The addition of repeated bottleneck with period  $T$  is also possible within the theory (Lambert and Stadler (2013), Proposition 7) by thinning the original process (Figure 3.9). Each bottleneck at time  $iT, i = 1, \dots, n$  may remove independently each branch of the CPP with probability  $(1 - \delta)$  (Grey in Figure 3.9). Removing a branch in the past (at time  $iT$ ) may result in removing several branches in the present, and requires an adjustment to branch length (green in Figure 3.9). The number of branches removed, and the adjustment

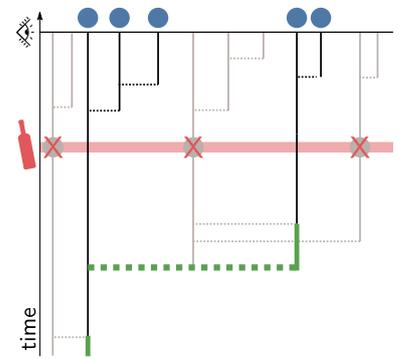


Figure 3.9: **Bottlenecks are modelled by thinning the process.** A bottleneck at some point in the past (red bottle) resulted in the extinction of some lineages (grey). As a consequence some branch length are changed (thick green lines).

to the branch length distribution can be computed from the law of the branch length  $H$ , the sampling probability  $\delta$  and the period of the bottleneck  $T$ . As a result:

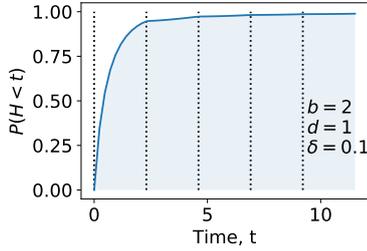


Figure 3.10: **Cumulative probability for the branch length.** Dotted lines represents bottlenecks.

**Proposition 6 (Coalescent tree of a lineage):**

Let  $\mathcal{T}_n$  be the random coalescent tree spawned by a single particle with extant descent just before the  $n$ -th dilution cycle.  $\mathcal{T}_n$  is a Coalescent Point Process (CPP) stopped in  $nT$ .

For super-critical particles ( $r > 0$ ) and critical dilution ( $\delta^* = e^{-rT}$ ). The inverse tail distribution of  $\mathcal{T}_n$  is:

$$F(kT + s) = 1 + \frac{b}{r} (e^{rs} - 1 + k(e^{rT} - 1)) \quad (3.7)$$

With  $k \in \mathbb{N}$  and  $s < T$ .

(Proof page 55.)

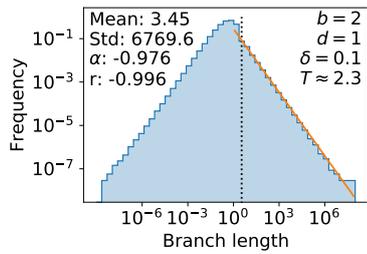


Figure 3.11: **Branch length distribution.** Sample of  $10^8$  realisation of the random variable by the inversion of the cumulative probability function method. Long branches follow a power law distribution with parameter  $\alpha = -1$  (orange line).

Proposition 6 gives the cumulative probability function for the branch length  $\mathbb{P}(H < t) = 1 - \frac{1}{F(t)}$  (Figure 3.10). Note that this function is defined by parts for each cycle.

$F$  has no finite limit for increasing times. Thus, the cumulative distribution function tends toward 1. As a consequence, there is no branch can have an infinite length (which happen with probability  $1 - \lim_{t \rightarrow \infty} H(t)$ ).

The random variable  $H$  can be easily sampled from its cumulative probability function. As illustrated in Figure 3.11, the variance of the branch length can be large. In this sample, the mean branch length is slightly longer than the cycle duration  $T$ , but the maximal branch length is in the order of  $10^6$  cycles. The distribution of long branches (longer than 1) can be fitted by a power law with parameter  $-1$ .

**Number of mutations**

In order to find the expected number of mutations within a droplet, the expected size of the full coalescent tree must be considered. This relies on the branch length distribution, conditioned to be lower than the duration of the experiment. It results in the following:

**Proposition 7 (Number of mutations):**

Suppose that a population of  $D$  droplets is seeded by independently sampling a proportion  $\delta$  of  $K$  cells. Let  $M_n$  be the number of mutation mutations we expect to observe in the population of droplets with carrying capacity  $K$  after  $n$  cycles of duration  $T$ , with optimal dilution

$\delta = \delta^* = e^{-rT}$ , birth rate of cells  $b$ , death rate  $d$ , ( $r = b - d$ ) and mutation rate  $\theta$ .

$$M_n = D\theta K\delta s_n L_n \quad (3.8)$$

With  $L_n$  the average length of the coalescent tree at cycle  $n$  of a surviving lineage started by one cell at cycle 0:

$$L_n = F(nT) \int_0^{nT} \frac{1}{F(t)} dt \quad (3.9)$$

With  $F$  the scale function of the associated CPP.

More precisely, when using the expression of  $F$  from Proposition 6:

$$L_n = \left(1 + \frac{b}{r}n(e^{rT} - 1)\right) \sum_{k=0}^n \frac{rT - \log\left(\frac{k(e^{rT}-1) + \frac{r}{b}}{(k+1)(e^{rT}-1) + \frac{r}{b}}\right)}{bk(e^{rT}-1) - d} \quad (3.10)$$

(Proof page 57.)

Proposition 7 shows first and foremost that the total expected neutral diversity in a serial transfer protocol is proportional to the number of droplets  $D$ , the carrying capacity of the droplets  $K$  and the mutation rate  $\theta$ . In other words, doubling the number of droplets, the mutation rate, or doubling the droplet size double the expected number of mutations in the whole population.

The expected number of mutations  $M_n$  is also proportional to the expected coalescent tree length of a single extant lineage  $L_n$  (weighted by the proportion of lineages that actually survive  $s_n$ ). This expected coalescent length is parametrised by the birth and death rates of the particles, but also the duration of the growth phase  $T$ .

Figure 3.12 shows that the expected number of mutations increases indefinitely with the number of cycles. However, the rate of increase is tied to the dilution bottleneck and tends to slow down when the number of cycles increase. Note that for a small number of cycles, the expected number of mutations increases with a higher dilution: one cycle with a dilution by two yields less diversity than one cycle with a dilution by one hundred. However, if ten cycles are performed, a dilution by two yields more diversity. This illustrates a trade-off between a harsh bottleneck, that allows a great number of cell divisions (and potentially a lot of mutations) but leads to loss of most extant mutations (due to founder effects) and a softer bottleneck that allows for fewer mutations to accumulate each cycle, but compounds more because fewer mutations are lost.

Figure 3.13 clarifies the link between the expected number of mutations and the dilution bottleneck. Note that smaller bottleneck sizes are compensated by longer cycles because the experiment is supposed to be performed in optimal conditions ( $\delta = e^{-rT}$ ). If there is only one cycle ( $n = 1$ ), maximal expected number of mutations is reached when the dilution bottleneck is vanish-

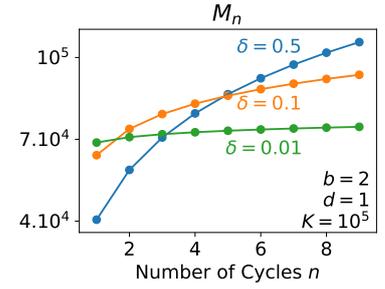


Figure 3.12: **Expected number of mutations through experimental cycles.** The expected number of mutations increases with the number of experimental cycles. For one cycle larger bottleneck give better results (The curve  $\delta = 0.5$  is higher than the curve for  $\delta = 0.001$ ) for more cycles however, larger bottlenecks yield more mutations.

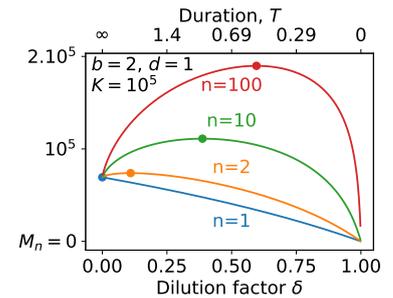


Figure 3.13: **Expected number of mutations for different bottleneck size.** For a single cycle, smaller bottleneck always yield more mutations. However, if more than one cycle is performed, there is a non null optimal bottleneck size that maximise the expected number of mutations found in the population.

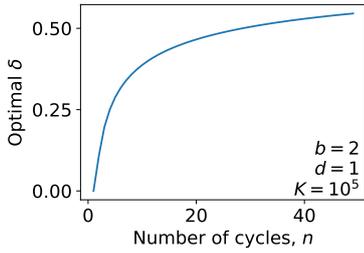


Figure 3.14: **Mutation-optimising bottleneck size as a function of the number of cycles.** The bottleneck size that optimises the expected number of mutations is increasing with the number of cycle performed.

ingly small ( $\delta \rightarrow 0$ ) and the cycle length adequately long ( $T \rightarrow \infty$ ). However, if there is more than one cycle ( $n > 1$ ) the expected number of mutations reaches a maximum value between  $\delta = 0$  and  $\delta = 1$ . This maximum-diversity dilution bottleneck value increases with the number of cycles (Figure 3.14). Thus, the dilution bottleneck should be adjusted to the expected duration of the experiment in terms of cycle number.

Overall, the expected number of neutral mutations accumulated by the population increases through time and can be optimised by appropriately choosing a bottleneck size that optimises the trade-off between accumulating new mutations and not losing old ones. However, the number of mutations is an information-poor descriptor of the diversity within the droplets. Indeed, some of those mutations could be born by a single individual, while others might be shared by the whole population. The next section address this problem by exploring the mutation frequency spectrum.

### Mutation Frequency spectrum

A more precise assessment of the neutral diversity structure involves differentiating between mutations that are carried by few individuals, and mutations that are widespread within the population. The mutation frequency spectrum presents the proportion of mutations that are carried by a given number of individuals. The expected mutation frequency spectrum of a coalescent point process can be deduced from the law of branch lengths  $H$  (Lambert (2008), Theorem 2.2). Indeed, the number of mutations carried by  $i$  individuals is proportional to the length of the coalescent tree subtending  $i$  leaves (Figure 3.15). As a result:

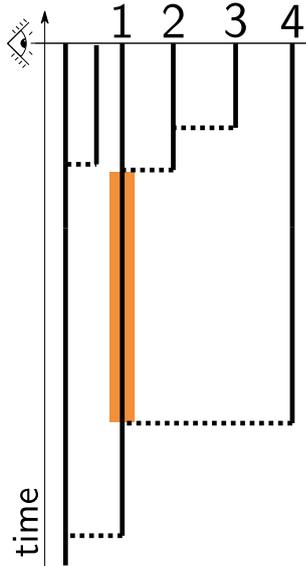


Figure 3.15: **Finding the mutation frequency spectrum.** Mutation shared by 3 individuals are the ones that arose within the orange region only. This region is delimited by  $\max(H_{i+1}, H_{i+2}) < t < H_{i+3}$ .

#### Proposition 8 (Mutation frequency spectrum):

Consider the Coalescent Point Process  $\mathcal{T}_n$ , with overlaying mutations following a Punctual Poisson Process with intensity  $\theta$ .

Let  $M_n^f$  be the expected number of mutation fixed in the population, that is mutations shared by all individuals. Then:

$$M_n^f = \theta \frac{1}{1 - F(nT)} \int_0^{nT} \frac{1}{1 - \frac{1}{F(s)} + \frac{1}{F(nT)}} ds \quad (3.11)$$

Let  $a_u$  be the expected frequency of mutations that are shared by  $u > 0$  individuals in the limit of large sample of the population.

$$a_u = \theta \int_0^{nT} \left( 1 - \frac{\frac{1}{F(x)} - \frac{1}{F(nT)}}{1 - \frac{1}{F(nT)}} \right)^{u-1} \left( \frac{\frac{1}{F(x)} - \frac{1}{F(nT)}}{1 - \frac{1}{F(nT)}} \right)^2 dx \quad (3.12)$$

(Proof page 59.)

Proposition 8 shows that the mutation frequency spectrum is proportional to the mutation rate  $\theta$  meaning that an increasing proportion of individuals

carry mutations if the rate increase, but that does not change the relative frequency of the size of groups carrying a given mutation.

Figure 3.16 shows mutation frequency spectra for one and a hundred cycles, and for three different dilution rate. Note that for a single cycle, harsher bottlenecks (i.e., smaller  $\delta$ , and correspondingly longer cycle duration  $T$ ) increase the tail of the distribution (there are more mutations that are shared by many individuals). This effect of  $\delta$  and  $T$  is not as simple when considering several cycles. For  $n = 100$ , the distribution is more heavy-tailed when the bottlenecks are soft ( $\delta = 0.5$ ) than when they are harsh ( $\delta = 0.001$ ). The information entropy of the mutation frequency spectrum can be used to systematically explore the effect of  $\delta$  on the shape of the distribution. Figure 3.17 shows that if more than one cycle are performed, there is a value of  $\delta$  that is expected to optimise the information entropy of the mutation frequency spectrum. This value is different from the value that optimises the number of mutations (Figure 3.13). Thus, there is a trade-off between accumulating many mutations, and having a diverse mutation frequency spectrum. The decision to fix  $\delta$  in order to optimise one or the other depend on the goal of the experiment.

To sum up, the neutral diversity found within parallel cultures in serial transfer depends mainly on the size of the coalescent tree of individuals. In general, higher dilution rates or longer duration of collective growth cycles result in longer trees and increased diversity, even though the population may risk going extinct. Extinction of the population marks the “death” of the culture, and it is the eventual fate of a serial transfer experiment in the limit of an infinity of cycles. However, in a nested population design, the cultures can also “reproduce” and replace the extinct ones. This has far-reaching consequences on the genealogy of the particles as shown in the next section.

### 3.2.2 Diversity in Recently Divided Droplets

Nested populations design differ from simple serial transfer in parallel cultures by the opportunity for the cultures (droplets, tubes or other compartments) to be subject to a birth-death process themselves. Each cycle, some cultures may be removed from the experiment, while others can be duplicated, usually by dispatching samples of the original collective in several new fresh medium compartments (rather than one in regular parallel serial transfer experiments).

This section focuses on the consequence of imposing a collective-level birth-death process on the neutral diversity. To this end, consider the simple scenario (depicted in Figure 3.18) of a pair of droplets that share a common “droplet ancestor” some cycles ago. The two droplets differs by the initial sampling performed in their common ancestor, and also by all new mutations accumulated since they became isolated.

In the following, the particles follow a super-critical linear birth-death process with parameters  $b - d = r > 0$ . The parameters of the population structure are supposed to be optimal in the sense of section 3.1.1: each cycle has a duration  $T^* = -r^{-1} \ln(cK^{-1})$  and each lineage as an independent probability of being sampled at a bottleneck of  $\delta^* = e^{-rT^*}$ .

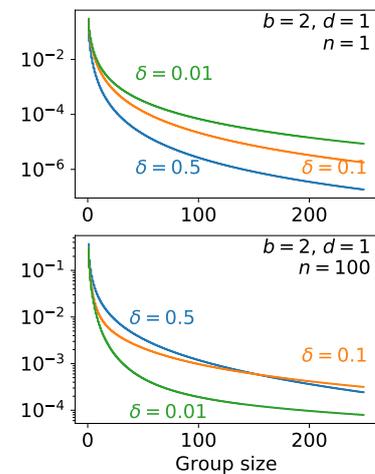


Figure 3.16: **Mutation frequency spectrum.** Give the frequency of mutations carried by a given number of individuals in a large sample of the population. *Top:* After  $n = 1$  cycle. *Bottom:* After  $n = 100$  cycles.

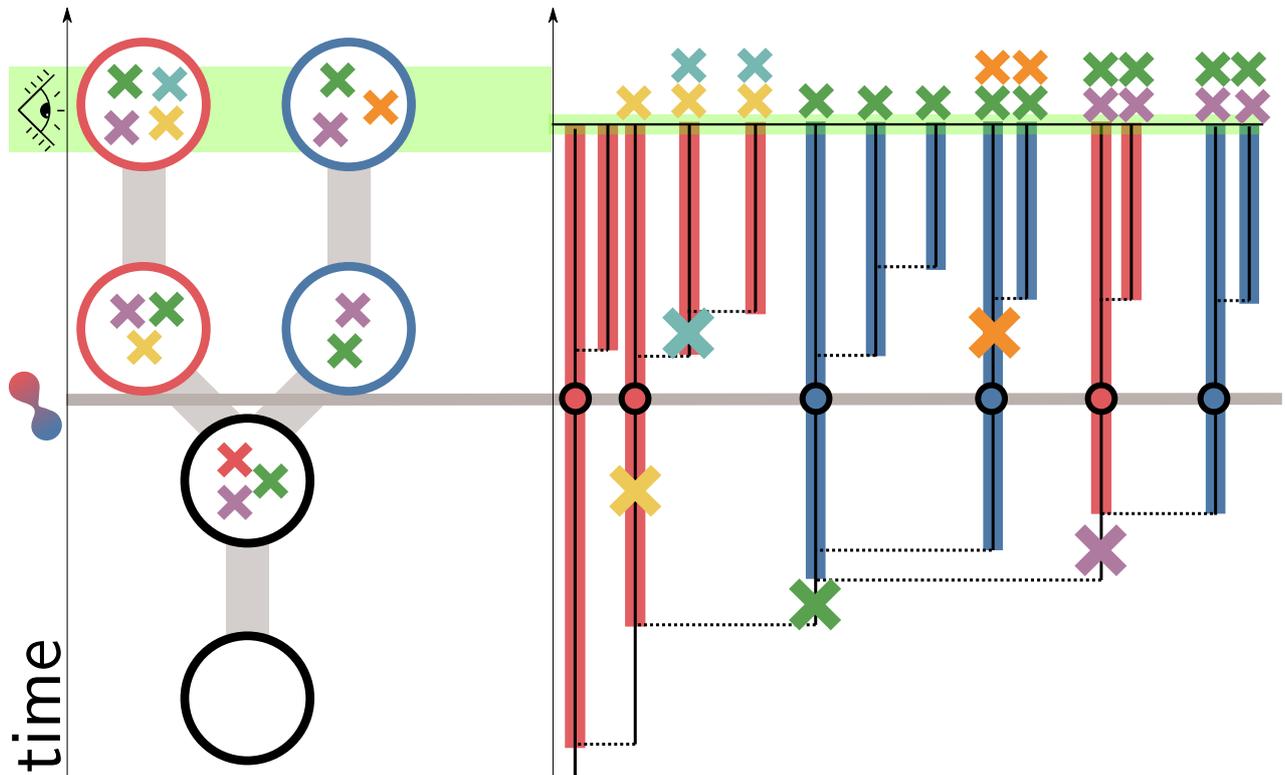


Figure 3.18: **Collective and Particle nested coalescent trees.** Left: coalescent tree of droplets. Right: coalescent tree of particles. An ancestral droplet lineage (black) is diluted into two offspring droplets (red, blue). At the time when the droplets are split, the particle lineages within the ancestral droplets are assigned a color (red or blue) that indicates the droplet in which they were sampled. Mutations appear along the genealogy of particles (crosses). Some mutations appear before the split (green, yellow, orange) and are found in both droplets (green, purple) if they were sampled by both droplet lineages, or within a single droplet (yellow) if they were segregated by the dilution. Other mutations appear after the split (light blue, yellow) and are only found in one of the droplet lineages

### Survival probability

First, let us consider the probability that a lineage spawned by a single cell  $n-m$  cycles in the past is not extinct within both droplets. The key to establish this probability is to recognise that the lineage undergoes a bottleneck with survival probability  $\delta$  at each cycle, except the cycle of the droplet split where each particle has a probability  $2\delta$  to survive. Indeed, two inoculation volumes are concurrently sampled from the ancestral droplet and dispatched into two offspring (Figure 3.19). Thus:

#### Proposition 9 (Survival probability - Split droplet):

*Cells within droplets in serial transfers are modelled by a linear birth-death process with constant parameters  $b$  and  $d$ , that is subject to periodic bottlenecks every duration  $T$ . Additionally, consider that at the  $m$ -th cycle, the dilution procedure is repeated to obtain  $k$  new droplets.*

*Let  $\bar{s}_{k,n,m}$  be the probability that lineage spawned by a single cell is not extinct at time  $nT$  just before the  $n$ -th dilution.*

$$s_{k,n,m} = 1 - h_{Q_{k,n,m}}(0),$$

Where  $h_A$  the linear fractional function with coefficient  $A$  defined in Equation 3.1. The matrix  $Q_{k,n,m}$  is the product:

$$Q_{k,n,m} = Q_1^{m-1} Q_k Q_1^{n-m-1} R$$

with,

$$\forall k \in \mathbb{N}^*, Q_k = \begin{bmatrix} k\delta(p-q) & p-k\delta(p-q) \\ k\delta(p-1) & p-k\delta(p-1) \end{bmatrix}; \quad R = \begin{bmatrix} \delta(p-q) & q \\ \delta(p-1) & 1 \end{bmatrix}$$

Where  $\delta$  is the survival probability of a particle at serial transfer,  $q$  is the extinction probability of a lineage during a cycle, and  $p$  the geometric parameter of the size of a non-extinct lineage as defined in Proposition 1.

(Proof page 60.)

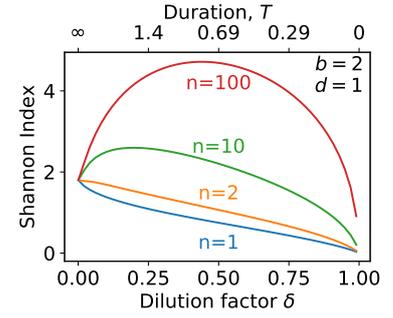


Figure 3.17: **Shannon Entropy of the mutation frequency spectrum.** Defined as  $S = -\sum_i a_i \log(a_i)$

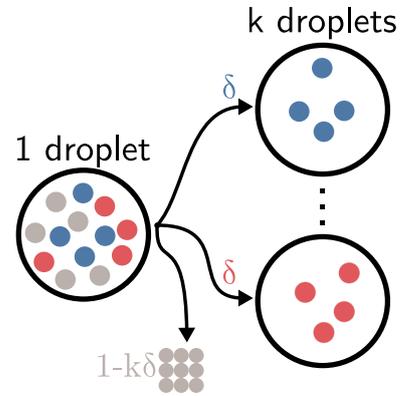


Figure 3.19: **Droplet Splitting Process.** When a droplet is split in  $k$  droplets, individual particles are independently selected to be transferred to the next cycle (with probability  $\delta$  for each new droplet) or sent to the waste (with probability  $1 - k\delta$ ).

Proposition 9 is similar in its conclusion Proposition 1, that treated of the serial transfer case. However, the expression is considerably less easy to handle, as the iteration does not simplify in a single matrix power.

### Total diversity

As seen in Proposition 7, quantifying the total neutral diversity in an infinitely-many sites model is a matter of finding the total length of the coalescent tree (or forest) of the population. Note that the full coalescent tree in Figure 3.20 can be decomposed into a stump, before the splitting of droplets, and

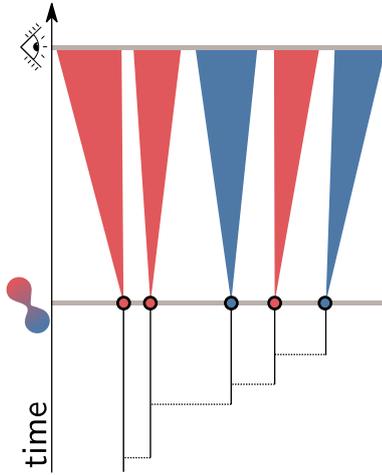


Figure 3.20: **Finding the total diversity at cycle  $n$  of droplets split at cycle  $m$ .** All the lineages (triangles) spawned from the particles dispatched in one of the  $k = 2$  droplets (here red and blue) have the same expected length  $L_{n-m}$ . The bottom part (or stump) of the tree (black) result from the sampling of a CPP stopped in  $mT$ , with probability  $\pi_{k,m,n}$ , the probability that an extant lineage at cycle  $m$  will be sampled in one of the two droplets, and survive until cycle  $n$ .

another set of CPP per droplet sampled in one or the other droplet lineage. As a result:

**Proposition 10 (Total Diversity - Split droplet):**

Consider an initial population of one droplet, seeded by sampling  $K$  cells with a probability  $\delta$ . Let a  $M_{k,m,n}$  be the expected number of unique mutations accumulated at cycle  $n$  into  $k = 1, 2 \dots \lfloor \frac{1}{\delta} \rfloor$  droplets that result from the splitting of the initial droplet at cycle  $m$ . Then:

$$M_{k,m,n} = K\delta\theta s_{k,m,n}L_{k,m,n}$$

With  $L_{k,m,n}$  the expected length of the coalescent tree of the population (conditional to survival):

$$L_{k,m,n} = F_{\pi_{k,m,n}}(mT) \left[ \int_0^{mT} \frac{1}{F_{\pi_{k,m,n}}(s)} ds + \int_0^{(n-m)T} \frac{F((n-m)T)}{F(s)} ds \right]$$

$F$  is the inverse tail distribution of the CPP with bottleneck,  $\pi_{k,m,n}(s) = k\delta s_{n-m}$  is the probability that a lineage at cycle  $m$  will have descent at cycle  $n$ .  $F_{\pi} = 1 - \pi + \pi F$  is the inverse tail distribution of a CPP with inverse tail distribution  $F$  submitted to sampling with probability  $\pi$ .

(Proof page 60.)

**Private Diversity**

In order to assess the divergence between split droplet, one can compute the expected number of *private* mutations, i.e., mutations that are only found in a single of the  $k$  split droplets. This number is the sum of all mutations that occur in the droplet after the splitting time  $mT$ , plus all the mutations that occur before the splitting time but in a lineage that only segregate in a single droplet. Since all the droplet are interchangeable, this value is identical for the  $k$  droplets. Overall this number is proportional to the red (or blue) part of the CPP in figure 3.18.

**Proposition 11 (Private Mutations - Split Droplet):**

Let a single droplet be split into  $k = 1 \dots \lfloor \frac{1}{\delta} \rfloor$  at cycle  $m$  out of  $n$ . Let  $\bar{M}_{k,m,n}$  be the expected number of mutations that are private to one of the  $k$  droplets.

$$K\delta\theta s_{k,m,n} \left[ \psi_{k,m,n} \bar{L}_{k,m,n} + \frac{1}{k} F_{\pi_{k,m,n}}(mT) L_{n-m} \right] \quad (3.13)$$

With  $\bar{L}_{k,m,n}$  the expected length of the stump tree (in black in Figure 3.20) and  $\psi$  the probability that a mutation occurring before cycle  $n$  is only carried by cells that are found in a single droplet,  $\pi_{k,m,n}$  the probability that a lineage extant before the dilution  $mT$  survives until time  $nT$ .

The stump tree is a Coalescent Point Process stopped in  $mT$ , with inverse tail distribution  $F$ , sampled with probability  $\pi_{k,m,n}$  at time  $mT$ . Thus, it has an expected number of leaves  $F_{\pi_{k,m,n}}(mT)$  and an expected length:

$$\bar{L}_{k,m,n} = F_{\pi_{k,m,n}}(mT) \int_0^{mT} \frac{1}{F_{\pi_{k,m,n}}(t)} dt \quad (3.14)$$

The expression of  $\psi$  is:

$$\psi = \sum_{i>1} \sum_{j=1}^{i-1} \frac{1}{F_{\pi}(mT)} \left(1 - \frac{1}{F_{\pi}(mT)}\right)^i \frac{1}{k^j} \theta \int_0^{\infty} dx \left(1 - \frac{1}{F_{\pi}(x)}\right)^{j-1} \left(\frac{i-j-1}{F_{\pi}(x)^2} + \frac{2}{F_{\pi}(x)}\right)$$

With  $\pi = \pi_{k,m,n}$ .

(Proof page 62.)

Droplet splitting main purpose is to select and duplicate a phenotype of interest. The last section explores, in the context of artificial selection, the advantage offered by a droplet-splitting process over the simple screening of parallel cultures in serial transfer.

### 3.3 Artificial selection of droplets

A practical application of a device that would allow the manipulation of small cultures of microbial organisms would be the artificial selection of phenotypes of interest. Suppose that a given phenotype of interest is reached after the accumulation of  $K$  mutations, and that it is possible to detect the number of mutation fixed so far, by sequencing or direct observation of the cultures.

To formalise, let  $D \in \mathbb{N}^*$  be the number of collectives. Each collective  $i$  is assigned a number  $e_i = 1, 2, \dots, K$ , corresponding to the number of fixed mutation. Suppose that the time for a collective to switch from  $e_i = j$  to  $e_i = j + 1$  is exponentially distributed with parameter  $p_j = \frac{\rho_j}{ND}$ . Where  $\rho_j$  is the rate of invasion of mutation  $j + 1$  (that could be deduced from Proposition 8), scaled by the number of droplets  $D$  and the number of cycles  $N$ . Initial conditions are that the  $D$  collectives are in state 0. The only possible transition is to accumulate a new mutation, no reversion is possible, as illustrated in Figure 3.21.

In order to assess the advantage of droplet splitting, consider two scenario, illustrated in Figure 3.22:

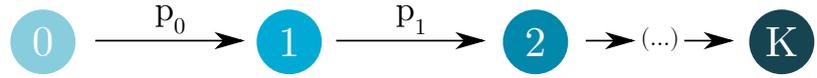


Figure 3.21: **Phenotypes.** There are  $K + 1$  possible phenotypes. A  $j$ -collective switch to the next phenotype  $j + 1$  with rate  $p_j$ .



Figure 3.22: **Propagation of mutations.** Without collective selection, all the lineage accumulate mutations independently. With collective selection, once a mutation fixation is first detected, the droplet is split into  $D$  lineages.

1. **Without collective selection**  $D$  collective lineages are started in state 0 at  $t = 0$  and undergo serial transfer independently of each other.
2. **With collective selection**  $D$  collective lineages are started in state 0 at  $t = 0$ , once a mutant is detected in a lineage, all the other collectives are killed and this lineage is split in  $D$  new lineages.

Let  $\Gamma$  (respectively  $\Gamma^*$ ) be the random variable encoding the first time for a lineage to get to the state  $K \in \mathbb{N}$  in the scenario without collective selection (respectively with collective selection). To compare them, consider their respective cumulative distribution functions:

**Proposition 12 (Cumulative distribution functions):**

The cumulative distribution function of  $\Gamma^*$  is:

$$\mathbb{P}(\Gamma^* \leq x) = 1 - e^{-x} \left( \sum_{u=0}^{K-1} \frac{x^u}{u!} \right) \quad (3.15)$$

The cumulative distribution function of  $\Gamma$  is:

$$\mathbb{P}(\Gamma \leq x) = 1 - e^{-x} \left( \sum_{u=0}^{K-1} \frac{x^u}{D^u u!} \right)^D \quad (3.16)$$

When the number of mutational steps tends to infinity, the two cumulative distribution function are equivalent. However, for any finite number of mutational steps  $K$ , the selective regime is faster than the serial transfer regime.

(Proof page 63.)

Proposition 12 shows that collective level selection, i.e., the process of splitting a droplet in which an intermediate mutation was fixed, lead to reduce the time to reach the  $K$ -th mutation. Figure 3.23 shows the shape of the cumulative probability function for both regime, illustrating this advantage. This constitutes a simple use-case for a device that allow the automated high-throughput manipulation of numerous cultures, such as the digital millifluidic analyser currently under development (Boitard et al., 2015).

### 3.4 Discussion

This chapter has laid the foundation for a theoretical understanding of the evolution of neutral diversity in massively parallel microbial evolution experiments. It was heavily inspired by ongoing engineering efforts to bring experimental evolution to digital millifluidics (Cottinet, 2013; Boitard et al., 2015; Dupin, 2018; Doulier, 2019).

In experimental microbiology, one desirable feature can be to maximise the number of mutations accumulated within the cultures, for instance in order to screen phenotypes of interest. The result presented above showed that, in an optimal growth setting, where cells are growing with a constant birth and death rate, without density dependant competition, the population should be submitted to cycles whose duration is tailored to compensate the bottleneck imposed at each serial transfer. The choice of the bottleneck should be made according to the expected duration of the experiment: in order to optimise the expected number of mutations, small bottlenecks (killing most of the lineages) should be used when the number of cycles is small, while larger bottlenecks (lower dilution rate) should be used for long term experiments. Additionally, the expected number of mutations increases linearly with increasing droplet volume and with increasing number of droplets which is a matter of technological progress as automation and larger droplet sizes are under consideration (Dupin, 2018). The mutation rate also increases linearly the number of expected mutations and can be manipulated by choosing mutator lineages or adding mutating chemicals to the culture broth. However, the potentially deleterious effects of this method might prevent using it in practical cases.

In long term evolution experiment, serial transfer is imposed by the need to replenish nutrients available to the cells. It is possible to build devices ensuring that a continuous flow of nutrient washes over the culture (for large volumes, see chemostats, or morbidostats (Toprak et al., 2013), in microfluidics, see mother machines (Potvin-Trottier et al., 2018)). However, these methods are usually more prone to contamination. In contrast, periodically diluting the culture in fresh medium is simple and robust.

The nested population design differs from traditional serial transfer of parallel cultures because it allows a collective birth-death process at the level of collective. Serial transfer is pervasive in experimental evolution (Kawecki et al., 2012), and has received extensive theoretical treatment. The first part of this Chapter constitute an example of such a study that only focus on neutral diversity. The problem of beneficial mutation, and the probability of loosing them because of the repeated bottleneck is well-known for serial transfer (Wahl and Krakauer, 2000; Wahl et al., 2002; Wahl and Gerrish,

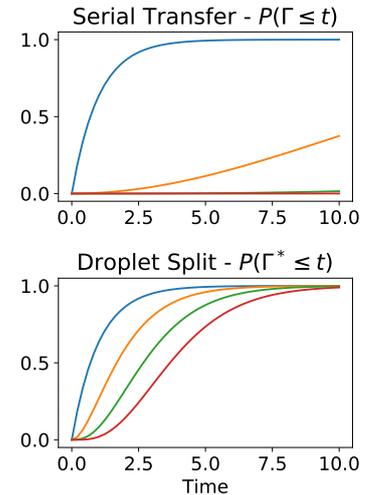


Figure 3.23: **Cumulative probability distribution** of the time to accumulate  $K$  mutations. With droplet splitting, accumulation of mutation is faster.  $D = 100$ .

2001; Wahl and Zhu, 2015) and should be extended to the nested population design in the future.

In practice, the collective birth-death process can come from the fact that some cultures are effectively empty because of high dilutions in the previous cycle, and may be replaced in the next cycle by cells from a non-empty culture. The collective birth-death process can also come from as a consequence of the culture technology. Milli and microfluidics compartments are usually produced in large numbers, while measurements are performed on all compartments, the retrieval of all the compartments content might not be practically possible or even desirable when they are too numerous. Finally, the collective birth-death process may stem from an active effort of the operator to select some populations based on some characteristics.

The use of a non-saturating population dynamics in this chapter is a simplification that should be carefully taken into account when transposing the results of this work to the design of experiments. Nonetheless, if the cycle duration are short enough so that the population is stopped during exponential phase, the heuristics developed in this chapter should hold. There are however two phenomena that were not modelled here and that will probably muddy the neutral pattern that was described. First, the absence of mutations affecting birth and death rates. If most point-mutations can be safely considered neutral, rare mutations affect the ability to reproduce of the cells. If the mutation are beneficial, they will increase in proportion within the population, and will change the relative frequency of all neutral mutations, by favouring the ones bore by the same strand of DNA. This is a well documented phenomenon known as hitchhiking (Fay and Wu, 2000). Second, horizontal gene transfer might allow the uncoupling of the mutation transmission from the genealogy (Dutta and Pan, 2002), muddying the pattern even more.

The nested population design also differs from trait groups (Wilson, 1975a) or transient compartments (Blokhuis et al., 2018) population structure because migration between compartments is prevented. As a consequence, it is possible to construct a non ambiguous genealogy of the cultures —hereafter refereed to as the collective genealogy and studied in more details in Chapter 4. In practice, serial transfer design offers a natural way to implement the birth-death process, by diluting some cultures into several new compartments (the droplet splitting) and discarding other.

Finally, this chapter touches briefly the problem of artificial selection using a nested population design. This was done by considering the accumulation of neutral mutation, in a subversion of classical cancer-evolution models (Nowak, 2014), indeed, here the objective is to accumulate mutations as fast as possible whereas those models are interested in population structures that decrease the rate of mutation accumulation. A more complete model of artificial selection would, however, take into account interaction between individuals, and potentially the selection of whole communities. Chapter 6 present an example of such a model, in the literature, community-level selection has been the subject of both experimental (Swenson et al., 2000a,b; Panke-Buisse et al., 2015) and theoretical inquiries (Arias-Sánchez et al., 2019; Xie et al., 2019).

Overall, the results presented in this chapter should be considered as a way to build intuition about the experimental system, while providing a null-model for diversity that could be compared to the actual patterns. Inevitable

---

differences will appear, but the point of comparison that is offered by neutral evolution will allow a better description of the observed diversity. Focusing on the part of the patterns that differ from this naive theoretical prediction will surely be fruitful: it shows that other mechanism than drift must be invoked. The next chapter continues the exploration of collective-level selection, by focusing on experimentally established collective-level genealogies.

## Appendix: Proofs

In this section we give the mathematical proofs for the results used in the main text.

### Parameters range

#### Proof (Proposition 1 - Survival Probability):

Let  $s_n$  be the survival probability at  $t = nT$  of a lineage started from a single cell at  $t = 0$ .

##### One cell, one cycle

Consider one cell at time  $t = 0$ . This cell follows a Linear Markov Branching Process  $(Z_t)_{t \in [0, T]}$  with constant rates  $b, d$  until the dilution time  $T$ . The branching process goes extinct ( $Z_t = 0$ ) with probability  $q(b, d, t) := \mathbb{P}(Z_t = 0)$ , and conditionally to non-extinction  $Z_t$  follow a geometric distribution with parameter  $p(b, d, t)$  ( $\mathbb{P}(Z_t = k | Z_t \neq 0) = p(1 - p)^{k-1}$ ).

		$p(b, d, t)$	$q(b, d, t)$
Subcritical particles	$r < 0$	$\frac{d-b}{d-be^{rt}}$	$\frac{d(1-e^{-rt})}{d-be^{rt}}$
Critical particles	$r = 0$	$\frac{1}{1+bt}$	$\frac{bt}{1+bt}$
Supercritical particles	$r > 0$	$\frac{(b-d)e^{-rt}}{b-de^{-rt}}$	$\frac{d(1-e^{-rt})}{b-de^{-rt}}$

The probability generating function of  $Z_T$  is:

$$\begin{aligned}
 f_{Z_T}(s) &= \sum_{k \geq 0} \mathbb{P}(Z_T = k) s^k && \text{Definition of } f \\
 &= \mathbb{P}(Z_T = 0) s^0 + \sum_{k \geq 1} \mathbb{P}(Z_T = k) s^k \\
 &= \mathbb{P}(Z_T = 0) s^0 + \sum_{k \geq 1} \mathbb{P}(Z_T \neq 0) P(Z_T = k | Z_T \neq 0) s^k \\
 &= q + \sum_{k \geq 1} (1 - q) p (1 - p)^{k-1} s^k && \text{Definition of } p, q \\
 &= q + \frac{p(1 - q)}{1 - p} \sum_{k \geq 1} ((1 - p)s)^k \\
 &= q + \frac{p(1 - q)}{1 - p} \left( -1 + \sum_{k \geq 0} ((1 - p)s)^k \right) \\
 &= q + \frac{p(1 - q)}{1 - p} \left( -1 + \frac{1}{1 - (1 - p)s} \right) && \text{Geometric series, } (1 - p)s < 1 \\
 &= q + \frac{p(1 - q)}{1 - p} \frac{s(1 - p)}{s(p - 1) + 1} \\
 &= q + \frac{sp(1 - q)}{s(p - 1) + 1} \\
 &= \frac{s(p - q) + q}{s(p - 1) + 1}
 \end{aligned}$$

Notice that, as expected,  $f_{Z_T}(1) = 1$  and  $f_{Z_T}(0) = P(Z_T = 0) = q$ .

#### Several cycles, including the last bottleneck

At the dilution time  $t = T$ , each descendant of this cell is sampled in the new droplet with probability  $\delta$ .

Let  $(B_i)_{i \geq 0} \sim B$  be a collection of independent Bernoulli random variables with parameter  $\delta$ . The probability generating function of  $B$  is  $f_B(s) = 1 - \delta + \delta s$ .

Let  $U_T$  the number of descent of a cell starting at time  $t = 0$  after one dilution at time  $t = T$ :

$$U_T = \sum_{i=1}^{Z_T} B_i \quad (3.17)$$

The number of descendant of one cell just after the  $n$ -th dilution follow a discrete time branching process  $(Y_n)_{n \in \mathbb{N}}$  with reproduction generating function  $f_{U_T}$ .

$$\begin{aligned} f_{U_T}(s) &= f_{Z_T} \circ f_B && (B_i)_{i \geq 0} \text{ iid.} \\ &= \frac{(1 - \delta + \delta s)(p - q) + q}{(1 - \delta + \delta s)(p - 1) + 1} = \frac{s\delta(p - q) - \delta(p - q) + p}{s\delta(p - 1) - \delta(p - 1) + p} \end{aligned}$$

Thus,

$$f_{U_T}(s) = \frac{(s - 1)\delta(p - q) + p}{(s - 1)\delta(p - 1) + p} \quad (3.18)$$

The survival probability of a lineage after  $n$  dilutions, is noted  $s_n$  and is given by the survival probability of the discrete branching process  $(Y_i)_{i=1 \dots n}$ . Thus:

$$s_n = 1 - \mathbb{P}(Y_n = 0) = 1 - f_{U_T}^n(0) \text{ where } f^k(s) = f^{k-1}(f(s)) = f(f^{k-1}(s)).$$

Note that  $f_{U_T}$  is linear fractional.

We can associate any linear-fractional function  $h_M$  with a coefficient matrix  $M$  as follow:

$$M = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \Leftrightarrow h_M = \frac{as + b}{cs + d}$$

It can be proven by simple induction that  $h_M^n = h_{M^n}$ .

We thus have  $1 - f_{U_T}^n = 1 - h_{Q^n}$  with:

$$Q = \begin{bmatrix} \delta(p - q) & p - \delta(p - q) \\ \delta(p - 1) & p - \delta(p - 1) \end{bmatrix}$$

Thus,

$$s_n = 1 - f_{U_T}^n(0) = 1 - h_{Q^n}(0)$$

### Several cycles, not including the last bottleneck

Let  $X_n$  be the number of descendant of one cell just before the  $n$ -th cycle. This number is given by the sum of the descent  $D_i$  of each of the  $i = 1 \dots (Y_{n-1})$  cells that descent from the ancestor at the beginning of the  $n$ -th cycle. By the branching property the  $(D_i)_{i=1 \dots Y_n}$  are independent of one another and follow the same law as  $U_1$ .

$$X_n = \sum_{i=1}^{Y_n} D_i \quad (3.19)$$

The survival probability  $\bar{s}_n = \mathbb{P}(X_n = 0)$  is:

$$\bar{s}_n = 1 - h_{Q^{n-1}R}(0)$$

With:

$$R = \begin{bmatrix} \delta(p-q) & q \\ \delta(p-1) & 1 \end{bmatrix}$$


---

**Proof (Proposition 2 - Long Term Survival Probability):**

Let us look for the fixed points of  $f_{X_\tau}$ :

$$\begin{aligned} f_{X_\tau}(s) = s &\Leftrightarrow (s-1)\delta(p-q) + p - s((s-1)\delta(p-1) + p) = 0 \\ &\Leftrightarrow (s-1)[\delta(p-q) - s\delta(p-1)] - (s-1)p = 0 \\ &\Leftrightarrow \delta(p-q) - s\delta(p-1) - p = 0 && \text{or } s = 1 \\ &\Leftrightarrow s\delta(p-1) = -p + \delta(p-q) && \text{or } s = 1 \\ &\Leftrightarrow s = \frac{p - \delta(p-q)}{\delta(1-p)} := s^* && \text{or } s = 1 \end{aligned}$$

$f_{X_\tau}(s) - s$  is a second order polynomial for  $s$  meaning that  $f_{X_\tau}$  has at most two fixed points. One of them is 1, in accordance with the definition of characteristic functions. The other is  $s^*$ .

$$s^* < 1 \Leftrightarrow p < \delta(1-q)$$

Since  $f_{X_\tau}$  is continuous and  $f_{X_\tau}(0) \in [0, 1]$ , the sequence  $(u_n)_{n \geq 1} = f_{X_\tau}(u_{n-1})$  with  $u_0 = 0$  converges toward 1 if  $p > \delta(1-q)$  and  $s^*$  otherwise.

- In **Subcritical regime**,  $p < \delta(1-q) \Leftrightarrow e^{-r\tau} < \delta$ , is always false since  $e^{-r\tau} > 1$  and  $\delta \leq 1$ .
- In **Critical regime**,  $p < \delta(1-q) \Leftrightarrow \frac{1}{1+b\tau} < \delta \left(1 - \frac{b\tau}{1+b\tau}\right) = \delta \frac{1}{1+b\tau}$ , is always false since  $\delta \leq 1$ .
- In **Supercritical regime**,  $p < \delta(1-q) \Leftrightarrow e^{-r\tau} > \delta$ ,

Thus,  $s^* > 1$  except if  $b > d$  and  $\delta < e^{-r\tau}$ . In this case,

$$\begin{aligned} s_\infty = 1 - s^* &= 1 - \frac{p - \delta(p-q)}{\delta(1-p)} \\ &= \frac{(b-d)(\delta - e^{-(b-d)T})}{b\delta(1 - e^{-(b-d)T})} \end{aligned}$$


---

**Proof (Proposition 4 - Critical dilution):**

We define  $\delta^*$  as the dilution parameter preventing the number of cells to grow unbounded. More precisely, we want the expected number of cells to be constant between two dilutions.

Let  $X_\tau$  be the number of descent just after one dilution ( $t = \tau_+$ ) of a cell starting a  $t = 0$  as defined in Eq 3.17. The generating function of  $X_\tau$  is  $f_{X_\tau}$  as defined in Eq 3.18. Note that  $\mathbb{E}(X_\tau) \Leftrightarrow f'_{X_\tau}(1)$ .

When  $\delta = \delta^*$ , we expect the number of descent after one cycle to be constant, thus  $\mathbb{E}(X_\tau) = 1$ .

$$\begin{aligned} f'_{X_\tau}(1) &= \delta^* \frac{1-q}{p} = 1 \\ \delta^* &= \frac{p}{1-q} \end{aligned}$$

When the particles are supercritical,  $0 < d < b$ :

$$\begin{aligned}\delta^* &= \frac{(b-d)e^{-(b-d)\tau}}{b-de^{-(b-d)\tau}} \frac{b-de^{-(b-d)\tau}}{b-de^{-(b-d)\tau}-d(1-e^{-(b-d)\tau})} \\ &= e^{-(b-d)\tau}\end{aligned}$$


---

**Proof (Proposition 5 - Critical survival Probability):**

Let  $\delta = e^{-rT}$ ,  $r = b - d$ ,  $r > 0$ .

Then, the probability of extinction of the branching process is:

$$q = \frac{d(1-\delta)}{b-d\delta} \quad (3.20)$$

The parameter of the geometric distribution of the number of branches, conditionally to non extinction is:

$$q = \frac{r\delta}{b-d\delta} \quad (3.21)$$

Thus:

$$Q = \frac{\delta}{b-\delta d} \begin{bmatrix} \delta b - d & b(1-\delta) \\ -b(1-\delta) & b(1-\delta) + r \end{bmatrix} := \frac{\delta}{b-\delta d} R$$

By induction, we can show that

$$s_n^* = \frac{Q_{01}^n}{Q_{11}^n} = \frac{R_{01}^n}{R_{11}^n} = \frac{1}{\frac{b}{r}n(1-\delta) + 1}$$


---

## Coalescent Point Processes

**Proof (Proposition 6 - Coalescent tree):**

Consider the CPP of the population just before the  $n$ -th cycle of dilution. The population experienced  $n - 1$  bottlenecks at times  $t_1 = (n - 1)T, \dots, t_{n-2} = 2T, t_{n-1} = T$ .

Let  $\{H_{ij}, (i, j) \in \mathbb{N}^2\}$  be a set of i.i.d. random variables following same law as  $H$ , defined by its inverse tail distribution  $F(t) = \frac{1}{\mathbb{P}(H > t)}$ .  $F$  is the scale function of the CPP with bottlenecks.

Let  $\tilde{F}$  be the scale function of the CPP without bottleneck. Using Proposition 7 in Lambert and Stadler (2013) with  $\epsilon_i = \delta$  and  $s_i = iT$ ,  $t = nT + s$ , with  $n = \lfloor \frac{t}{T} \rfloor \in \mathbb{N}$  and  $s < T$ , it is possible to write  $F$  as:

$$F(t) = \delta^n \tilde{F}(nT + s) + (1 - \delta) \sum_{j=0}^{n-1} \delta^j \tilde{F}(jT)$$

Just after the sampling, the scale function is  $F_s$ :

$$F_s(t) = 1 - \delta + \delta F(t)$$

The expression of  $\tilde{F}$  is known for the most common birth-death processes:

	Parameters	Scale function
Pure Birth	$0 = d < b$	$\tilde{F}(t) = e^{bt}$
Non critical	$0 < d \neq b > 0$	$\tilde{F}(t) = \frac{b}{r}(e^{rt} - 1) + 1$
Critical	$0 < d = b$	$\tilde{F}(t) = bt + 1$

**Non critical case**  $r := b - d \neq 0, b > 0, d > 0$ :

$$\begin{aligned}
F(nT + s) &= \delta^n \left[ 1 - \frac{b}{r} + e^{r(nT+s)} \frac{b}{r} \right] + (1 - \delta) \sum_{j=1}^n \delta^{j-1} \left[ e^{rjT} \frac{b}{r} + 1 - \frac{b}{r} \right] \\
&= \delta^n \left( 1 - \frac{b}{r} \right) + \frac{b}{r} e^{rs} (\delta e^{rT})^n + (1 - \delta) \left[ \left( 1 - \frac{b}{r} \right) \sum_{j=0}^{n-1} \delta^j + \frac{b}{r} e^{rT} \sum_{j=0}^{n-1} (\delta e^{rT})^j \right] \\
&= \delta^n \left( 1 - \frac{b}{r} \right) + \frac{b}{r} e^{rs} (\delta e^{rT})^n + \left( 1 - \frac{b}{r} \right) (1 - \delta^n) + \frac{b(1 - \delta)}{r} e^{rT} \sum_{j=0}^{n-1} (\delta e^{rT})^j \quad (\text{Geometric series, } \delta \neq 1) \\
&= 1 + \frac{b}{r} \left[ e^{rs} (\delta e^{rT})^n - 1 + (1 - \delta) e^{rT} \sum_{j=0}^{n-1} (\delta e^{rT})^j \right]
\end{aligned}$$

If  $\delta \neq \delta^*$  then,

$$F(nT + s) = 1 + \frac{b}{r} \left[ e^{rs} (\delta e^{rT})^n - 1 + (1 - \delta) e^{rT} \frac{1 - (\delta e^{rT})^n}{1 - \delta e^{rT}} \right] \quad (\text{Geometric series, } \delta^* e^{rT} \neq 1)$$

Otherwise, if  $\delta = \delta^* = e^{-rT}$ ,

$$F(nT + s) = 1 + \frac{b}{r} (e^{rs} - 1 + n(e^{rT} - 1))$$

**Pure-birth case**  $d = 0$

$$\begin{aligned}
F(nT + s) &= \delta^n e^{b(nT+s)} + (1 - \delta) \sum_{j=1}^n \delta^{j-1} e^{jbT} \\
&= e^{bs} (\delta e^{bT})^n + (1 - \delta) \sum_{j=0}^{n-1} \delta^j e^{(j+1)bT} \\
&= e^{bs} (\delta e^{bT})^n + (1 - \delta) e^{bT} \sum_{j=0}^{n-1} (\delta e^{bT})^j
\end{aligned}$$

If  $\delta \neq \delta^* = e^{-bT}$ , then since  $\delta e^{bT} \neq 1$ , we can apply the geometric series expression:

$$F(nT + s) = e^{bs} (\delta e^{bT})^n + (1 - \delta) e^{bT} \frac{1 - (\delta e^{bT})^n}{1 - \delta e^{bT}}$$

Otherwise,  $\delta e^{bT} = 1$ :

$$F(nT + s) = e^{bs} + n(e^{bT} - 1)$$

**Critical case  $b = d \neq 0$**

$$\begin{aligned}
F(nT + s) &= \delta^n (b(nT + s) + 1) + (1 - \delta) \sum_{j=1}^n \delta^{j-1} (bjT + 1) \\
&= \delta^n (b(nT + s) + 1) + (1 - \delta) \sum_{j=0}^{n-1} \delta^j (b(j+1)T + 1) \\
&= \delta^n (b(nT + s) + 1) + (1 - \delta) \left[ (1 + bT) \sum_{j=0}^{n-1} \delta^j + bT \sum_{j=0}^{n-1} j\delta^j \right] \\
&= \delta^n (b(nT + s) + 1) + (1 - \delta^n)(1 + bT) + (1 - \delta)bT \sum_{j=0}^{n-1} j\delta^j \quad (\text{Geometric series, } \delta \neq 1) \\
&= \delta^n b(T(n-1) + s) + 1 + bT + (1 - \delta)bT \sum_{j=0}^{n-1} j\delta^j \\
&= 1 + \delta^n b(T(n-1) + s) + bT \left[ 1 + (1 - \delta) \sum_{j=0}^{n-1} j\delta^j \right]
\end{aligned}$$


---

**Proof (Proposition 7 - Length of the coalescent tree):**

Let  $M_n$  be the expected number of mutations within the serial transfer experiment after  $n$  cycles but before the  $n$ -th dilution. With  $K$  the carrying capacity of the droplets, and  $T$  the duration of the growth phase. Suppose that  $\delta = \delta^*$ , and that each initial droplet is seeded from diluting a stock of  $K$  cells by  $\frac{1}{\text{delta}}$ .

The number of initial lineages within the droplet  $C$  is a Binomial random variable with parameters  $(\delta, K)$ . The expected number of is  $\mathbb{E}(C) = K\delta = Ke^{-rT}$ .

Each initial lineage survives to cycle  $n$  with probability  $\bar{s}_n$  (see Proposition 1). Since all lineages are independent the expected number of extant lineages after  $n$  cycles is  $K\delta s_n$ .

Each extant lineage after  $n$  cycle spawn an independent coalescent tree with expected length  $L_n$ . Each tree contains an expected Poisson number of mutations, with parameter proportional to the length of the tree. Since the length of the tree and the mutation rate are independent, the number of mutation accumulated on one tree is  $\theta L_n$ .

Suppose that there are  $D$  droplets, since all droplets are independents, the expected number of mutations within the experiment is simply:

$$M_n = Ke^{-rT} \bar{s}_n L_n \quad (3.22)$$

Let us now compute  $L_n$ .

Let  $\{H_{ij}, (i, j) \in \mathbb{N}^2\}$  be a set of i.i.d. random variables following same law as  $H$ , defined by its inverse tail distribution  $F(t) = \frac{1}{\mathbb{P}(H > t)}$ .

Let  $\tau_n$  be a CPP with branches  $H$  stopped in  $nT$ . Let  $N(\tau)$  be the number of leaves of the random tree  $\tau_n$ . The length of the tree  $\mathcal{L}(\tau_n)$  is the random variable:

$$\mathcal{L}(\tau_n) = nT + \sum_{j=1}^{N(\tau_n)-1} H_j$$

Where  $nT$  is the length of the spine and  $H_j$  are the length of the other  $N(\tau_n) - 1$  branches.

Its expected value is:

$$L_n = \mathbb{E}L(\tau_n) = nT + (\mathbb{E}(N(\tau_n)) - 1)\mathbb{E}(H|H < nT)$$

**Number of leaves:**  $N(\tau)$  is a geometric random variable:

$$\mathbb{P}(N(\tau) = k) = P(H > nT)P(H \leq nT)^{k-1} = \frac{1}{F(nT)} \left(1 - \frac{1}{F(nT)}\right)^{k-1}$$

Indeed, if  $k$  is the index of the first  $H_i$  such that  $H_i > nT$ , there are  $k - 1$  branches, plus the spine, for a total of  $k$  leaves.

Thus, the expected number of leaves of  $\tau_n$ , is:

$$\mathbb{E}(N(\tau_n)) = F(nT) \tag{3.23}$$

Note that if we do not take the spine into account:

$$\mathbb{E}(N(\tau_n - 1)) = F(nT) - 1 = \frac{1 - \frac{1}{F(nT)}}{\frac{1}{F(nT)}} = \frac{\mathbb{P}(H < t)}{\mathbb{P}(H > t)} \tag{3.24}$$

**Length of branches:**

We recall that for a positive r.v.  $X$ ,  $\mathbb{E}(X) = \int_0^{+\infty} \mathbb{P}(X > x)dx$ .

We can rescale the tail-distribution to take into account the conditional:

$$\mathbb{P}(H > x|H < y) = \begin{cases} 0 & \text{if } x > y, \\ \frac{\mathbb{P}(H > x) - \mathbb{P}(H > y)}{\mathbb{P}(H > 0) - \mathbb{P}(H > y)} & \text{otherwise.} \end{cases} \tag{3.25}$$

Thus:

$$\begin{aligned} \mathbb{E}(H|H < nT) &= \int_0^{+\infty} \mathbb{P}(H > x|H < nT)dx \\ &= \int_0^{nT} \frac{\mathbb{P}(H > x) - \mathbb{P}(H > nT)}{1 - \mathbb{P}(H > nT)} dx \\ &= \frac{1}{1 - \frac{1}{F(nT)}} \left[ \int_0^{nT} \frac{1}{F(x)} dx - \frac{nT}{F(nT)} \right] \quad \left( \text{By definition, } \mathbb{P}(H > x) := \frac{1}{F(x)} \right) \\ &= \frac{F(nT)}{F(nT) - 1} \left[ \sum_{k=0}^n \int_{kT}^{(k+1)T} \frac{1}{F(x)} - \frac{nT}{F(nT)} \right] \end{aligned}$$

Moreover,  $\forall k \in \mathbb{N}$  and  $\forall s \in \mathbb{R}, s < T$ :

$$F(kT + s) = 1 + \frac{b}{r}(k(e^{Tr} - 1) + e^{rs} - 1)$$

$$\begin{aligned} \int_k^{k+1} \frac{1}{F(t)} dt &= \int_0^T \frac{1}{1 + \frac{b}{r}(k(e^{Tr} - 1) + e^{rs} - 1)} ds \\ &= \left[ \frac{rs - \log(k(e^{rT} - 1) + e^{rs} - 1 + \frac{r}{b})}{bn(e^{rT} - 1) - b + r} \right]_0^T \\ &= \frac{rT - \log\left(\frac{k(e^{rT} - 1) + \frac{r}{b}}{(k+1)(e^{rT} - 1) + \frac{r}{b}}\right)}{bk(e^{rT} - 1) - d} \end{aligned}$$

Finally,

$$\begin{aligned}
L_n &= \mathbb{E}L(\tau_n) = nT + \mathbb{E}(N(\tau) - 1)\mathbb{E}(H|H < nT) \\
&= nT + \frac{\mathbb{P}(H < nT)}{\mathbb{P}(H > nT)}\mathbb{E}(H|H < nT) \\
&= nT + (F(nT) - 1)\frac{F(nT)}{F(nT) - 1} \left[ \int_0^{nT} \frac{1}{F(t)} dt - \frac{nT}{F(nT)} \right] \\
&= F(nT) \int_0^{nT} \frac{1}{F(t)} dt
\end{aligned} \tag{3.26}$$

Replacing the expression of  $F$ :

$$L_n = \left(1 + \frac{b}{r}n(e^{rT} - 1)\right) \sum_{k=0}^n \frac{rT - \log\left(\frac{k(e^{rT}-1) + \frac{r}{b}}{(k+1)(e^{rT}-1) + \frac{r}{b}}\right)}{bk(e^{rT} - 1) - d} \tag{3.27}$$

### Proof (Proposition 8 - Mutation Frequency Spectrum):

#### Fixed mutations

Fixed mutations are the one found between the root of the coalescent tree ( $t = nT$ ) and the first branching. Let  $Y = nT - X$  be the random variables encoding this length.

Let  $H_1 \dots H_{\tau_n - 1}$  be the branch lengths of the CPP stopped in  $nT$ . The variables  $H_i$  are independent and identically distributed as  $H$ , conditioned to be smaller than  $nT$ .  $X$  is the maximum of those values:

$$X = \max\{H_1, H_2 \dots H_N(\tau_n - 1)\} \tag{3.28}$$

$$\mathbb{P}(X > s) = \begin{cases} 0 & (\text{If } s > nT) \\ \sum_{k=0}^{\infty} \mathbb{P}(N(\tau) = k)\mathbb{P}(H > s|H < nT)^k & (\text{otherwise}) \end{cases}$$

$$\begin{aligned}
\sum_{k=0}^{\infty} \mathbb{P}(N(\tau) = k)\mathbb{P}(H > s|H < nT)^k &= \sum_{k=0}^{\infty} \frac{1}{F(nT)} \left(1 - \frac{1}{F(nT)}\right)^{k-1} \left(\frac{\frac{1}{F(s)} - \frac{1}{F(nT)}}{1 - \frac{1}{F(nT)}}\right)^k \\
&= \frac{1}{F(nT)} \left(1 - \frac{1}{F(nT)}\right)^{-1} \sum_{k=0}^{\infty} \left(\frac{1}{F(s)} - \frac{1}{nT}\right)^k \\
&= \frac{1}{1 - F(nT)} \frac{1}{1 - \frac{1}{F(s)} + \frac{1}{F(nT)}}
\end{aligned}$$

$$\mathbb{E}(X) = \int_0^{\infty} \mathbb{P}(X > x) dx \tag{3.29}$$

#### Full spectrum

For a sample of  $v$  individuals in a CPP stopped at time  $nT$ , Theorem 2.2 in Lambert (2008) gives the expression of  $\mathbb{E}(S_v(u))$  the expected number of mutant sites that are carried by exactly  $1 \leq u \leq v - 1$ .

$$\mathbb{E}(S_v(u)) = \theta \sum_{j=0}^{\infty} \left(1 - \frac{1}{W(x)}\right)^{u-1} \left(\frac{v-u-1}{W(x)^2} + \frac{2}{W(x)}\right) dx \tag{3.30}$$

With  $W$  the inverse tail distribution of the branch length in the stopped CPP.

The expected value of  $S_v(u)$  converges toward a limit (Lambert, 2008) when the sample size increases, provided  $\mathbb{E}(H|H < nT) < \infty$ , which is the case here, since the CPP is stopped in  $nT$ .

$$\lim_{v \rightarrow \infty} v^{-1} \mathbb{E}(S_v(u)) = \theta \int_0^\infty \left(1 - \frac{1}{W(x)}\right)^{u-1} \frac{1}{W(x)^2} dx \quad (3.31)$$

$W$  is expressed, for  $x < nT$ , as a rescaling of  $F$ :

$$W(x) = \frac{1}{\mathbb{P}(H > x | x < nT)} = \frac{1 - \frac{1}{F(nT)}}{\frac{1}{F(x)} - \frac{1}{F(nT)}} \quad (3.32)$$

Thus,

$$a_u = \lim_{v \rightarrow \infty} v^{-1} \mathbb{E}(S_v(u)) = \theta \int_0^{nT} \left(1 - \frac{\frac{1}{F(x)} - \frac{1}{F(nT)}}{1 - \frac{1}{F(nT)}}\right)^{u-1} \left(\frac{\frac{1}{F(x)} - \frac{1}{F(nT)}}{1 - \frac{1}{F(nT)}}\right)^2 dx \quad (3.33)$$

## Split droplets

### Proof (Proposition 9 - Survival Probability (2 drops)):

The proof is the same as the one for Proposition 1 (p. 1) but the survival probability at the end of cycle  $m$  is  $k\delta$  instead of  $\delta$ .

### Proof (Proposition 10 - Total Diversity (2 drops)):

Let  $t = nT$  the duration of the experiment. Let  $(Z_s)_{s \in [0, t]}$  be the continuous time branching process encoding the population size. The initial population size is  $Z_0 = 1$ . The population is submitted to a sampling with probability  $\delta$  every period  $T$ , i.e. at times  $(t_1 = T, \dots, t_{n-1} = (n-1)T)$

#### Number of branches within the tree, without droplet splitting

Let  $X_m$  with  $1 \leq m \leq n$ , be the random variable encoding the number of extant lineages just before the  $m$ -th bottleneck of a CPP stopped in  $t_m = mT$ . Conditioned on non extinction,  $X_m$  is geometric with parameter  $\mathbb{P}(H > t_m)$ .

Moreover, let  $Y_{m,n}$  with  $1 \leq m < n$  be the random variable encoding the number of extant lineages just after the  $m$ -th bottleneck in a CPP stopped at  $t = nT$ . Conditioned on  $Y_m$ ,  $Y_{m,n}$  is binomial with parameter  $\pi_m = \delta \mathbb{P}(Z_{t_n-m} \neq 0)$ , the probability that a lineage extant at time  $t_m$  will not be extinct at time  $t$ . To survive, a lineage have to be sampled (with probability  $\delta$ ) and survive up to time  $t$ .

Thus, for all  $k > 0$  the law of  $X_{m,n}$  can be expressed as:

$$\begin{aligned} \mathbb{P}(Y_{m,n} = k) &= \mathbb{P}(Z_{t_m} \neq 0) \mathbb{P}(Y_{m,n} = k | Z_{t_m} \neq 0) && \text{(Conditional Probability)} \\ &= \mathbb{P}(Z_{t_m} \neq 0) \sum_{j \geq 1} \mathbb{P}(X_m = j | Z_{t_m} \neq 0) \mathbb{P}(Y_{m,n} = k | X_m = j) && \text{(Total Probability)} \\ &= \mathbb{P}(Z_{t_m} \neq 0) \sum_{j \geq 1} \mathbb{P}(H > t_m) \mathbb{P}(H < t_m)^{j-1} C_j^k (\pi_m)^k (1 - \pi_m)^{j-k} && (3.34) \end{aligned}$$

The expected value is:

$$\mathbb{E}(Y_{n,m}) = \mathbb{P}(Z_{t_m} \neq 0) \frac{\pi_m}{\mathbb{P}(H > t_m)}$$

There are two ways that leads to have no branches alive at time  $t_0$  and time  $t$ . First if there are no branches at time  $t_0$ , second if all the lineages extant at time  $t_0$  went extinct before  $t$ :

$$\begin{aligned} \mathbb{P}(Y_{m,n} = 0) &= \mathbb{P}(Z_{t_m} = 0) + \mathbb{P}(Z_{t_m} \neq 0)\mathbb{P}(Y_{n,m} = 0|Z_{t_m} \neq 0) && \text{(Total Probability)} \\ &= \mathbb{P}(Z_{t_m} = 0) + \mathbb{P}(Z_{t_m} \neq 0) \sum_{j \geq 1} \mathbb{P}(H > t_m) \mathbb{P}(H < t_m)^{j-1} (1 - \pi_m)^j && \text{(From 3.34, k=0)} \\ &= \mathbb{P}(Z_{t_m} = 0) + \mathbb{P}(Z_{t_m} \neq 0) (1 - \pi_m) \mathbb{P}(H > t_m) \sum_{j \geq 1} (\mathbb{P}(H < t_m) (1 - \pi_m))^{j-1} \\ &= \mathbb{P}(Z_{t_m} = 0) + \mathbb{P}(Z_{t_m} \neq 0) \frac{\mathbb{P}(H > t_0) (1 - \pi_m)}{1 - \mathbb{P}(H < t_m) (1 - \pi_m)} && \text{(Geometric Series)} \end{aligned}$$

Thus, the probability of having at least one branch alive at time  $t_m$  that has still descent at time  $t$  is given by:

$$\begin{aligned} \mathbb{P}(Y_{m,n} \neq 0) &= 1 - P(Y_{n,m} = 0) \\ &= \mathbb{P}(Z_{t_m} \neq 0) \left( 1 - \frac{\mathbb{P}(H > t_m) (1 - \pi_m)}{1 - \mathbb{P}(H < t_m) (1 - \pi_m)} \right) \\ &= \frac{\mathbb{P}(Z_{t_m} \neq 0) \pi_m}{1 - \mathbb{P}(H < t_0) (1 - \pi_m)} \end{aligned}$$

Note that the probability of having at least one extant lineage at cycle  $m$  that is still extant at cycle  $n$  is the same as the probability of having at least one extant lineage at cycle  $n$ :  $\mathbb{P}(Y_{m,n} \neq 0) = \mathbb{P}(Y_n \neq 0)$ . Thus, this probability is independent of  $m$ .

Let us partition the expected number of branches by the event  $Y_{m,n} = 0$ :

$$\mathbb{E}(Y_{m,n}) = \underbrace{\mathbb{P}(Y_{m,n} = 0) \mathbb{E}(Y_{m,n} | Y_{m,n} = 0)}_{=0} + \mathbb{P}(Y_{m,n} \neq 0) \mathbb{E}(Y_{m,n} | Y_{m,n} \neq 0)$$

Finally, the expected number of branches at time  $t_m$  that still have a descent at time  $t$ , conditional to non extinction is:

$$\begin{aligned} \mathbb{E}(Y_{m,n} | Y_{m,n} \neq 0) &= \frac{\mathbb{E}(Y_{m,n})}{\mathbb{P}(Y_{m,n} \neq 0)} \\ &= \left[ \frac{\mathbb{P}(Z_{t_m} \neq 0) \pi_m}{\mathbb{P}(H > t_m)} \right] \left[ \frac{1 - \mathbb{P}(H < t_m) (1 - \pi_m)}{\mathbb{P}(Z_{t_m} \neq 0) \pi_m} \right] \\ &= \frac{1 - \mathbb{P}(H < t_m) (1 - \pi_m)}{P(H > t_m)} \\ &= 1 - \pi_m + \pi_m F(t_m) \quad \left( \text{Since } F(t) = \frac{1}{\mathbb{P}(H > t)} \right) \end{aligned}$$

Notice that it is the same expression as the expected number of branches of a CPP with inverse tail distribution  $F$ , where the leaves are independently sampled in the present with probability  $\pi_m$  (Lambert and Stadler (2013), Proposition 2).

### Number of branches within the tree, with droplet splitting

Consider now that the droplet was diluted into  $k$  droplets at time  $t_m$ . Let  $Y_{k,m,n}$  be the random variable encoding the number of extant lineages just after the split at time  $t_m$  that are still extant at time  $t_n$ . For  $k = 1$ ,  $Y_{k,m,n}$  follows the same law as  $Y_{m,n}$ . For all values of  $k = 1, 2 \dots \lfloor \frac{1}{\delta} \rfloor$ ,  $\pi_m$  must be substituted by:

$$\pi_{m,k} = k\delta\mathbb{P}(Z_{t_n-t_m} \neq 0) \quad (3.35)$$

### Tree length

Let  $\mathcal{L}_{k,m,n}$  be the random variable encoding the length of the coalescent tree of the population of particles within  $k$  droplets, submitted to a bottleneck  $\delta$  every  $T$  unit of time, where at cycle  $m$ , the population was diluted into  $k$  droplets instead of one.

The length of this tree, conditioned on non extinction ( $Y_{k,n,m} \neq 0$ ) is the sum of the length of the "stub"  $\mathcal{S}_{k,m,n}$ , the tree before  $t_m$ , and the length of all the  $Y_{k,n,m} \neq 0$  trees spawned by the lineages that are extant after the dilution at cycle  $m$ .

First, the stub is a CPP with inverse tail distribution  $F$  submitted to a Bernoulli sampling  $\pi_{k,m}$  at time  $t_m$ . Thus, it is a CPP with inverse tail distribution  $F_{\pi_{k,m}} := 1 - \pi_m + \pi_m F$ . Its expected length is given by Equation 3.26:

$$\mathbb{E}(\mathcal{S}_{k,m,n}) = S_{k,m,n} = \int_0^{mT} \frac{F_{\pi_{k,m}}(mT)}{F_{\pi_{k,m}}(s)} ds \quad (3.36)$$

Additionally, each of the sampled leaves at time  $t_m$  give rise to a CPP of expected length  $L_{m-n}$  (Equation 3.26). Finally:

$$\mathbb{E}(\mathcal{L}|Y_{m,n} \neq 0) = \int_0^{mT} \frac{F_{\pi_{k,m}}(mT)}{F_{\pi_{k,m}}(s)} ds + F_{\pi_{k,m}}(mT) \int_0^{(n-m)T} \frac{F((n-m)T)}{F(s)} ds \quad (3.37)$$

### Proof (Proposition 11 - Private Mutations - Split Droplet):

All mutations that occurs after the droplet split are private to the droplet.

Mutations that occurs before the droplet split (in the "stump", i.e. the black part of Figure 3.20) can be ultimately found in a single or several droplets (as explained in the Figure 3.18). Let  $F_\pi$  be the inverse tail distribution of the stump CPP.

Let a single droplet be split into  $k = 1 \dots \lfloor \frac{1}{\delta} \rfloor$  at cycle  $m$  out of  $n$ . Let  $P$  be the event that a mutation occurring on the stump tree is private to a droplet, i.e. that this mutation is only carried by cells that are found in a single droplet.

Let  $C$  be the random variable that encodes the number of individual carrying the focal mutation. Conditional on the value of  $C = j$ , the probability of  $P$  is the probability that the  $j$  individuals are sampled in the same droplet, namely  $(\frac{1}{k})^j$ .

Let  $Y_{m,n}$  be the number of leaves of the stump. Conditional on non extinction, it follows a geometric distribution with parameter  $\frac{1}{F_\pi(mT)}$ .

Conditional on  $Y_{m,n}$ ,  $C$  follow a probability distribution given in Theorem 2.2 of Lambert (2008). Indeed, if  $S_i(j)$  is the random variable encoding the number of mutations carried by  $j$  individual in a sample of  $i$  individual, then:

$$\begin{aligned} P(C = j|Y_{m,n} = i) &= \sum_{u=1}^i \frac{\mathbb{P}(S_i(j) = u)u}{i} = \mathbb{E}(S_i(j)) \\ &= \theta \int_0^\infty dx \left(1 - \frac{1}{F_\pi(x)}\right)^{j-1} \left(\frac{i-j-1}{F_\pi(x)^2} + \frac{2}{F_\pi(x)}\right) \quad \text{Theorem 2.2 of Lambert (2008)} \end{aligned}$$

To sum up:

$$\begin{aligned}
\mathbb{P}(P) &= \sum_{j>1} \mathbb{P}(C = j) \mathbb{P}(P|C = j) && \text{(Total Probabilities)} \\
&= \sum_{i>1} \sum_{j=1}^{i-1} \mathbb{P}(Y_{m,n}|Y_{m,n} \neq 0) \mathbb{P}(C = j|Y_{m,n} = i) \mathbb{P}(P|C = j) && \text{(Total Probabilities)} \\
&= \sum_{i>1} \sum_{j=1}^{i-1} \frac{1}{F_\pi(mT)} \left(1 - \frac{1}{F_\pi(mT)}\right)^i \frac{1}{k^j} \theta \int_0^\infty dx \left(1 - \frac{1}{F_\pi(x)}\right)^{j-1} \left(\frac{i-j-1}{F_\pi(x)^2} + \frac{2}{F_\pi(x)}\right)
\end{aligned}$$


---

## Artificial Selection of Droplets

**Proof (Proposition 12 - Cumulative distribution functions):**  
**Time to accumulate  $K$  mutations**

Let  $(e_{p_{ij}}^{(i,j)})_{i,j \in \mathbb{N}^2}$  be independant exponential random variables with parameter  $p_{ij}$ . Without collective selection,  $\Gamma$  is the minimum of a set of sum of i.i.d. exponential variables:

$$\Gamma = \inf_{j=1 \dots D} \left( \sum_{i=0}^K e_{\frac{1}{D}}^{(i,j)} \right)_j$$

**With collective selection**,  $\Gamma^*$  is a sum of a minimum of i.i.d. exponential variables, (i.e. a sum of exponential variables):

$$\Gamma^* = \sum_{i=0}^K \left( \inf_{j=1 \dots D} e_{\frac{1}{D}}^{(i,j)} \right)_i = \sum_{i=0}^K e_{\sum_{j=1}^D \frac{1}{D}}^{(i)} = \sum_{i=0}^K e_1^{(i)}$$

cumulative distribution function of  $\Gamma^*$  is:

$$\mathbb{P}(\Gamma^* \leq x) = 1 - e^{-x} \left( \sum_{u=0}^{k-1} \frac{x^u}{u!} \right)$$

Note that it does not depend on  $D$  anymore: if the number of collectives is in the order of one over the scaled invasion rate, the time to accumulate  $K$  mutations with collective selection does not depend on the invasion rate anymore.

**Without collective selection**, the cdf of  $\Gamma$  is:

$$\begin{aligned}
\mathbb{P}(\Gamma \leq x) &= 1 - \mathbb{P}(\Gamma > x) \\
&= 1 - \mathbb{P}\left(\bigcap_{j=1}^D \left[ \left( \sum_{i=0}^K e^{\frac{i}{D}} \right)_j > x \right]\right) \\
&= 1 - \left[ \mathbb{P}\left( \left( \sum_{i=0}^K e^{\frac{i}{D}} \right)_j > x \right) \right]^D && \text{(The lineages are independants)} \\
&= 1 - \left[ 1 - \mathbb{P}\left( \left( \sum_{i=0}^K e^{\frac{i}{D}} \right)_j \leq x \right) \right]^D \\
&= 1 - \left[ 1 - 1 + e^{-\frac{1}{D}x} \left( \sum_{u=0}^{k-1} \frac{x^u}{D^u u!} \right) \right]^D && \text{(sum of indep. exp. r.v.)} \\
&= 1 - \left[ e^{-\frac{1}{D}x} \left( \sum_{u=0}^{k-1} \frac{x^u}{D^u u!} \right) \right]^D \\
&= 1 - e^{-x} \left( \sum_{u=0}^{k-1} \frac{x^u}{D^u u!} \right)^D
\end{aligned}$$

### Comparison of the two regimes

It boils down to comparing:

$$f_k(x) := \sum_{u=0}^{k-1} \frac{x^u}{u!} \quad \text{and} \quad h_k(x, D) := \left( \sum_{u=0}^{k-1} \frac{x^u}{D^u u!} \right)^D$$

**Infinite number of mutational steps**, if we let  $K \rightarrow \infty$ , the time to accumulate  $K$  mutation increase toward infinity (it is not possible to accumulate an infinity of mutation in a finite time).

$$\begin{aligned}
\lim_{K \rightarrow \infty} \mathbb{P}(\Gamma^* \leq x) &= 1 - e^{x-x} = 0 && \left( \text{because } \sum_u \frac{x^u}{u!} = e^x. \right) \\
\lim_{K \rightarrow \infty} \mathbb{P}(\Gamma \leq x) &= 1 - e^x \left( e^{\frac{x}{D}} \right)^D = 0 && \text{(idem).}
\end{aligned}$$

Moreover, the two cdf are equivalent:

$$\lim_{k \rightarrow \infty} \frac{h_k(x, D)}{f_k(x)} = \frac{e^x}{e^x} = 1$$

### Any number of mutations steps

Let us prove that  $h_k(x, D) \geq h_k(x, 1) = f_k(x)$  for all  $(x, D, k) \in \mathbb{R}_+ \times (\mathbb{N}^*)^2$ ,  $k \geq 2$ ,

*Case 1*,  $0 \leq x \leq (k-1)D$ :

$$f'_k(x) = \sum_{u=1}^{k-1} \frac{(u)x^{u-1}}{u!} = \sum_{j=0}^{k-2} \frac{x^j}{j!} = f_k(x) - \frac{x^{k-1}}{(k-1)!}$$

Let us define  $A(D) := \ln(Dh(x, k, n)) = D \ln\left(f_k\left(\frac{x}{D}\right)\right)$

$$\begin{aligned} A'(D) &= \ln\left(f_k\left(\frac{x}{D}\right)\right) + D \frac{\left(f_k\left(\frac{x}{D}\right)\right)'}{f_k\left(\frac{x}{D}\right)} \\ &= \ln\left(f_k\left(\frac{x}{D}\right)\right) - \frac{x}{D} \frac{f_k'\left(\frac{x}{D}\right)}{f_k\left(\frac{x}{D}\right)} \\ &:= B\left(\frac{x}{D}\right) \end{aligned}$$

$$\begin{aligned} B(x) &= \ln(f_k(x)) - x \frac{f_k'(x)}{f_k(x)} \\ &= \ln(f_k(x)) - x + \frac{x^k}{(k-1)!f_k(x)} \end{aligned}$$

Note that  $B(0) = 0$ .

$$\begin{aligned} B'(x) &= \frac{f_k'(x)}{f_k(x)} - 1 + \frac{kx^{k-1}(k-1)!f_k(x) - (k-1)!f_k'(x)x^k}{((k-1)!f_k(x))^2} \\ &= \frac{1}{f_k(x)} \left[ \left( f_k(x) - \frac{x^{k-1}}{(k-1)!} \right) - f_k(x) + k \frac{x^{k-1}}{(k-1)!} - \frac{x^{k-1}}{(k-1)!} \frac{x \left( f_k(x) - \frac{x^{k-1}}{(k-1)!} \right)}{f_k(x)} \right] \\ &= \frac{x^k}{(k-1)!f_k(x)} \left[ k - 1 - x + \frac{x^k}{(k-1)!f_k(x)} \right] \end{aligned}$$

Since  $\forall(x, k) \in \mathbb{R}_+ \times \mathbb{N}^*$ ,  $\frac{x^k}{(k-1)!f_k(x)} \geq 0$

$$\begin{aligned} x \leq (k-1)D &\Rightarrow B'\left(\frac{x}{D}\right) \geq 0 \text{ and moreover } B(0) = 0 \\ &\Rightarrow B\left(\frac{x}{D}\right) \geq 0 \\ &\Rightarrow A'(D) \geq 0 \\ &\Rightarrow h_k(x, D) \text{ is increasing in } D. \end{aligned}$$

*Case 2,  $x \geq (k-1)D$ :*

Let  $(D, k) \in (\mathbb{N}^*)^2$ ,  $k \geq 2$ ,

$$\begin{aligned} \frac{x}{D} &\geq (k-1) \geq 1 \\ \frac{x^u}{D^u u!} &\geq \frac{(k-1)^u}{u!} \quad (\forall u \in \mathbb{N}, \text{ since } x \mapsto x^u \text{ increasing } \forall x \geq 1). \\ \sum_{u=0}^{k-1} \frac{x^u}{D^u u!} &\geq \sum_{u=0}^{k-1} \frac{(k-1)^u}{u!} \\ \left( 1 + \sum_{u=1}^{k-1} \frac{x^u}{D^u u!} \right)^D &\geq \left( 1 + \sum_{u=1}^{k-1} \frac{(k-1)^u}{u!} \right)^D \quad (\text{Since } x \mapsto x^D \text{ increasing } \forall x \geq 1 \text{ and } D \in \mathbb{N}). \\ h_k(x, D) &\geq \left( 1 + \sum_{u=1}^{k-1} \frac{(k-1)^u}{u!} \right)^D \end{aligned}$$

Moreover,

$$\left(1 + \sum_{u=1}^{k-1} \frac{(k-1)^u}{u!}\right)^D \geq \left(1 + \sum_{u=1}^{k-1} \frac{(k-1)^u}{u!}\right) \quad (\text{Since } x \leq x^D \text{ increasing } \forall x \geq 1 \text{ and } D \in \mathbb{N}).$$

$$h_k(x, D) \geq h_k(x, 1) = f_k(x)$$


---

## Chapter 4

# Locating Mutations in Collective-level Genealogies

“Voyez-vous cet œuf ? C’est avec cela qu’on renverse toutes les écoles de théologie et tous les temples de la terre.”

—DENIS DIDEROT, *Entretien entre D’Alembert et Diderot* (1769)

A TRANSITION TO MULTICELLULARITY happens when ancestral free living cells become part of collectives that eventually evolve mechanisms for autonomous reproduction (Okasha, 2006). It implies the existence of a multicellular life-cycle with its own development and reproduction mechanisms that can span several life-cycles at the cell level. Multicellular states may be optional for cells, such as in aggregative slime molds (Bonner, 1998), or become obligate as in metazoans. Overall, multicellularity evolved at least twenty-five times in the history of life (Bonner, 1998; Grosberg and Strathmann, 2007). Such a spectacular example of convergent evolution is a potent argument in favour of the existence of one or several general mechanisms promoting it. For these reasons multicellular transitions are more accessible to experimental inquiries than, say, the origin of the eukaryotic cell. Experimental approaches to the evolution of multicellularity often focus on organisms that already display some multicellular life cycle (such as social amoeba *Dictyostelium*, Bonner (2015) or green algae *Volvox*, Kirk (1997)) or start from unicellular organisms (such as budding yeast *Saccharomyces cerevisiae*, Ratcliff et al. (2012) or bacteria such as *Pseudomonas fluorescens*, Hammerschmidt et al. (2014)) and impose selective conditions thought to favour the emergence of multicellular structures. At any rate, the problem is to understand how collectives of initially free living individuals become Darwinian entities on their own right (De Monte and Rainey, 2014), and possibly stay that way thanks to ratcheting mechanisms (Libby and Ratcliff, 2014) that can culminate with effective “de-Darwinization” of the individual level (Godfrey-Smith, 2009).

Population structure is a key parameter of ecological and evolutionary dynamics: the interplay between local and global populations have far-reaching consequences for diversity (MacArthur and Wilson, 1967), adaptation (Wade, 2016) and speciation (Wright, 1949, 1982). As such, population structure is

a likely candidate one may consider when looking for the mechanisms underpinning the emergence of multicellularity. Consider a meta population of individual cells in an environment where the habitat is structured in patches because of rare resources, the existence of physical support, or boundaries. Additionally, suppose that the cells have means to colonise new patches by dispersal. Under specific conditions, this could bestow marginal Darwinian properties upon collectives of unwitting cells. This would provide the opportunity for natural selection to effect the transition from marginally Darwinian collectives to fully integrated, paradigmatic Darwinian organisms. This process is called ecological scaffolding (Black et al., 2019).

By experimentally controlling population structure, and any transfer of matter from one deme to the other, it is possible to put a population of micro-organisms in close to ideal conditions for ecological scaffolding of multicellularity. In a very elegant experiment using *Pseudomonas fluorescens* Hammerschmidt et al. (2014) showed that collective-level selection could act on a developmental programme that underpinned a two-phase life cycle involving a collective (soma-like) mat-forming stage and a dispersing (germ line-like) dispersing phase. A continuation of this experiment, conducted by Philippe Remigi and Daniel Rixin in Massey University (New Zealand) followed more precisely the collective-level genealogy of the system. These experiments have produced a wealth of data including records of population dynamics and meta genome sequences. Additionally, since samples from all lineages and time points have been preserved in cold storage, it is possible to “replay the tape” of evolution and obtain replications of key events. Overall, synthesising the various data sets and finding relevant measures for quantifying the level of collective level adaptation is a challenge.

The aim of this chapter is to visualise and analyse collective-level genealogies established experimentally with the goal of quantifying collective-level adaptation through time. This approach involves identification of those places in the genealogy where collectives improve most with respect to the selective pressure. The combination of statistical modelling with sequence data helps shed new light on possible molecular bases of adaptation in this system, and will help to orient and motivate future experimental research.

## 4.1 The cheat as first propagule hypothesis

Of all the potential obstacles to the evolution of cooperative structures such as multicellularity, the emergence of cheating types is maybe the one that received the most attention (Michod and Roze, 2001; Strassmann and Queller, 2011). From parasitic types of polymerase (Takeuchi and Hogeweg, 2012) to cancers (Aktipis, 2016), the idea that a sub population of individuals would benefit greatly from going rogue, and would eventually bring any collective structure down is pervasive across evolutionary biology. This apparent paradox is studied by a wealth of literature of evolutionary game theory with ramifications that goes beyond the domain of biology into social sciences (Hardin, 1968; Axelrod and Hamilton, 1981) and addressed specifically in the case of structured populations (Ferriere and Michod, 1996; Hochberg et al., 2008; Garcia et al., 2015). In this chapter however, the focus is resolutely non-game-theoretic and more in line with multi-level selection of Darwinian

populations (Okasha, 2006; Godfrey-Smith, 2009). The thought experiment of the “pond-scum” is developed and gives a clear illustration of the concept of ecological scaffolding.

Aerobic bacteria, such as *Pseudomonas fluorescens*, are limited in their growth by the supply of oxygen. In this context, colonising the air-liquid interface brings a marked advantage to any individual because dynamics of convection and diffusion are more efficient in air than in water to counteract the local depletion of oxygen caused by respiration. Individual cells do not usually stay within the air-liquid interface. However, some mutants are able to produce an extra-cellular mesh of cellulose that allows formation of a microbial mat at the air-liquid interface (Rainey and Travisano, 1998; Spiers et al., 2002; Ardré et al., 2019). These mutants, known as wrinkly spreader (WS) types on account of their colony morphology on agar plates, have a longer doubling time compared to cells of the ancestral type. This reduced growth rate arises as a consequence of the costs associated with cellulose over-production. Nonetheless, despite costs to individual WS cells, the group reaps a benefit that individual cells cannot, namely, access to oxygen because of ability to colonise the air-liquid interface (Figure 4.1a).

The success of WS mutants is attributable to the cooperative interactions among cells. Theory predicts that cooperating groups are prone to invasion by cheating types (Hardin, 1968). Such a scenario unfolds in WS groups propagated for extended periods of time (Rainey and Rainey, 2003). Mutants that lose ability to over-produce cellulose arise within the mat. These mutants avoid the cost of producing cellulose, yet retain the advantages that come from access to oxygen via existence at the air-liquid interface. These mutants, known as smooth (SM) types, grow unchecked among the WS mats (Figure 4.1b), and as they contribute nothing toward mat strength, their presence weakens the mat leading to its premature collapse (Figure 4.1c). In a widespread anthropomorphic analogy SM mutants are colloquially called “cheating” or “defecting” cells, while WS are called “cooperating” cells (Rainey and Rainey, 2003).

Under certain ecological circumstances the eco-evolutionary dynamics arising from the frequency dependent interactions among WS and SM types can generate a life cycle that allows a Darwinian process to unfold at the level of mats. An example of these circumstances, colloquially named “pond-scum” scenario, is shown in Figure 4.2. Consider a pond where reeds or other physical support provide an anchoring for bacterial mats (Figure 4.2.1). In this context, the collapsing of a mat corresponds to the opening of a niche in the form of the now free support (Figure 4.2, 1 and 2). Extant collectives may shed cells in the environment that could colonise those empty niches. SM mutant cells are more likely to be washed out from their collective and thus more likely to disperse to new niches, provided that the WS mutation arise once again in their descent (Figure 4.2, 2 and 3). If the niches are sufficiently numerous to support the population (see Chapter 5), and distant enough to be initially colonised by one or a few cells, it is possible to identify parent-offspring relationship between collectives. Overall, this defines a genealogy of collectives (Figure 4.2, left) with its own birth events (the colonisation of a new support) and death events (the collapse of a mat). Ecological conditions that result in collective-level Darwinian properties being exogenously imposed

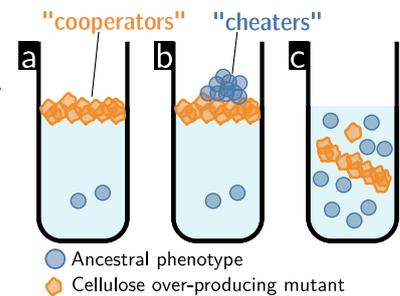


Figure 4.1: **Tragedy of the commons in a microcosm.** a. Mutant cells colonise the air-water interface by cooperatively producing a cellulose mat (“cooperators”). b. Ancestral phenotype cells can benefit from the oxygen supply without paying the cost of cellulose production (“cheaters”). c. The mat collapses if the cellulose production is too low compared to the number of cells (“tragedy”). This is described in Rainey and Rainey (2003)

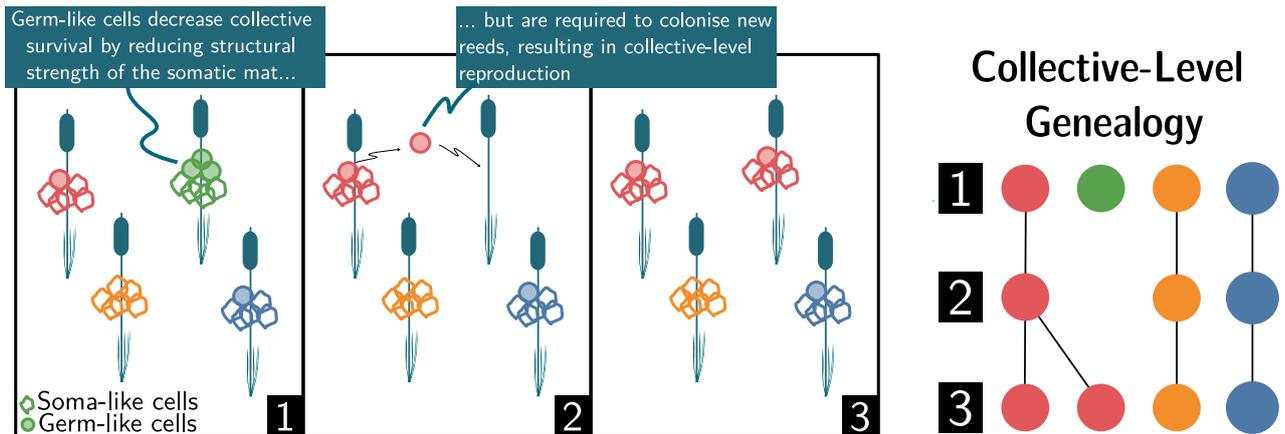


Figure 4.2: **The pond-scum scenario.** Bacterial cells colonise physical supports (reeds) at the air-liquid interface of a pond by producing a mat structure. The collapse of such structures can be identified as a collective-death event while the colonisation of a new support can be identified as a collective-birth. This establishes collective-level genealogies. Cells can be more adapted to participate in the formation of the mat (soma-like) or in the colonisation of new niches (germ-like)

on collectives such as the WS mats, has been termed “ecological scaffolding” (Black et al., 2019).

Thus, SM cells, while imposing on the mat an undeniable strain, also improve the mat’s reproductive output. This reframing of the phenomenon from an individual tragedy of the commons to a collective trade-off between survival and reproduction (Rainey and Kerr, 2010) shifts a major impediment to the evolution of multicellularity (the presence of cheats) to a solution of an even greater problem (the reproduction of collectives). It allows to explore the idea that SM and WS cells form different lineages of a proto-multicellular organism. WS cells are adapted to the collective function of survival and constitute a soma-like lineage, while SM cells are adapted to the collective function of reproduction, constituting a germ-like lineage. The transition from soma-like to germ-like cells becomes a requirement for the closing the proto-life cycle of the mat (Figure 4.3). Note that this life cycle requires two mutation events to be completed (to switch from germ-like phenotype to the soma-like phenotype, and back) contrary to cell differentiation in metazoan that relies mainly on epigenetics mechanisms.

## 4.2 Establishing collective level genealogies

The reality of ecological scaffolding of bacterial mats in nature has yet to be demonstrated. However, this system can be studied in ideal conditions in the laboratory. By substituting reeds with microcosms (i.e., tubes) and by controlling the establishment of new colonies. The artificial boundaries of this meta population structure discretise the collectives and impose a certain collective-level individuality. The following presents the general experimental setup used by Remigi and Rexin, and showcases a new contribution to the

analysis of this system: the visual display of replacement events as a collective-level genealogy.

### Experimental set-up

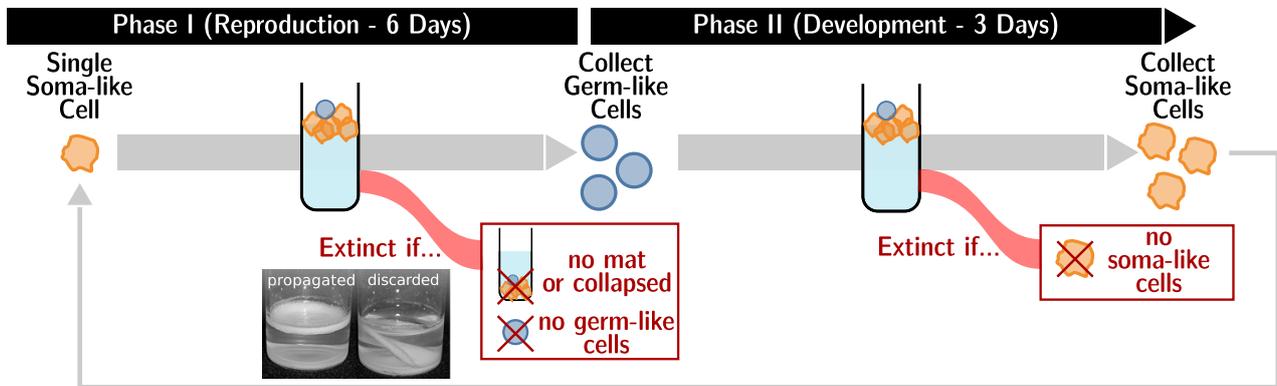


Figure 4.4: **Experimental collective-selection regime.** Each lineage of the population goes repetitively through two successive selection steps. After a period of 6 days they are tested for the presence of a mat sitting at the air-liquid interface and germ-like cells. Germ-like cells are collected and inoculate collectives in the next phase. After a period of 3 days they are tested for the presence of soma-like cells. Extinct collectives are removed from the population and replaced by new collective seeded with cells from surviving collectives (see main text for details).

Collectives undergo successive cycles comprised of two phases of selection (Figure 4.4). The first phase is initiated by a single soma-like cell and ends with the selection for soma-like mats that are both viable (i.e., non collapsed) and fecund (i.e., production of germ-like cells). The second phase is initiated by placing germ-like cells collected from the previous phase in a fresh microcosm and ends with the selection for the reversion to the soma (i.e., production of soma-like cells). On completion of a cycle, a single colony of the most dominant soma-like cell is transferred to a fresh microcosm. This constitutes the beginning of the first phase of the next cycle and ensures that the population goes through a single cell bottleneck.

Ten cycles are performed on 48 parallel cultures. As in the pond-scum thought experiment, collectives lineages that fail to complete the life cycle and go extinct provide an opportunity for viable collective lineages to reproduce by freeing niches. Upon the demise of a particular collective lineage, a viable lineage is chosen at random and allowed to replace the extinct type (Figure 4.5).

This experiment was repeated for two microcosm diameters, small (S) and large (L), and for two genotypes. One of the genotypes was recovered from the end of the [Hammerschmidt et al. \(2014\)](#) experiment (Line 17) and contained a mutation in the DNA mismatch repair system (*mutS*) that underpinned a genetic switch that allowed rapid transition between soma and germ phases, the second strain is isogenic to the first, with the *mutS* mutation reverted to wild type (WT). This gives four combinations (WT-S, WT-L, 17-S, 17-L). The exact biological motivations and conclusion of those different treatments are

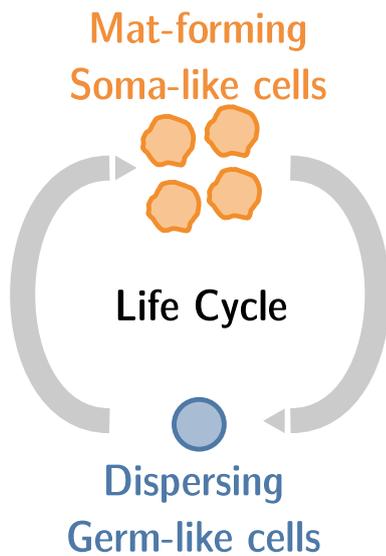


Figure 4.3: **A proto life-cycle.** Cell over-producing cellulose (also known as Wrinkly spreader colony phenotype, WS or cooperators in the literature) may be considered as the proto-soma of a life cycle which also involve non over-producing strain (Smooth colony phenotype, SM, or cheaters in the literature) as a proto germ-line. Several mutational pathways exists for the transition from one phenotype to the other.

not the focus of this chapter and will only be briefly discussed. The results will mainly serve as illustration of general statements on the visualisation and interpretation of the genealogies. This is an ongoing project and the complete results will be the subject of a future publication.

### Visualisation of collective-level genealogies

Each new cycle of the experiment produced a set of complex data: for each collective, the record of germ and soma-like cell census, of the presence of a mat, and additionally the record of the collective birth and death events. Finding a graphical representation that allow to synthesise these data while not discounting their complexity was required.

Collective-level genealogies are defined as a directed acyclic graph where vertices represent cultures and edges represent the flow of cells from one generation to the other. They constitute the natural representation of this data set that stem from the idea that collectives could be seen as Darwinian individuals (Figure 4.2). A second advantage of the tree representation is that edges and vertices can be coloured to denote both qualitative and quantitative information about the culture. A special purpose interactive visualisation interface was developed to allow the quick visualisation and cross-referencing of these rich data sets. Compact static representations are more suited to the publication of the results and will be presented in the context of this manuscript.

Figure 4.6 shows the genealogies obtained by analysing the data produced by Remigi and Rexin for the four different experimental conditions. Vertices are coloured according to the fate of the culture at the end of the phase (survived, or extinct), and presented in parallel with the proportion of extinction events represented as stacked bar plots. The tree structure is more information-rich than proportions because it allows fast inspection of the correlation between the genealogy and the causes of extinction. Indeed, a first observation is that individuals of the same clutch (sharing the same parent) seem to share similar fates. Of course, patterns derived from visual inspection need to be confirmed by statistical analysis, but the benefit of quick exploration of the data set is that this costly procedure will only be applied to promising traits. Additionally, highlighting of the coalescent tree of the population, that is the descent of all extant collectives at the end of the experiment, paints a vivid picture of the eco-evolutionary dynamics of the population. For instance the genealogy of wild type cells in large microcosms (WT-L) is completely coalesced by the end of the experiment, meaning that all collectives are comprised of cells that descend from a single ancestral collective within cycle 3. In contrast, when using the previously selected lineage (17-S), half of the initial lineages still have a descent at the end of the experiment. The difference in average mortality can be assessed from the proportion of extinct collectives, but the knowledge of the full genealogy allows recognition of the heterogeneity between lineages. It is also a first indication of potential beneficial mutations: it is possible that a lineage that ended up being fixed in the population might bear genetic differences that enhance its chance of survival.

Overall, clear visualisation of complex experimental results is a critical

first step toward their interpretation. Collective genealogies are rich representations that illustrate patterns and act as initial guide for further statistical analyses, or even additional experiments. The next section presents such an example of further analysis, and a model for the inference of survival probability across the whole genealogy.

### 4.3 Survival probability estimation

Pairing each collective with an adaptive value, or fitness measure, is a complex task but central to any evolutionary interpretation of the system. Different methods exist derived from evolutionary models or from additional experiments. Model-fitting based methods may extract biologically relevant parameters from the existing data. In contrast, additional fitness assay may be performed by repeating the culture from frozen stock sampled from each collective. The Bayesian-Network model presented in this section aims to be a flexible method with minimal biological assumptions that allow assignment to each collective a most probable survival probability. This model is designed to use the knowledge of the full collective-level genealogy while being able to incorporate new fitness assay data that may be produced later.

The proposed method relies on a minimal probabilistic model of the observation of collective survival, and the transmission of the survival probability from one collective to its collective offspring. The output of the model is both a list of survival rate for each node of the genealogy, and the fit of a single mutation rate parameter issued from the most probable configuration of the model.

As a pre-processing step, the genealogy is split into two data sets. The two alternating phases involve very different processes and are likely to yield different survival probabilities. As a consequence, two sub-trees were extracted of each genealogy, one for the phase one and one for the phase two (Figure 4.7) and analysed separately.

#### Mathematical model

The goal of the model was to assign to each node a most probable adaptive value. In technical terms, the “most probable adaptive value” refers to a maximum likelihood estimation (MLE) that is a way to assign, under a defined graphical probabilistic model, a value to a parameter that maximise the probability of the observed result. The adaptive value is here a hidden variable: one does not simply observe the adaptive value of a collective, but performs experiments that provide information on its true value.

This endeavour requires encoding of both the observations and the hidden characteristics of the collectives in a model, that takes the form of a joint probability distribution. Second, a method to find the most likely assignation of an adaptive value to each culture within the genealogy, for a given mutation rate, will be presented. Conversely, parameter inference allows assignment of the most likely value of a mutation rate for a given configuration. Finally, the iterative method of Expectation-Maximisation will be used to concurrently fit to the data both the adaptive value and the mutation rate.

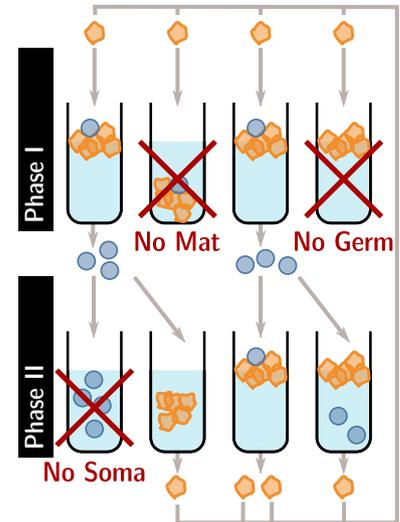


Figure 4.5: **Experimental birth and death events.** Collectives lineages that fail to either produce a viable and fecund mat (phase 1) or to revert to soma-like cells (phase 2) free niches (collective death). This provides opportunities for viable collective lineages to export their success and replace the extinct lineage in the next phase (collective birth). The total number of niches is constant

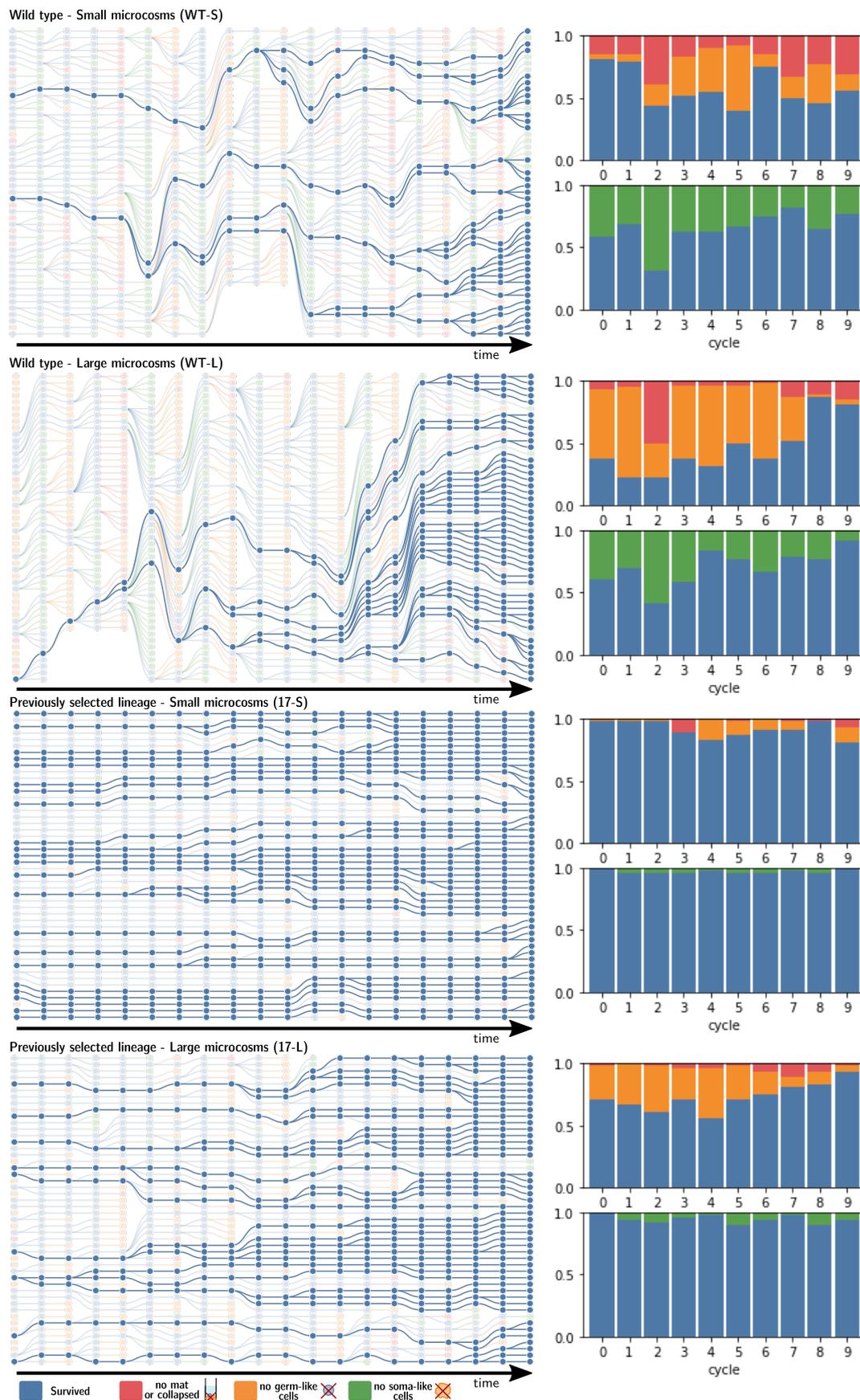


Figure 4.6: **Experimental collective-level genealogies.** The collective-level selection experiment yields a rich data set of 48 parallel microcosms (the collectives) with associated extinction (colours) and replacement events (edges). A collective-level genealogy in the shape of a directed acyclic graph (left) offer a synthetic but more granular visualisation than looking at the proportion of events at each phase (right). See main text for details. (Data collected by Remigi and Rexin).

### Model structure

Each collective genealogy  $\mathcal{T} = (V, \mathcal{E})$  is a forest containing  $N$  numbered vertices:  $V = 1 \dots N$  which are the collectives, and an edge linking each collective  $i$  to its parent  $p_i$ :  $\mathcal{E} = (i, p_i)_{i=1 \dots N}$ .

The model assigns to each collective an observed and a hidden variable. For a collective  $i$ , let  $E_i$  be the random variable encoding whether this collective went extinct ( $E_i = 1$ ) or not ( $E_i = 0$ ). In the field of probabilistic graphical models the observed values are called “evidence”

The first assumption of the model is that there exists a hidden random variable for each collective  $i$ , called  $X_i$ , representing its survival probability. In probabilistic terms, this means that the conditional probability of extinction of a collective  $i$  knowing the value of the survival probability  $x_i$  is exactly the value of  $x_i$ :

$$\mathbb{P}(E_i = e_i | X_i = x_i) = \begin{cases} x_i & \text{if } e_i = 0 \\ 1 - x_i & \text{if } e_i = 1 \end{cases} \quad (4.1)$$

The second assumption of the model is that the survival probability of a collective is transmitted from one collective generation to the next, with an unbiased small variation. In probabilistic terms it means that if the parent collective has survival probability  $x_{p_i}$ , then the survival probability of its offspring collective follows a truncated Gaussian distribution (between 0 and 1) with mean  $x_{p_i}$  and standard deviation  $\sigma$ . The value of  $\sigma$  is the only parameter of the model:

$$\mathbb{P}(X_i \in A | X_{p_i} = x_{p_i}) = C \int_A \exp\left(-\frac{(x - x_{p_i})^2}{2\sigma^2}\right) dx \quad (4.2)$$

With  $A \subset [0, 1]$ , and  $C$  a scaling factor ensuring that  $\int_0^1 \mathbb{P}(X_i = x | X_{p_i} = x_{p_i}) dx = 1$ .

During the experiment, the values of  $E_i = e_i$  (extinct or not) are observed whereas the values of  $X_i$  are “hidden”.

### Simple Bayesian inference

Consider a collective  $i$  with associated survival probability  $X_i$  and experimental outcome  $E_i$ . The prior probability distribution of  $X_i$  is given by Equation 4.2. Consider for an instant that  $x_{p_i}$ , the adaptive value of its parent is a fixed parameter. Bayes’ Theorem gives the expression of the probability of  $X_i$ , conditional on the observed value of  $E_i$ :

$$\mathbb{P}(X_i | E_i) = \frac{\mathbb{P}(E_i | X_i) \mathbb{P}(X_i)}{\mathbb{P}(E_i)} \quad (4.3)$$

Where  $\mathbb{P}(X_i)$  is the prior probability of  $X_i$ ,  $\mathbb{P}(E_i | X_i)$  is the observation model from Equation 4.1 and  $\mathbb{P}(E_i)$  is the probability of the evidence. The maximum likelihood estimate (or more precisely the maximum a posteriori estimate, MAP) of the value of  $X_i$  is given by taking the value of  $x$  that maximises  $\mathbb{P}(X_i = x | E_i)$ , note that the value of the MAP does not depend on  $\mathbb{P}(E_i)$ , which is a simple scaling constant. Figure 4.8 shows the consequence

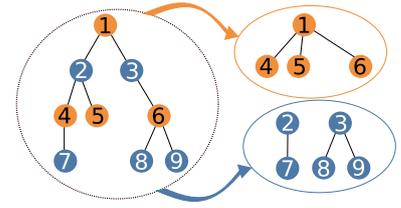


Figure 4.7: **Subtree extraction procedure.** Phase 1 (selection for a somatic mat and germ-like cells) and phase 2 (selection for the reversion to soma-like cells) are qualitatively different and are treated separately. In order to achieve this, each collective genealogy is split into a phase 1 and a phase 2 tree that are analysed on their own, and re-assembled for visualisation.

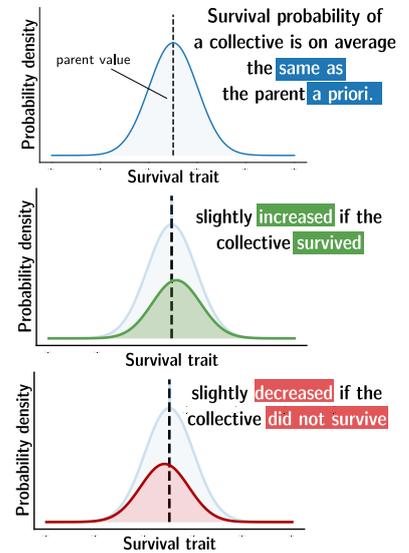
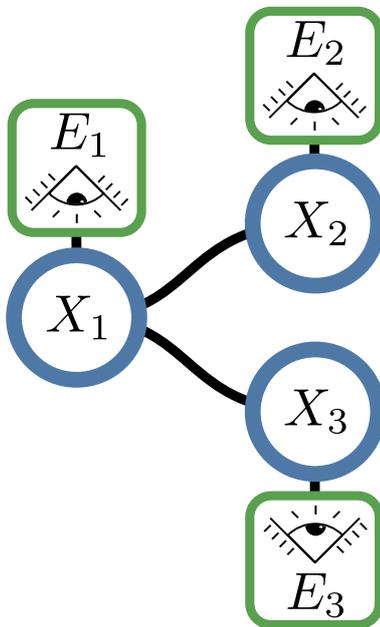


Figure 4.8: **Bayesian Model of Heredity.** In blue  $\mathbb{P}(X_i | X_{p_i})$ , in green  $\mathbb{P}(X_i | X_{p_i}) \mathbb{P}(E_i = 0 | X_i)$ , in red  $\mathbb{P}(X_i | X_{p_i}) \mathbb{P}(E_i = 1 | X_i)$ .

of observing the outcome of the experiment (survival or extinction) on the posterior: the whole distribution (and consequently the MAP) is shifted toward higher values if the collective survived, or lower values if the collective went extinct.

The problem lies, as often in Bayesian statistics, in the prior probability of  $X_i$  (Efron, 2013). Indeed, in the context of the experiment, this probability depends on the state of all the ancestors of the collective  $i$ . Applying Bayes theorem to all those interdependent observations quickly becomes a complex undertaking without a proper method. Evidence propagation, presented hereafter, is this optimal method.

The rest of this section details the technical steps required to find the most probable values of  $X_i$  throughout the genealogy.



### Finding the most likely configuration of the genealogy

The configuration of the full genealogy is the knowledge of the state of all vertices. Hence, it is encoded as a joint distribution that maps each configuration to a probability:

$$(\mathbf{x}, \mathbf{e}) \mapsto \mathbb{P}(\mathbf{X} = \mathbf{x}, \mathbf{E} = \mathbf{e}) \quad (4.4)$$

Where bold letters are used to represent vectors of dimension  $N$ ,  $\mathbf{x} = (x_i)_{i \in V}$ ,  $\mathbf{X} = (X_i)_{i \in V}$ ,  $\mathbf{E} = (E_i)_{i \in V}$ ,  $\mathbf{e} = (e_i)_{i \in V}$ .

The likelihood of a given set of adaptive values  $\mathbf{x}$  is defined as the probability of observing this value, conditional on the state of all collectives:

$$L(\mathbf{x}) = \mathbb{P}(\mathbf{X} = \mathbf{x} | \mathbf{E} = \mathbf{e}) \quad (4.5)$$

Figure 4.9: **Bayesian Network.** Bayesian network corresponding to a genealogy of three collectives. Collective 1 has two offsprings: collectives 2 and 3. Vertices are random variables, edges are conditional dependence. The hidden variables  $X$  encode the survival probability of the collectives, while the observed variables  $E$  encode the outcome of the experiment (survived or not).

The maximum a posteriori configuration of the network is given by finding the value of  $\mathbf{x}$  that maximises  $L(\mathbf{x})$ . The naïve, brute force method to find this value is to compute the  $L$  for all  $\text{Card}(\omega)^N$  possible configurations (where  $\text{Card}(\omega)$  is the number of possible value of  $X$ . If the adaptive value is discretised in a hundred of values, this gives  $100^{48}$  computations (in the order of the number of atom of earth  $\approx 10^{50}$ , and way too much even for modern computers).

However, the probability distribution can be efficiently factorised thanks to the definition of conditional probabilities whose structure is represented in the graph: each pair of nodes not linked by an edge is conditionally independent of one another. As an example consider the small network in Figure 4.9. The factorisation will go like this:

$$\begin{aligned}
\mathbf{x}^* &= \arg \max_{\mathbf{x}} \mathbb{P}(X_1, X_2, X_3 | E_1, E_2, E_3) \\
&= \arg \max_{\mathbf{x}} \frac{\mathbb{P}(X_1, X_2, X_3, E_1, E_2, E_3)}{\mathbb{P}(E_1, E_2, E_3)} \quad (\text{Definition of conditional probability}) \\
&= \arg \max_{\mathbf{x}} \mathbb{P}(X_1, X_2, X_3, E_1, E_2, E_3) \quad (\text{Denominator constant with respect to } \mathbf{x}) \\
&= \arg \max_{\mathbf{x}} \mathbb{P}(X_1) \mathbb{P}(E_1 | X_1) \mathbb{P}(X_2 | X_1) \mathbb{P}(E_2 | X_2) \mathbb{P}(X_3 | X_1) \mathbb{P}(E_3 | X_3) \quad (\text{Conditional independence}) \\
&= \arg \max_{X_1} \left[ \underbrace{\mathbb{P}(X_1) \mathbb{P}(E_1 | X_1)}_{\phi_1(X_1)} \underbrace{\arg \max_{X_2} (\mathbb{P}(X_2 | X_1) \mathbb{P}(E_2 | X_2))}_{\phi_2(X_1)} \underbrace{\arg \max_{X_3} (\mathbb{P}(X_3 | X_1) \mathbb{P}(E_3 | X_3))}_{\phi_3(X_1)} \right] \\
&= \arg \max_{X_1} \phi_1(X_1) \phi_2(X_1) \phi_3(X_1) \quad (\text{Variable elimination of } X_1, X_2)
\end{aligned}$$

Additionally, the “maximum” operation is distributive, which allows for an efficient computation of local maxima ( $\phi_2, \phi_3$ , called messages) that depend on a single variable. This constitutes the essence of the Max-Product Belief propagation algorithm (Koller and Friedman (2009), Chapter 13).

### Expectation-Maximisation procedure

Now, how can one find the value of the parameters of the model? The Expectation-Maximisation (EM) algorithm (Dempster et al., 1977), which might be one of the most influential advances in modern computational statistics and is widely used in Gaussian Mixtures, and Hidden Markov Models, provides an iterative solution.

First, note that for any given configuration of the network (values of  $\mathbf{X}$  and  $\mathbf{E}$ ), finding the maximum likelihood estimate of the parameter (mutation rate  $\sigma$ ) is a simple matter of computing the value of the parameter that maximises the likelihood  $\mathbb{P}(\mathbf{X}, \mathbf{E})$ .

Now, starting from an arbitrary value of parameters (say,  $\sigma = \frac{1}{2}$ ), one may compute the Maximum A Posteriori configuration of the network, that assigns a value to each hidden variable. This constitutes the Expectation (E) step. Then, a new value of the parameter is computed using the Maximum of Likelihood estimate. This constitutes the maximisation step (M). The procedure can be iterated until convergence. This algorithm is guaranteed to reach a local maximum in likelihood, but may not reach the global maximum. However, its conceptual simplicity means that it is often the best heuristic available for practical applications.

### Results

The four genealogies were separated into eight phase-trees (p1, p2) and independently analysed using the expectation maximisation method presented above. Figure 4.10 shows that the E-M method results in an increase in likelihood. However, the optimisation method resulted in minimising the parameter  $\sigma$  within the limit of the discretisation grid used (200 values linearly distributed between 0 and 1). As a consequence, the global likelihood maximum may not have been reached, and it is impossible to interpret the differences

in  $\sigma$  from one tree to another. Further analyses are required to refine these results.

Figure 4.11 presents the trajectory of the four experiments within the survival value space. Note that the four experiments end up improving the survival probability in both life-cycle phases. Strain that already underwent several cycles of selection in the original experiment (17-S and 17-L), present a lower relative improvement than wild type strains (WT-S, WT-L). Experiment WT-S show large improvement in the phase 2, coupled to initial decrease in the survival probability in phase 1, this might be the signature of an underlying trade-off in the biological functions required to survive both phases. Experiment WT-L presents the most spectacular relative improvement in both phases.

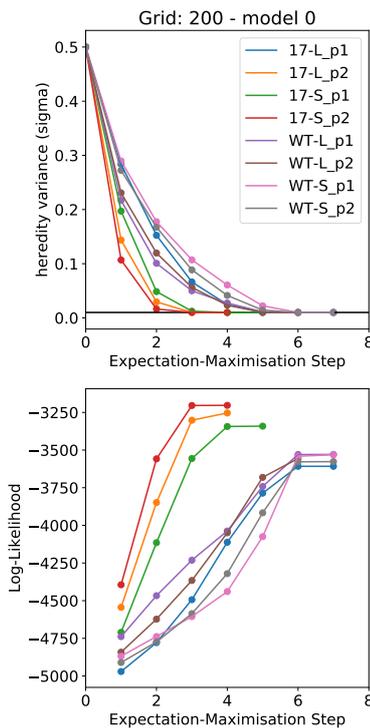


Figure 4.10: **Convergence of the expectation maximisation algorithm.** Top: Value of the parameter through EM cycles, the black horizontal line is the minimal value allowed. Bottom: Associated Log-Likelihood. Convergence of the E-M algorithm is currently limited by the discretisation grid of the parameter.

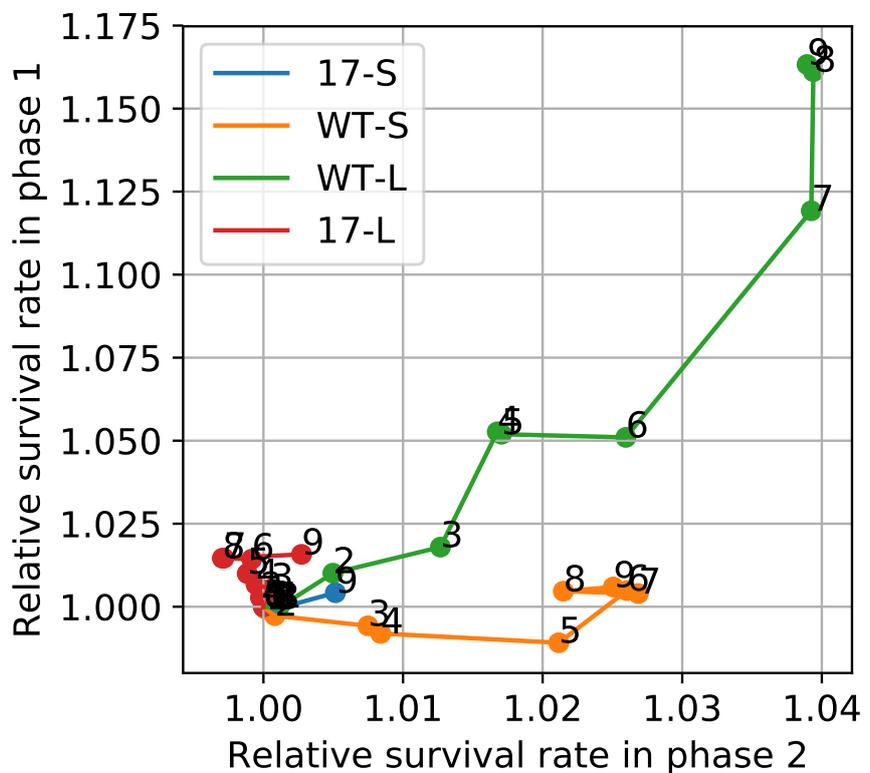


Figure 4.11: **Average Inferred Survival rates through time.** Numbers represent the cycle. Values have been normalised to start in (1,1) and facilitate comparison.

Figure 4.12 presents the histogram of survival probability change for the four experiments and the two phases. Note that, at the resolution given by the current discretisation grid, the typical offspring has the same survival probability than its parent (the mode corresponds to no changes). Phase one (selection for a viable mat, and the presence of germ-like cells) seems to be

more prone to improvement than Phase two (selection for the reversion to soma-like cells), except in the experiment 17-L where the survival probability decreased several times.

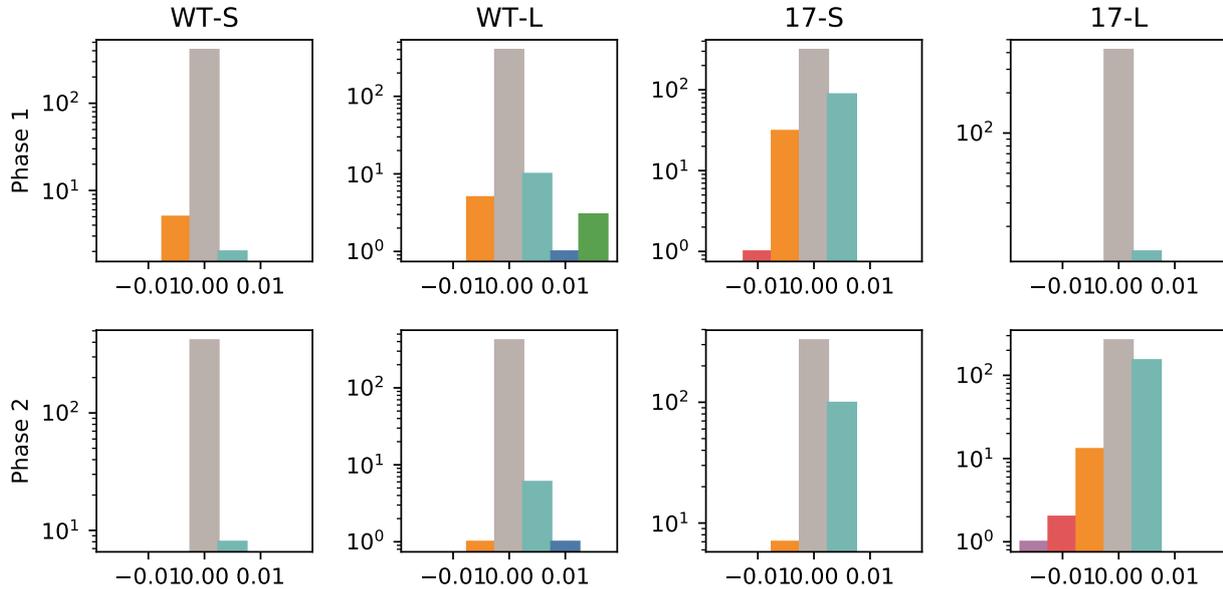


Figure 4.12: **Distribution of the change in survival probability.** The top row concern the phase 1 (selection of germ-like cells as well as a viable mat), while the bottom row concern the phase two (selection for the reversion to soma-like cells). The columns represent different treatment. WT: wild type, 17: Lineage 17 from [Hammerschmidt et al. \(2014\)](#), S: small microcosms, L: large microcosms.

The relative survival improvement can only be really appreciated when put into the proper context of the collective genealogy. Figure 4.13 shows an example for the WT-L experiment. Note how the survival probability improvements are localised in the beginning of the genealogy. The best improvements are localised at or before the coalescence of the whole population. Every collective associated with a decrease in survival probability has no descendent in the last cycle. This shows that “unfit” phenotypes are readily purged by the collective-level selection procedure.

#### 4.4 Propagating sequencing information to the full genealogy

Collective genealogies make manifest eco-evolutionary dynamics of collectives. Survival probabilities assigned by the model from the previous section are a quantitative measurement of adaptation. However, if they reveal evolution they do not contribute to explanation for the underlying molecular mechanisms of adaptation. The genetic mechanisms associated with the over-production of cellulose and eventually bacterial mats has been traced to

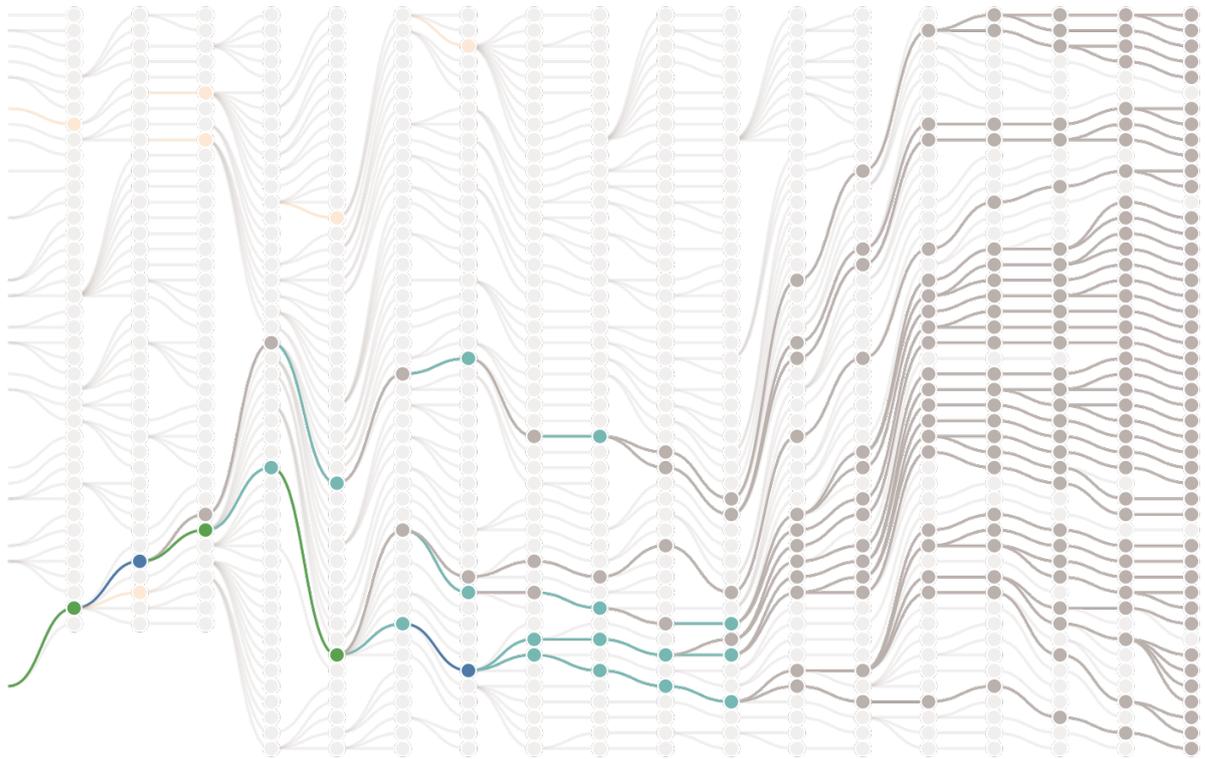


Figure 4.13: **Position of survival probability improvement in the genealogy**, color code from Figure 4.12

key mutations, for instance in the *wss* operon regulating cellulose synthesis (Bantinaki et al., 2007; McDonald et al., 2009; Lind et al., 2017). Additionally, the lineage improvement in the collective-level selection experiment has been associated with mutations in the DNA mismatch repair system, increasing mutation rate and thus the ability to complete the proto-life cycle, such as the one affecting the *mutS* locus (Hammerschmidt et al., 2014).

A similar effort has been produced by Remigi and Rexin who sequenced the communities of all endpoints of collective lineages, aligned the genomes and identified mutations. For instance, 3686 mutations from the ancestral wild type (WT) have been identified across all collectives at the end of the 10th cycle of the large-microcosm experiment. The presence of sequenced genetic mutations is an additional information that may be represented on the genealogy. Material constraints mean that it is neither cost effective nor practical to sequence all the microcosms at each cycle. However, when a mutation is found in several cultures, it might be possible to infer the position of its first appearance in the collective genealogies. The method should ideally be a way to point out where in the genealogy additional sequencing should be focused, and allow seamless incorporation of new data as it becomes available.

The belief propagation algorithm presented in the previous section aptly fits these requirements. The first step is to change meaning of the hidden variable  $X_i$  from the adaptive value of collective  $i$  to the presence ( $X_i = 1$ ) or

absence ( $X_i = 0$ ) of the mutation within the collective. Then, the observed variable  $E_i$  which encodes “the collective  $i$  went extinct” is replaced by a new conditionally Bernoulli variable  $O_i$  encoding the event “the mutation of interest was found in sequences obtained from collective  $i$ ” for the collective that were indeed sequenced. Then, the observation model must be specified:

$$\mathbb{P}(O_i = i | X_i = j) = \begin{bmatrix} p & 1-p \\ 1-q & q \end{bmatrix}_{i,j} \quad (4.6)$$

Where  $p$  and  $q$  represent, respectively, the probability of true negative and true positive reporting. In the following  $p = q = 1$ . Finally, the heredity model is specified as:

$$\mathbb{P}(X_i = i | X_{p_i} = j) = \begin{bmatrix} 1-\mu & \mu \\ \mu_r & 1-\mu_r \end{bmatrix}_{i,j} \quad (4.7)$$

Where  $\mu$  is probability that the mutation appears within one cycle and  $\mu_r$  the probability of losing the mutation. Ideally,  $\mu$  and  $\mu_r$  could be estimated from a mechanistic model of mutation (see chapter 3), or fitted using the expectation-maximisation procedure presented in the previous section. In the following, arbitrary values of  $\mu = 0.001$  and  $\mu_r = 0.00001$  have been selected.

Figure 4.14 presents the result of the belief propagation for two mutations of interest within the large microcosm, wild type treatment (WT-L). The two mutations have been selected for illustrative purpose, and a more systematic study of the full set of mutation is ongoing.

First, mutation 25 (in green, and red) corresponds to a mutation on the *mutL* locus, which encodes a protein necessary for the DNA mismatch repair system. This mutation is found in all the sequenced cultures at the end of the experiment. Naturally, the model predicts that it emerged at or before the most recent common ancestor of the whole population, at cycle 2. New sequencing effort should be concentrated around this position in the genealogy to refine this prediction.

Second, mutation 85 (in red) corresponds to a mutation on the *wssE* locus, part of the *wss* operon that regulates cellulose synthesis. Mutation 85 is shared by 22 sequenced collectives and its predicted first appearance within the genealogy is at cycle 5, in a collective that is concurrently predicted to have witnessed a significant improvement in survival probability (Figure 4.13).

These promising preliminary results exemplify the interest of combining sequencing and statistical modelling approaches in order to efficiently exploit the wealth of data produced by the experimental setup.

## 4.5 Discussion

This chapter has formalised the concept of collective genealogy and introduced tools, both graphical and statistical, for their study. This work, motivated and illustrated by the rich data set produced by Remigi and Rexin’s ongoing experiment addresses three main points: how to efficiently visualise collective genealogies, how to estimate a collective-level adaptive value in the context of these experiments, and finally how to propagate mutational information

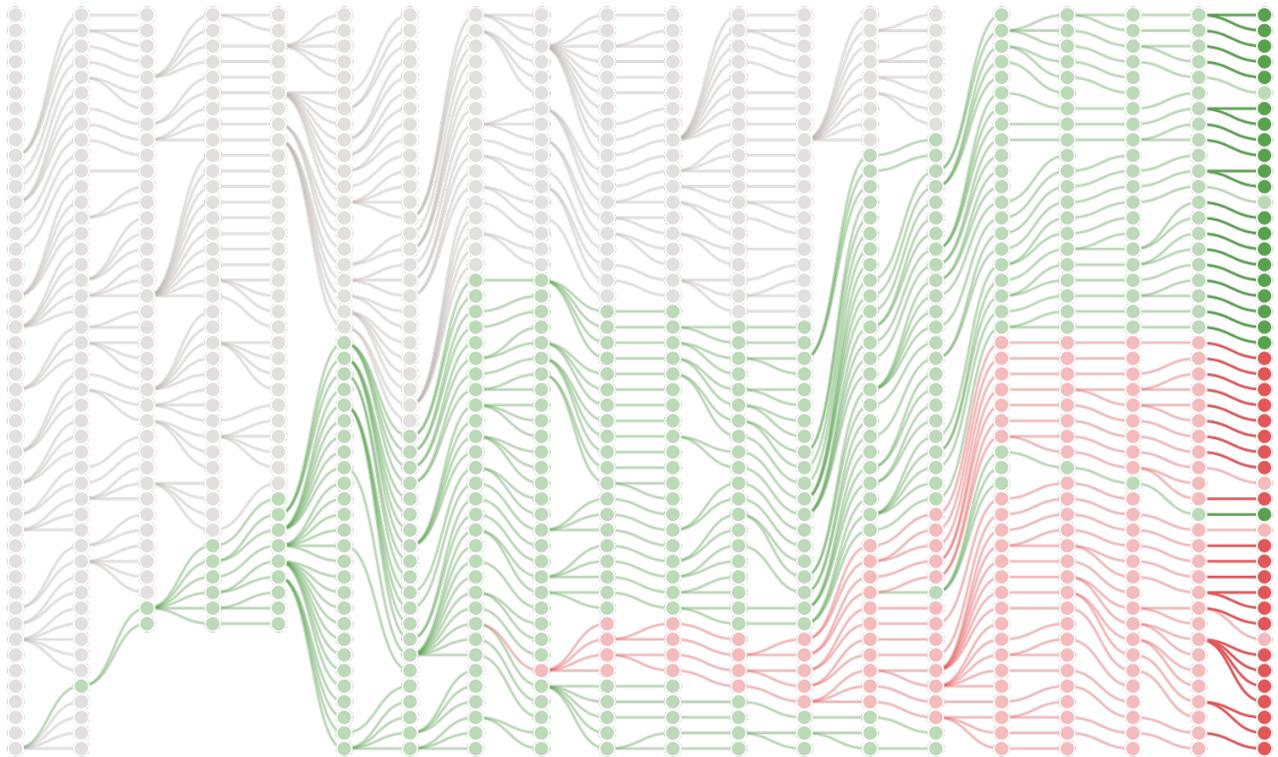


Figure 4.14: **Inferring the existence of mutation in the collective genealogy**, Opaque vertices are the sequenced collectives, transparent vertices are inferred by the model. Collectives in green are predicted to bear the point mutation MUT0025 that corresponds to a point mutation from a T to a G at within gene *mutL*. Collectives in red are predicted to additionally bear the point mutation MUT0085 that corresponds to a point mutation from a C to a T within gene *wssE*.

to the whole genealogy. Collective genealogies are naturally represented as trees or forests, where the relevant information can be highlighted by varying the colour and opacity of vertices. A compact layout is preferred for printed publication (as here) but interactive exploration (developed, but not presented here) offers real advantages to the practitioner. Graphical probabilistic models (such as Bayesian networks) for statistical inference have been used in haplotyping of pedigree (Fishelson et al., 2005), gene network inference (Needham et al., 2007) and genetic diseases risk assessment (Nuel et al., 2017). This flexible method presents several advantages. Firstly, it is designed to work with partial data: inference can be performed even when sequence data is missing from collectives at certain time points. Secondly, it allows the seamless integration of new data points: additional sequencing or fitness assays may be added to refine the inference. Thus, it provides a means for feedback between data collection and inference of the next promising experiment according to current knowledge. Finally, the statistical model of observation and transmission of characters is relatively simple to define and

can be adjusted for different purposes, as presented for the tracking of mutations or the estimation of survival probability. While the models used above were minimal in their assumptions, it is possible to consider more complex mechanistic models when the question requires it.

Preliminary analysis of the experimental results generally confirms the conclusions of [Hammerschmidt et al. \(2014\)](#) with a new level of granularity. First, adaptation results from imposition of the collective-level selection regime alternating between selection of mats that are viable (i.e., non collapsed) and fecund (i.e., producing germ-like cells) and selection for the reversion to soma-like from germ-like cells. Second, these improvements can be quantified in term of survival probability by our statistical model. Third, the improvement correlates with, and might originate from, mutations in key genes in the regulation of cellulose production and DNA mismatch repair mechanisms.

Characterising, let alone quantifying, collective-level adaptation is a complex problem. If few authors disagree that a group could in theory constitute units of evolution on their own right, the impact of this process in evolution in general is still debated. To paraphrase [Williams \(2018\)](#), the question is to distinguish a collective of adapted bacteria from an adapted collective of bacteria. Several theoretical frameworks exist to address this question (Reviewed in [Rose and Rainey, unpublished](#)), but let us limit the discussion to (Nested) Darwinian Populations ([Godfrey-Smith \(2009\)](#), reviewed in Chapter 2).

In this framework, “a transition in individuality involves the appearance of a new kind of Darwinian population” ([Godfrey-Smith, 2009](#)), meaning that the collectives themselves must display some minimal level of Darwinian properties. The level of Darwinian properties can in theory be measured on a scale that qualitatively goes from marginal to paradigmatic and be defined in several dimensions. There is no general way to quantify those properties yet, but this objective already inspired the work that was presented in this chapter. Two dimensions of the Darwinian Properties Space are of particular interest when considering a transition in multicellularity: the fidelity of heredity ( $H$ ) and the dependence of reproductive differences on intrinsic characters ( $S$ ). Without one or the other, an entity may not qualify as Darwinian, and there is no reason to expect the evolution of adapted trait values by the mean of natural selection.

A marginal level of heredity ( $H$ ) proceeds from the population structure under consideration. Indeed, the protocol prevents any transfer of matter between collectives. This sidesteps the mixing that may disqualify numerous biological systems such as genotypes under meiosis ([Williams, 2018](#)) or groups in traits-groups models ([Wilson, 1975a](#)). A more complete discussion on how ecological scaffolding might affect marginal level of heredity and potentially increase it toward the paradigmatic end of the spectrum is the subject of Chapter 6.

A marginal level of  $S$  is also ensured by the protocol since collective birth and death events are effected by the experimenter based on phenotype assessment (presence of germ or soma-like cells, presence of a mat). From this premise, ecological scaffolding theory predicts that the collective level selection will favour lineages with improved  $S$ , meaning that the marginal

Darwinian properties granted by the population structure may be refined by natural selection itself. Testing this prediction requires a way to assess  $S$ . As mentioned before, there is no general method to do so yet, but by definition, a measurement of  $S$  must link intrinsic characteristics of collectives (i.e., traits) to reproductive differences. The choice of collective traits to assess is a first thorny issue, due to the classical counterargument that they are mere byproducts of lower level processes (Williams, 2018). However, it has been established that the fact that collective traits are causally dependant from individual traits is not an obstacle to the existence of a collective-level selection process (Okasha, 2012). In a recent development, Rose and Rainey (Unpublished) propose to turn the problem on its head and only assign a trait to a level of organisation once it has been proved to be selected at this given level. The switching rate from germ-like to soma-like cells is a natural candidate in the experiment under consideration. Once a trait is chosen, it is necessary to measure its impact on reproductive differences. The proportion of extinct lineages is a good first approximation (Hammerschmidt et al., 2014). However, fitting a survival probability to each collective, as presented in this chapter, offers a more granular alternative. Moreover, survival probability is an unmediated characteristic of demographic events that are the mechanistic basis of evolutionary trajectory (Doebeli et al., 2017). Overall, the tools presented in this chapter will help further inquiries by allowing to easily map traits to collective-level genealogies, while assessing reproductive success in the population in a single, unified framework.

For all those reasons the establishment and study of experimental collective genealogies may well be the most exciting short term prospect in the study of major evolutionary transitions by ecological scaffolding. Microbiology experiments are however limited in time: even if they can last decades, they cannot explore the same time scales as biological life on earth. Consequently, the next chapter presents a model of the long term evolution of the pond-scum thought experiment with a focus on the trade-off between survival and reproduction at the collective level.

## Chapter 5

# From Particles Traits to Collective-level Demography

“Dans un être vivant, tout est agencé en vue de la reproduction. Une bactérie, une amibe, une fougère, de quel destin peuvent-elles rêver sinon de former deux bactéries, deux amibes, plusieurs fougères ?”

— FRANÇOIS JACOB, *La Logique du Vivant: Une histoire de l'hérédité* (1970)

A TRANSITION IN INDIVIDUALITY is the evolutionary process by which a collection of entities (molecules, cells, individuals) get to form a Darwinian individual on its own right (cell, multicellular organism, eusocial community) (Maynard Smith and Szathmary, 1995; Okasha, 2006). An ecological starting point for such transitions is ecological scaffolding (Black et al., 2019). This scenario postulates a metapopulation structure, with patchily distributed resources, limited migration between patches and a selective process at the level of the patches. In this context, evolution by natural selection is predicted to favour the refinement of collective-level mechanisms of reproduction and survival, leading to potentially ever-increasing integrated organisation levels.

As discussed in Chapter 4, artificial selection imposed at the level of collectives can result in adaptive changes that allow the completion of a collective-level proto-lifecycle. This lifecycle involves the reliable production of a proto-somatic "mat" structure and associated proto-germ line (phase 1), as well as the re-establishment of somatic cells from propagules after dispersal (phase 2). This experiment is an instance of a more general thought experiment (the pond-scum scenario described in section 4.1) that illustrates the emergence of multicellularity by ecological scaffolding. In the former, most of the collective population dynamics is simplified compared to the thought experiment, in order to offer a clear demonstration of the effects of selection. The main differences are that in the actual experiment collective population dynamics is highly simplified, and no selection is imposed on the ability for a collective to disperse its propagules. First, since collective-level birth and death events are acted by the experimenter, the population can be held at a constant size, without risk of extinction by the concurrent death of all collectives. Second,

since selection for dispersal is limited to the sole germ-like cells in phase 2, there is no competition between collectives for the quantity of propagules produced. Another feature of the experimental design is that there is no mixing between collectives, preventing direct competition between cell lineages in a single tube.

This chapter is focused on the long-term evolutionary dynamics of nested populations in a pond-scum-like scenario. It clarifies the link between particle traits, particle eco-evolutionary dynamics, collective life-history traits and collective eco-evolutionary dynamics.

We start by building a deterministic model of the discrete-generation process used in experimental evolution, with fixed collective population size and colonization of empty demes (following collective death) with propagules from a well-mixed pool (Hammerschmidt et al., 2014). We show that, in the absence of constraints derived from particle ecology, evolution comes to a halt with the invasion of "perfect collective lineages", that never go extinct and therefore leave no space for further dispersal events. In reality however, particle-level ecology constrains collective evolution, so that immortal lineages are not practically achievable. In the following, therefore, we explore different hypotheses on the ecological underpinning of collective functions and examine the coupling between evolutionary trajectory and emergent trade-offs between survival and reproduction. In a second time, we relax the constraint that the collective population size is constant, first introduced for adherence to the experimental system. We modify the model to explicitly take into account the population dynamics of collectives within a fixed set of environmental niches. The resulting model shares its basic structure with classical epidemiology models and provides simple conditions on the niche density for the population to be viable. Finally, we allow the possibility of invasion of cells from one collective to another. This changes the evolutionary endpoint of the population, making it more dependent to the density of niches.

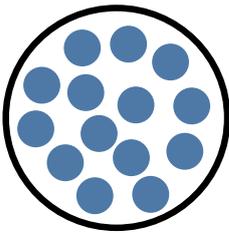


Figure 5.1: **An isolated collective.** Blue discs represent the particles, the black outline represents the boundary. In the experiment from chapter 4 particles are bacteria, boundaries are the test-tube walls.

## 5.1 Evolution of collective survival and reproduction

Consider a set of collectives comprised of particles. For instance, a collective could be the community of cells living attached to a single physical support in the pond-scum thought experiment or the content of a test tube in an actual microbiology experiment. As in the evolutionary experiments from chapter 4, each collective occupies a niche that is completely isolated from the others and undergoes discrete, synchronised generations (Figure 5.2). The number of collectives is constant through time, and is noted  $D$ . At regular intervals of duration  $T$ , the operator assesses a feature of the collective that determines the probability it will survive (Figure 5.1). Each collective is indexed by a number  $i = 1 \dots D$ . A collective  $i$  is marked for extinction and is trashed with a probability  $1 - \sigma_i \in [0, 1]$ , where  $\sigma_i$  is the survival probability of collective  $i$ . The probability of extinction is different from one collective to another and generally depends on both the selection rule (for instance only viable and fertile mats survive in the experiment from chapter 4) and the composition of the collective (some collectives might contain cells that produce more cellulose, resulting in stronger mats, and a higher survival probability).

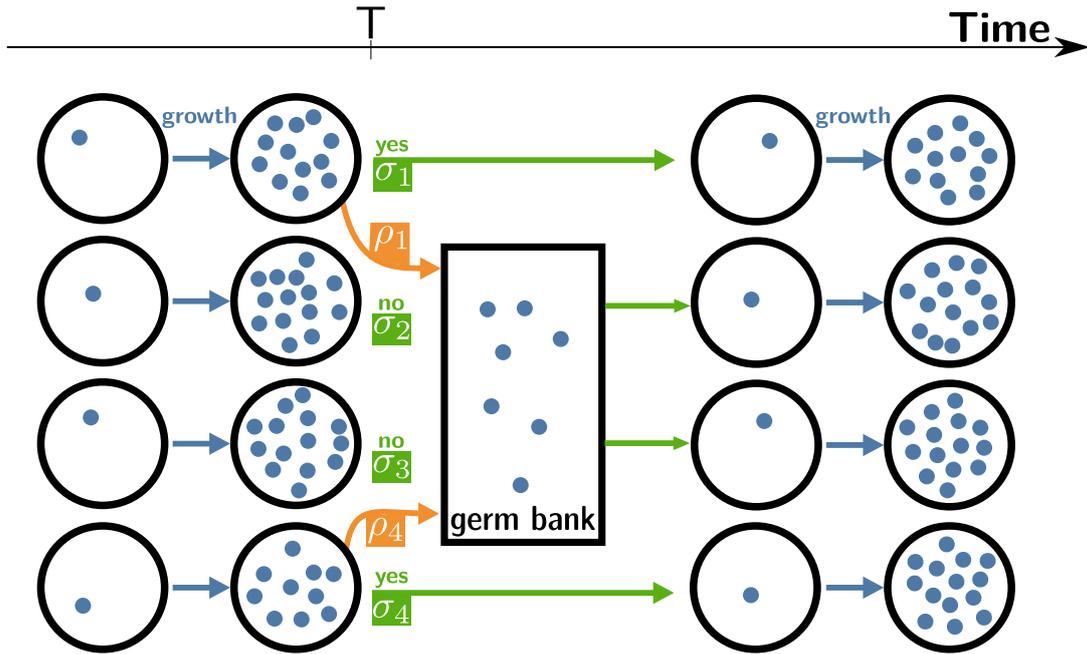


Figure 5.2: **A model of non-overlapping collective generations.** A fixed-size population of  $D$  collectives undergoes discrete generations of duration  $T$ . Each collective  $i$  is characterised by a pair of  $(\sigma_i, \rho_i)$  life-history traits. At the end of each generation, a collective may survive with probability  $\sigma_i$ , and seed its own niche in the next generation or die with probability  $1 - \sigma_i$ . Collectives that survive contribute to a global germ bank with a quantity  $\rho_i$  of propagules. Niches emptied by the death of a collective are seeded a single propagule particle drawn uniformly from the germ bank.

The following “collective generation” starts by establishing a new population of  $D$  collectives. Each new collective is seeded with exactly one particle. First, each collective that survived seeds a single new collective. Thus, collectives that survive the selection step keep occupying their niches without interference, even though (just like in the experiment) the composing particles population is reset to a single cell at each generation. In the perspective of the collective genealogy (see Chapter 4), survival corresponds to the continuation of a collective lineage. The empty niches remaining after collective death are filled randomly, drawing particles from a global germ bank to which every surviving collective  $i$  contributes according to its propagule output  $\rho_i$ . Thus, the probability of drawing a founding particle originating from such collective is proportional to  $\rho_i$  normalised by the size of the gene bank (sum of all propagule outputs for surviving collectives). Note that  $\rho_i$  may depend on the state of the collective but is not necessarily proportional to the collective size.

In the experiment presented in Chapter 4, there is no competition between collectives in propagule quantity: as long as a collective is able to produce even a few germ-like cells during phase 1, it will have the same chance as any other collective to be selected to fill empty niches in the next generation. In the formalism of the model, it means that the reproductive output of all

surviving collectives  $i$  is identical:  $\rho_i = 1$ . Moreover, the model discussed here does not distinguish between phase 1 and phase 2: each generation consists in a single round of selection.

In the rest of the chapter  $\rho$  is referred to indifferently as reproductive output, or as propagule production. Propagule production  $\rho$  and survival probability  $\sigma$  are referred to as the *life history parameters* of the collectives. Their values are assumed to be deterministic functions of the collective state (particle composition, particle traits...). In the rest of this section, I consider the case when life-history traits are generic functions of a collective trait. In the following section, they will be constrained by trade-offs (section 5.2). Finally, section 5.3 considers cases when collective life-history traits derive from specifying the underpinning particle-level ecological model.

### Invasion of a mutant collective

Experiments (described in Chapter 4) showed that a lineage that is more adapted is able to invade and get fixed in the collective population (Figure 5.3). In this simple case one can consider that collective populations are most of the time monomorphic. It is then possible to establish the conditions for successful invasion when the relationship between collective life-history traits  $(\sigma, \rho)$  and particle traits or collective composition are unspecified.

Suppose that the survival probability  $\sigma$  and reproductive output  $\rho$  are deterministic functions of an underlying heritable collective trait  $\theta \in \mathbb{R}^p$ . For instance the collective trait can represent the number of cells in the collective, or be a more complex function of collective composition such as the proportion of germ-like cells, or the overall production of structural molecules. In general, these features are expected to vary along a collective lineage unless they are highly regulated. As a first step, however, let us suppose that they are transmitted from parent to offspring collective. Consider a large population of  $D - 1$  identical collectives, with "resident" trait value  $\theta$ . A single "mutant" collective has instead a trait value of  $\theta'$ . The per capita growth rate of the mutant within this monomorphic population is called its Invasion fitness. Adaptive dynamics theory is a general framework for modelling adaptation by natural selection. It supposes that a mutant lineage will replace the resident if its invasion fitness, the per capita growth rate when mutants replace a vanishing fraction of the resident population, is positive (Geritz et al., 2002). By assuming that the population is large enough, it is possible to give an expression for the invasion fitness as a function of  $\theta$  and  $\theta'$  (Proposition 13).

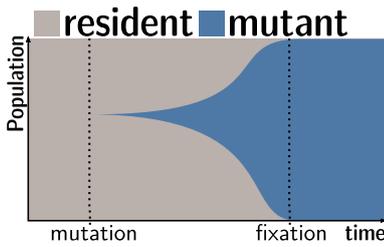


Figure 5.3: **Selective Sweep in a Population.**

#### Proposition 13 (Invasion fitness):

The invasion fitness of rare collectives bearing trait  $\theta'$  in a large  $D \gg 1$  monomorphic population of collectives bearing trait  $\theta$  is:

$$f(\theta, \theta') = \ln \left( \sigma(\theta') \left[ 1 + \rho(\theta') \frac{1 - \sigma(\theta)}{\sigma(\theta)\rho(\theta)} \right] \right) \quad (5.1)$$

(Proof page 111.)

Proposition 13 shows that if the resident population is composed of immortal collective lineages ( $\sigma(\theta) = 1$ ), no mutant collective can ever invade (the invasion fitness is always negative, as  $\forall\theta, \sigma(\theta) \in [0, 1]$  and  $\forall\theta, \rho(\theta) > 0$ ). Furthermore, proposition 13 shows that a mutant lineage that does not produce any propagule for the germ bank ( $\rho(\theta') = 0$ ) will always go to extinction (the invasion fitness will be negative) unless it is immortal ( $\sigma(\theta') = 1$ ). In this case it will survive indefinitely, without being able to invade another collective population (null invasion fitness).

Knowledge of the invasion fitness allows to predict the long-term evolutionary dynamics as a process of successive substitutions of residents by successful mutants. This is addressed in the next section.

### Evolution of collective traits

Suppose that mutations are rare enough so each new mutant can either disappear or go to fixation before a new mutation arises (Figure 5.4 a). It results that the long term evolutionary trajectory of the population is a step-wise trait substitution sequence (Figure 5.4 b). If we additionally consider that the effect of mutations on the trait  $\theta$  is small enough, the resulting trait substitution sequence can be approximated by a continuous trajectory (Figure 5.4 c). This trajectory is modelled by the canonical equation of adaptive dynamics (Dieckmann and Law, 1996), which predicts that the trait substitution sequence will follow the steepest ascent of the invasion fitness surface, that determines the local adaptive landscape of the population. Evolutionary singular strategies are then trait values where the invasion fitness gradient  $g$  is null. Proposition 14 gives the expression of  $g$  corresponding to the invasion fitness defined in Equation 5.1.

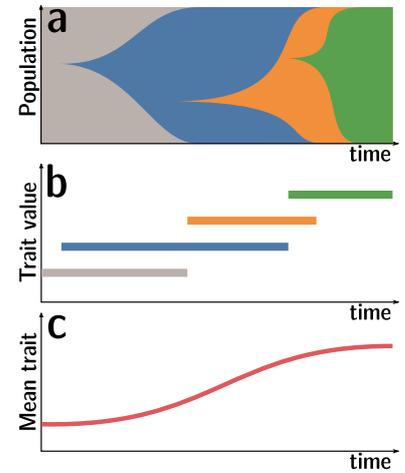


Figure 5.4: **Continuous approximation of successive invasions.** The canonical equation of adaptive dynamics models long-term evolution as a continuous average trait value change due to successive invasion by small phenotypic mutations.

**Proposition 14 (Invasion fitness gradient):**

The invasion fitness gradient around trait  $\theta$  is:

$$g(\theta) := \nabla f(\theta', \theta)|_{\theta'=\theta} = \frac{\nabla\sigma(\theta)}{\sigma(\theta)} + \frac{\nabla\rho(\theta)}{\rho(\theta)}(1 - \sigma(\theta)) \quad (5.2)$$

Where  $\nabla$  denotes the gradient with respect to the first variable, evaluated in  $\theta' = \theta$ .

(Proof page 111)

The invasion fitness gradient (Equation 5.2) is the sum of two terms, the first one being the relative increase in survival probability, the second being the relative increase in contribution to the germ pool, weighted by the probability of death. The advantage in increased reproductive output (in  $\rho$ ) is thus relatively more important when the survival probability is low ( $1 - \sigma$  close to one), than when it is high. In the limit where collective lineages are

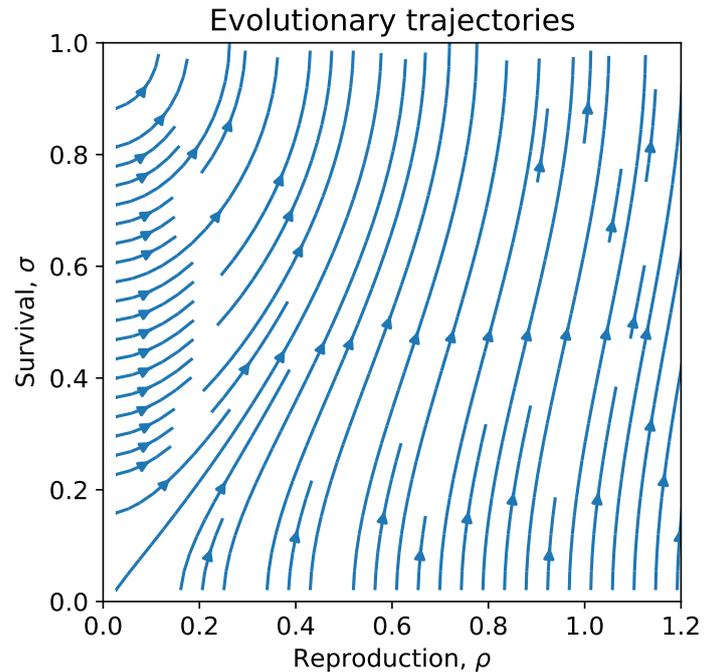


Figure 5.5: **Flow induced by the canonical equation of adaptive dynamics in the absence of evolutionary trade-off between reproduction and survival.** The canonical equation of adaptive dynamics implies that the temporal derivative of the average trait value in the population is proportional to the invasion fitness gradient (Equation 5.2). Here the trait is the vector  $\theta = (\rho, \sigma)$ . Mutations affect both dimensions independently. The average trait in the population is expected to follow the flow lines (in blue) as it is changed by successive invasion of mutants.

immortals ( $\sigma = 1$ ), the relative advantage in reproductive output is irrelevant, and the dynamics is independent of  $\rho$ .

In order to examine the predictions of Equation 5.2, assume now that the collective life-history parameters depend on two independently evolving traits, and that these traits map directly to survival probability and reproductive output:  $\theta = (\theta_1, \theta_2)$ ,  $\rho(\theta) = \theta_1$  and  $\sigma(\theta) = \theta_2$ . Figure 5.5 shows the evolutionary flow resulting from the invasion fitness gradient in this situation. For any meaningful trait value, the local direction of evolution is toward ever-increasing values of  $\rho$  and  $\sigma$ . When the trajectory reaches  $\sigma = 1$  (immortal collective lineages), the flow is null, and the population is at an evolutionary endpoint. There, all collective lineages are immortal and no new variant can invade. On the line  $\sigma = 1$ , reproductive output is neutral. However, since the flow is positive for small perturbation around  $\sigma = 1$ , stochastic effects would favour ever-increasing reproductive output: if, for any reason, a niche was emptied, collectives with the larger propagule output would have an advantage in the competition to colonise it.

The canonical equation is applicable when the conditions of adaptive dy-

namics (rare infinitesimal mutations, heredity of the collective traits) are verified at the collective level, and positive invasion fitness implies fixation. However, the results on the trajectory qualitatively holds even if the condition that mutations are infinitely small is relaxed. For instance, Figure 5.6 shows the result of a stochastic simulation of the model described in Figure 5.2 where each time a collective is founded, it either inherits the parental life history parameters  $\sigma$  and  $\rho$  or mutates one of them by a normally distributed effect. Note how the trajectory of the average trait value in the population is in agreement with the flow induced by the canonical equation of adaptive dynamics (Figure 5.5).

The evolutionary endpoint of this model, consisting in a population of immortal collective lineages, is not the most realistic outcome. In the bacterial mat experiment, any new cell puts a strain on the structural stability of the mat, leading inevitably to a rupture. Mechanical constraints prevents the existence of a mat which is both immortal and contains numerous cells. The next section introduces a trade-off between survival and reproduction to account for this kind of phenomenon. For a start, section 5.2 consider this trade-off to be established directly at the collective level by supposing a common dependence of life-history parameters on a collective trait. Then, in section 5.3 the more general case where the trade-off derives from the properties of the underlying particles is explored.

## 5.2 Trade-off between survival and reproduction

Collective-level evolution would grind to a halt if the collective lineages were immortal. In this case, expansion of a lineage would be impossible, and the collective genealogy would correspond to the parallel serial transfer classically used in experimental evolution. There are however good reasons to think that this is an unrealistic limit case. In an experimental setting, immortal lineages can be avoided by setting the survival criterion accordingly (for instance by only selecting a fixed proportion of the population, as seen in Chapter 6). More generally, trade-offs between survival and reproduction are pervasive across biology (Stearns, 1989), rendering immortal collective lineages unlikely.

### Evolutionary equilibria require a trade-off

Evolutionary singular strategies (ESS) are traits values for which the invasion fitness gradient is null. They are the equilibria of the evolutionary dynamics modelled by the canonical equation. In general, the existence of an evolutionary singular trait value depends on the functional form of the survival probability  $\sigma$ . Immortal collective lineages are the consequence of the non-existence of an “internal” evolutionary equilibrium for  $\theta$ , i.e., an equilibrium within the bound  $\sigma \in (0, 1)$ . For this reason, clarifying the conditions under which an internal ESS exists is valuable.

The invasion fitness gradient (Proposition 14, Equation 5.2) is a sum of two terms representing the relative change in  $\sigma$  and  $\rho$  due to the mutation on  $\theta$ . Since both  $\sigma$  and  $\rho$  are positive, the sign of the two terms of  $g$  depends only on the sign of the derivatives of  $\sigma$  and  $\rho$  with respect to  $\theta$  respectively. As a consequence, no evolutionary stable strategy ( $g(\theta) = 0$ ) exists within

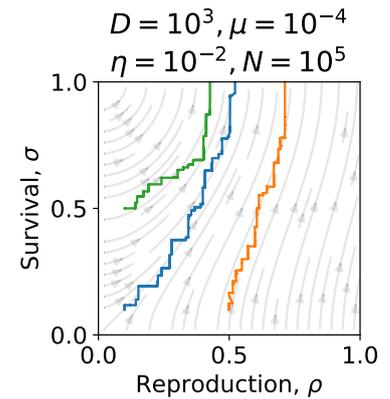


Figure 5.6: **Stochastic simulation of the model without constraints.** The model is described in Figure 5.2. Each time a collective is founded, it either inherits the life history parameters of its parent with probability  $(1 - \mu)$ , or mutates with probability  $\mu$ . Mutations affect either  $\sigma$  or  $\rho$  with equal probability, by a random normally distributed effect with zero mean and variance  $\eta$ . The trajectory of the mean  $\rho$  and  $\sigma$  on  $D$  collectives is shown for three different initial conditions and  $N$  generations. The flow induced by the canonical equation of Adaptive Dynamics (Figure 5.5) is overlaid in grey.

the bound  $\sigma \in (0, 1)$  for traits that concurrently increases survival ( $\partial_\theta \sigma > 0$ ) and reproductive output ( $\partial_\theta \rho = \alpha > 0$ ). This is also true for a trait that would concurrently decrease  $\sigma$  and  $\rho$ .

If the propagule output increases and concurrently the collective survival probability decreases around a trait value, then an internal ESS is possible. When the system is at such an internal ESS, collectives reproduce both by re-colonizing their own niche, and by producing dispersing propagules. How many propagules are produced and how much niches turn over depends on the specific functional form of the dependence of life-history parameters on the collective trait.

A simple way to model the constraints acting concurrently on collective survival and reproduction is to suppose that both life-history parameters depend on a single internal state of the collective, represented by a one-dimensional trait  $\theta$ . Suppose that the propagule output is linearly increasing with the trait value, while the survival probability is a deterministic function of the trait value:

$$\begin{cases} \rho = \theta \mapsto \alpha\theta, & \alpha > 0 \\ \sigma = \theta \mapsto \sigma(\theta) \end{cases} \quad (5.3)$$

For instance,  $\theta$  could be the density or the number of particles composing the collective. This would mean that a fixed fraction  $\alpha$  of particles become propagules. Conversely, having survival of the collective depend on its size is a reasonable assumption. For instance, a larger collective-size is known to reduce the vulnerability to predation of multicellular aggregates such as the snowflake yeast (Ratcliff et al., 2012). Alternatively, the survival probability of the bacterial mat from Chapter 4 is thought to first increase with collective size (there is a minimum number of cells required to form the cooperative mat), but saturate or even decrease above a given value (as the numerous cells impose a higher constrain on the cellulose mesh).

A sufficient condition for the existence of an internal ESS can be derived from Equation 5.2 and Equation 5.3:

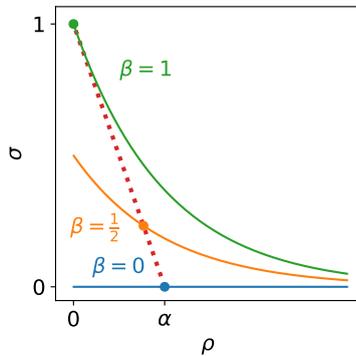


Figure 5.7: **Exponentially decaying trade-off** between survival  $\sigma(\theta) = \beta e^{-\theta}$  and reproduction  $\rho(\theta) = \alpha\theta$  for  $\alpha = 1$ . The red dotted line represent ESS  $\theta^*$  for all values of  $\beta$  (projected from figure 5.8).

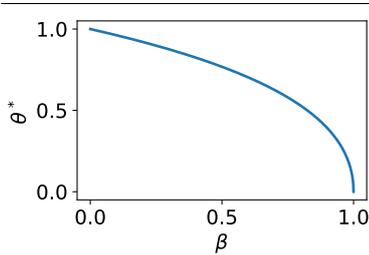


Figure 5.8: **ESS value for the exponential decay trade-off.**

#### Proposition 15 (ESS for linear reproduction):

Let  $\rho(\theta) = \alpha\theta$ ,  $\alpha > 0$  and  $\theta > 0$ .

Evolutionary singular strategies are values  $\theta^*$  of  $\theta$  where:

$$\partial_\theta \sigma(\theta^*) = -\frac{\sigma(\theta^*)(1 - \sigma(\theta^*))}{\theta^*} \quad (5.4)$$

Corollaries:

- There is no internal ESS for  $\sigma(\theta) = 1 - \beta\theta$ ,  $\beta \in \mathbb{R}^*$ .
- All positive values of  $\theta$  are ESS for  $\sigma(\theta) = \frac{1}{1+\theta}$ .
- There is a single internal ESS  $\theta^*$  for  $\sigma(\theta) = \beta e^{-\theta}$ ,  $\beta > 0$  and in this case  $\theta^* = 1 + W(-\beta e^{-1})$  with  $W(ze^z) = z$ .

(Proof page 112).

Proposition 15 precises the condition under which an ESS is possible. As stated before, change in trait value must locally have an opposite effect on survival and reproduction to qualify as ESS. Equation 5.4 shows that, if the survival probability is proportional (with coefficient  $\alpha$ ) to the trait value, the required (negative) local slope for the survival probability is independent of the proportionality coefficient  $\alpha$ .

A corollary of the condition 5.4 is that linear functional responses, cannot result in an internal ESS. Moreover, it shows that the inverse decay functional response is a pathological case where the model predicts neutral evolution. Indeed, as all trait values are evolutionary singular, the model cannot predict a favoured direction (increase or decrease in trait value) of successive mutant invasion. As a consequence, the average trait value is predicted to have the dynamics of an unbiased random walk in the trait space.

In the case of the exponential decay (Figure 5.7), there is a single internal evolutionary singular strategy, which is always stable. Numerically solving the position of the ESS shows that its corresponding trait value decreases when the maximum survival rate  $\beta$  decreases (Figure 5.8). When immortal collective are possible ( $\beta = 1$ ) the stable trait value corresponds to immortal collectives, and no investment in propagules ( $\theta = 0$ , hence  $\rho(\theta) = 0$  and  $\sigma(\theta) = 1$ ). As soon as the maximum survival rate is lower than one, some niches become available for replacement by propagules each generation, thus nonzero propagule investment is favoured. This investment increases as  $\beta$  decreases, and tends toward  $\rho(\theta^*, \beta = 0) = \alpha$ . This means that even when approaching the limit case where no collective can survive for a single generation  $\beta = 0$ , investment in reproduction does not grow unbounded.

To summarise, the existence of an evolutionary singular trait value corresponding to non-immortal collective lineages (internal ESS) requires a negative trade-off between survival probability and reproductive output. This is not a sufficient condition: if the reproductive output is proportional on the collective trait, linear trade-offs prevent internal ESS while some non-linear trade-off, such as an inverse decay, have an infinite number of internal ESSs. However, exponential decay in survival yields an internal ESS, whose value depends on the maximal survival probability. The next section explores other non-linear trade-offs between life-history parameters by restricting the functional forms to explore the effect of the trade-off curvature.

### Non linear trade-off

In order to control the non-linearity and curvature of the trade-off, suppose that survival probability and propagule output are deterministic functions of a collective state  $p(\theta) \in [0, 1]$ , that itself depends on the collective trait  $\theta$ . For example,  $p$  could be the proportion of cells of a given type. The reproductive output  $\rho$  as well as the survival probability  $\sigma$  are assumed to follow a power-law of  $p$  (or  $1 - p$ ). The use of the power-law family implies that the trade-off is scale independent with respect to  $p$ , which is a widespread pattern in evolutionary biology (Stevens, 2009). The scale constant  $d$  controls the curvature of the trade-off (Figure 5.9), allowing to explore a range of relationships between survival and reproduction.

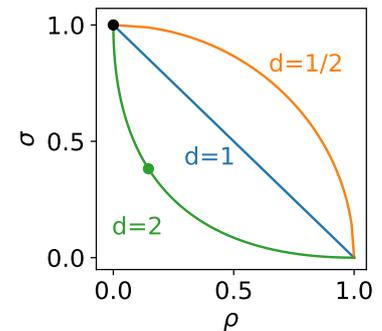


Figure 5.9: **Non linear trade-off corresponding to Equation 5.5..** The trade-off is convex if  $d \leq 1$ , and concave if  $d \geq 1$ . There is an ESS for  $d \geq 1$ , represented by disks.

**Proposition 16 (ESS in a simple non-linear trade-off):**

Let:

$$\begin{cases} \rho(\theta) = p(\theta)^d \\ \sigma(\theta) = (1 - p(\theta))^d \end{cases} \quad (5.5)$$

with  $p$  a continuous function with values in  $[0, 1]$  and  $d \in \mathbb{R}$  the non-linearity factor of the trade-off between survival and reproduction.

Then, evolutionary singular strategies are values of  $\theta$  where:

$$g(\theta) = 0 \Leftrightarrow \begin{cases} \partial_{\theta} p(\theta) = 0 \\ \text{or} \\ 1 - 2p(\theta) - (1 - p(\theta))^{d+1} = 0 \end{cases} \quad (5.6)$$

The second condition is always true for  $p(\theta) = 0$  and never verified for  $p(\theta) \neq 0$  if the trade-off is convex ( $d \leq 1$ ).

(Proof page 112.)

Proposition 16 shows that there are two conditions for a trait to be at evolutionary equilibrium (Equation 5.6). First, and maybe trivially, if local changes in the trait have not consequence for the relevant state of the collective ( $\partial_{\theta} p(\theta, T) = 0$ ), the adaptive dynamics model predicts no evolutionary change, and the trait is neutral. The second condition is always true for immortal collectives that do not disperse ( $p(\theta) = 0$ ). Additionally, this condition can be satisfied by non-immortal collectives if the trade-off has a strictly concave shape ( $d > 1$ ).

Numerically solving the condition in Equation 5.6 for  $p(\theta)$  gives the ESS displayed in Figure 5.10. This illustrates that the only stable evolutionary traits values for convex trade-off ( $d < 1$ ) are those that result in immortal collectives ( $p(\theta^*) = 0$ ). Note that immortal collectives are always an evolutionary singular point. However, if the trade-off is concave ( $d > 1$ ), this value becomes evolutionary unstable, and a new stable equilibrium appears (there is a trans-critical bifurcation in  $d = 1$ ).

To understand this difference, note that concave trade-offs are characterised by a steep slope (higher than  $-1$ ) around the point  $(\sigma, \rho) = (1, 0)$ , meaning that small mutations in the reproductive output have a comparatively larger effect on the survival probability than for a convex trade-off. Thus, near the point  $(\sigma, \rho) = (1, 0)$  where collective lineages are immortal, there are more opportunities for propagules-producing collectives (i.e.,  $\rho \neq 0$ ) to reproduce if the threshold is concave than if it is convex.

The previous results give some general intuition about the consequences of trade-offs between survival and reproductive output in this system. The functional form of the life-history parameters is however decoupled from the particle-level ecology within collectives. It is instead reasonable to expect particle dynamics to impose constraints on the trade-off exist. The next

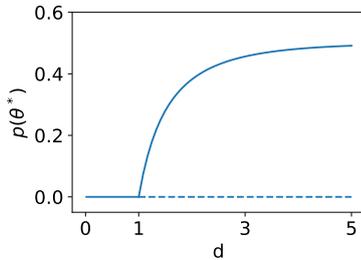


Figure 5.10: **Position of the ESS as a function of the non-linearity parameter  $d$ .** Plain lines are evolutionary stable, dotted lines are evolutionary unstable.

section is dedicated to explicitly take into account the ecological dynamics within collectives.

### 5.3 Particle dynamics

In general, in a nested model the state of a collective depends on the internal ecological dynamics of their composing particles. For instance, a bacterial mat slowly grows from a few cells colonising the air-water interface and extends from a few aggregation spots to the whole vial (Ardre et al., 2019). Hence, survival and reproduction of collectives are not constant in time: young collectives may be too small to produce a significant number of propagules while older collectives might be subject to ageing and have a lower survival probability. The underlying ecology is in turn governed by the traits of the particles. The birth rate of particles, as well as the quantity of various molecules they produce (such as cellulose in the pond scum experiment) influence their population dynamics and *in fine* the collective-level life history parameters. In the following, we present two simple models of population dynamics for the particles, and couple them to the evolutionary model presented in the previous section.

#### 5.3.1 Logistic Particle Dynamics

Consider finite-volume collectives supplied with a constant flux of resources. The density of particles at time  $t$  is noted  $n$ . The per capita birth rate of the cells is governed by the individual trait  $b$ . The death rate of the cells is proportional to their encounter rate, which is proportional to the square of particle density (without loss of generality the coefficient of proportionality is set to one). The birth-death rates are summarised in Figure 5.11. A collective is seeded by a density  $n_0$  of particles. Thus, the particle density within a collective is governed by the following well-known logistic differential equation:

$$\begin{cases} \frac{dn}{dt} = n(b - n) \\ n(0) = n_0, \end{cases} \quad (5.7)$$

where  $n_0$  is the density corresponding to one particle in the volume of the niche. Note that the population initially grows if  $n_0$  is lower than  $b$ . In the following,  $n_0$  is assumed to be always smaller than the initial birth rate  $b$ . The solution  $n(t)$  of eq. 5.7 describes the variation in time of the particle population size within a collective from its seeding to the time of reproduction  $T$ .

#### Isolated collective

Suppose that the collectives have a large lifespan. The particle population will approach its ecological equilibrium, defined as the asymptotic value of  $n$ . In this model, the asymptotic density is equal to the birth rate ( $\lim_{t \rightarrow \infty} n(t) = b$ , Figure 5.12), meaning that increasing the birth rate also increases the carrying capacity of the niche, as one would expect if the particles use their resources more efficiently.

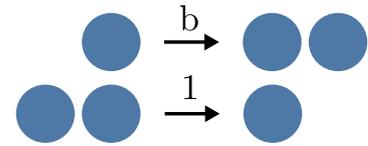


Figure 5.11: **Logistic birth-death events.** Particles reproduce at a constant birth rate  $b$ . Particles die from density dependant interactions, the rate is thus proportional to their squared density.

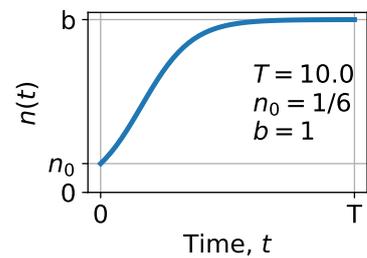


Figure 5.12: **Logistic population dynamics.** Solution of the differential equation 5.7.

The collective state is thus parametrized by the sole particle-level parameter  $b$ , whose evolutionary trajectory can be predicted by mutant invasion analysis:

**Proposition 17 (Logistic dynamics in isolation):**

Let  $b$  be the per capita birth rate of particles,  $1$  be their per-capita density-dependent death rate in a collective of lifespan  $T \gg 1$ .

Then, the invasion fitness of any mutant with an increased value of  $b$  is positive (conversely the invasion fitness of a mutation decreasing  $b$  is negative). There is no evolutionary singular strategy.

(Proof page 113).

The consequence of Proposition 17 is that the long-term evolutionary trajectory of the population is an ever-increasing value of  $b$  as illustrated in Figure 5.13. This phenomenon can be empirically observed in long term evolution experiments (Wiser et al., 2013). This arm race for ever faster growing particle might have deleterious consequences for collective-level survival: for instance, cancerous tumours emerge from this particle (cell) level selection within organisms. The next section will put this model back in the context of collective-level selection and show how the evolutionary equilibrium is modified when collectives have a finite lifespan, and their size determines the life-history parameters.

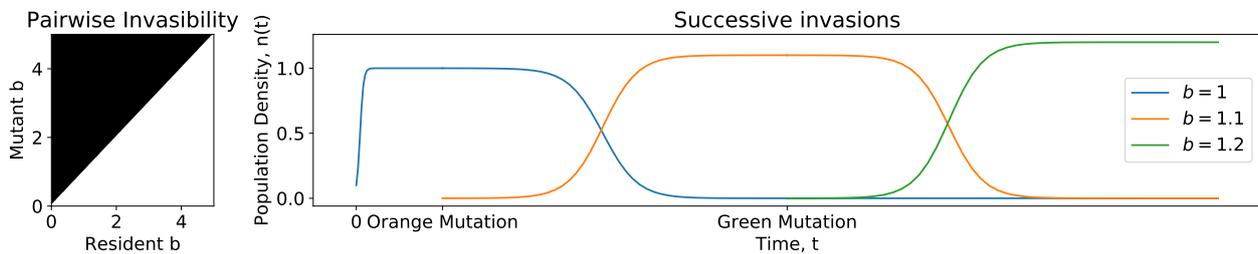


Figure 5.13: **Invasion by successive mutants in a collective of indefinite lifespan.** *Left:* Pairwise invasibility plot on  $b$ . *Right:* Population dynamics.

### Effect of collective-level selection

Logistic particle dynamics can now be used to constrain evolution at the collective level. In order to do so, we need to express the life-history parameters of collectives, survival probability and propagule output, as a function of the state of the particle population.

Consider, in analogy with the pond-scum thought experiment, that the reproductive output is increasing with the number of cells, as a fraction of

them detach from the mat and disperse in the pond. Suppose that fixed fraction ( $\alpha > 0$ ) of cells is shed by the collective. Conversely, the survival probability decreases with the particle density, as the mass of cells increases and inevitably supersedes surface tension. Since cell density is positive and potentially unbounded, let us suppose that survival decreases with particle density, and use the exponential decay trade-off from Proposition 15. The survival probability tends toward zero for infinitely large collectives, while it tends toward a maximal probability  $\beta$  for vanishingly small collectives.

A direct consequence of Proposition 15 is that there is an evolutionary stable population density, noted  $n^*$  that only depends on  $\beta$ , namely when  $\sigma(n^*) = 1 + W(-\beta e^{-1})$ . However, the population density  $n$  at the time of selection now depends on the duration of the growth phase  $T$ , the initial density  $n_0$ , but also the particle birth rate  $b$ . In summary:

$$\begin{cases} \rho(n(n_0, b, T)) &= \alpha n(n_0, b, T) \\ \sigma(n(n_0, b, T)) &= \beta e^{-n(n_0, b, T)} \end{cases} \quad (5.8)$$

With the expression of  $n(n_0, b, T)$  given by the solution of Equation 5.7:

$$n(n_0, b, T) = \frac{b}{1 - (1 - bn_0^{-1})e^{-bT}} \quad (5.9)$$

The particle evolutionary trait that we consider is the birth rate  $b$ . Numerically solving the position of the ESS by substituting  $\sigma, \rho$  (expressed as functions of  $n$  in Equation 5.8) and  $n$  (expressed as function of  $b$  in Equation 5.7) in the fitness gradient  $g$  (Equation 5.2) shows that a finite evolutionary stable value of  $b$  (noted  $b^*$ ) exists. This limits the maximum possible size of the particle population, in contrast with the potentially unbounded increase in isolation. Thus, imposing collective-level selection changes the evolutionary dynamics of the particles.

The value of  $b^*$  decreases when the duration between two selective events  $T$  increases, and tends toward a limit value as illustrated in Figure 5.14. Indeed, the intra-collective dynamics saturates when the time before reproduction grows larger, so that the collective state becomes largely independent of  $T$ .

When the maximal survival probability achievable by collectives  $\beta$  is decreased, collectives get bigger. Indeed, the evolutionary stable birth rate  $b^*$ , that is proportional to collective size, increases when  $\beta$  decreases (as pointed out in Proposition 15). As seen in Proposition 14, the effect of an investment in reproductive output  $\rho$  is weighted by the mortality probability of collectives ( $1 - \sigma$ ). As a consequence, an increase in birth rate, collective size, and thus a higher number of propagules, is selected when collective-level selection is harsher (smaller values of  $\beta$ ).

This model relates a simple particle ecological dynamics to the evolutionary fate of collectives. However, the ecology neglects the specificity of our motivating example: the existence of cells that have a propensity to contribute differently to reproduction and survival of collectives. This extension is discussed in the next section.

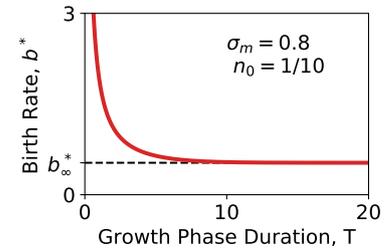


Figure 5.14: **Evolutionary stable birth rate  $b^*$ .** The evolutionary stable birth rate  $b^*$  decreases (and saturates) with the duration of the growth phase  $T$ .

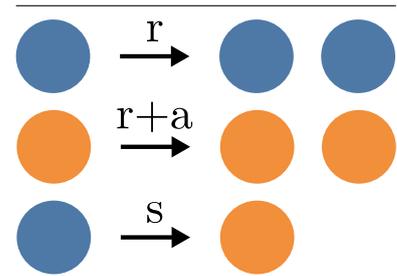


Figure 5.15: **Birth events in the germ-soma model.** Germ cells (blue) reproduce at a constant per-capita rate  $r$ , and switch to somatic cells at rate  $s$ . Somatic cells (orange) reproduce at rate  $r + a$ .

### 5.3.2 Soma-germ particle dynamics

Division of labour between particles is a hallmark of evolutionary transitions in individuality. In the pond-scum scenario presented in chapter 4, wrinkly and smooth cell types have different roles in the life cycle of the collective. Wrinkly, sticky cells contribute to the structural integrity of the mat, and form a proto-somatic line, while smooth cells disperse more easily, forming a proto-germ line. As well as for other fraternal transitions, cells can in principle change their role by mutation or epigenetic changes (contrasting with egalitarian transitions, such as to symbiosis where there is no likely mutational path from one type to the other — see Chapter 6). In the following, we introduce a simple particle-level ecological dynamics that models a two-type population with contrasting contributions to survival probability and propagule output.

#### Two-types "linear" ecology of particles

Let the collective be composed of two kinds of particles, proto-germ and proto-soma, in a pure birth process depicted in Figure 5.15. Proto-germ cells give birth at a constant rate  $r$ , while soma have a constant birth-rate offset  $a \in \mathbb{R}$  with respect to the germ cells, called the growth differential (advantage if  $a > 0$ , disadvantage if  $a < 0$ ). Additionally, germ cells produce soma cells at a rate  $s$ . The proto-germ particles are the only ones that can be transmitted to the next generation (by self-replacement of surviving collectives or via the germ bank). Survival of the collective is increasing with the proportion of soma cells, whereas their reproductive output increases with the number of germ cells.

Let  $\mathbf{x}(t) = (G, S)$  be the density of germ and soma cells in the collective at time  $t$ . Since germ cells are the only ones that are transmitted to the next generation  $\mathbf{x}(0) = (1, 0)$ . The population dynamics follows the ordinary differential equation:

$$\frac{\mathbf{x}(t)}{dt} = A \mathbf{x}(t), \quad \text{with } A = \begin{bmatrix} r - s & 0 \\ s & r + a \end{bmatrix} \quad (5.10)$$

The internal ecological dynamics of the collective can be derived from Equation 5.10. It can be equivalently expressed in terms of the dynamics of germ and soma  $(G, S)$  or in terms of total particle density  $n = G + S$  and proportion of germ-like particles  $p = \frac{G}{N}$ . This second set of coordinate is used below.

**Proposition 18 (Linear Ecology Trajectory):**

*The density of particles  $n(T, r, s, a)$  and the proportion of germ-like particles in a collective  $p(T, s, a)$  at the time of selection, expressed as functions of the duration of the growth phase  $T$ , the switch rate  $s$  and the growth differential of germs particles  $a$  are:*

$$\begin{cases} n(T, r, s, a) = \frac{s}{a+s} e^{(r+a)T} + \frac{a}{a+s} e^{(r-s)T} \\ p(T, s, a) = \frac{s+a}{s e^{(s+a)T} + a} \end{cases} \quad (5.11)$$

(Proof page 113.)

Figure 5.16 presents four qualitative ecological trajectories displayed by the germ-soma model (from Equation 5.11). Within a collective generation, the initial condition is always one unit of germ and no soma.

If the switch rate is lower than the growth rate ( $s < r$ , Figure 5.16 a, b), the density of germ increases within one generation. Otherwise ( $s > r$ , Figure 5.16 c, d), it decreases. If the density of germ decreases within one generation, the collective-level population is not viable across several generations as the number of available propagules will tend toward zero (preventing both dispersal and self-replacement). If, moreover, the soma as a growth disadvantage higher than the growth rate ( $a < -r$ ,  $s > r$ , Figure 5.16 d), the whole population (germ and soma) asymptotically goes extinct within one generation.

In the case of a viable population ( $s < r$ ), the density of germ and soma grows within one generation. If the soma has a growth advantage or a small growth disadvantage that is compensated by the production of new soma by the germ ( $a > -s$ ,  $s < r$ , Figure 5.16 a), the proportion of germ decreases with time. Conversely, if the soma has a growth disadvantage higher than the switch rate ( $a < -s$ ,  $s < r$ , Figure 5.16 b), the proportion of soma stabilises to  $1 + \frac{s}{a}$ .

Overall, this simple pure-birth ecological model hinges on two important parameters: the switch rate of the germ  $s$ , and the growth differential of the soma  $a$ . The next section lays out the respective contribution of germ and soma to the collective life history parameters.

### Effect of collective-level selection

Let us turn to the collective-level dynamics. Suppose now that life-history traits  $(\rho, \sigma)$  depend on the proportion of germ cells  $p$  and soma cells  $(1 - p)$ . Using the non-linear trade-offs for survival rate and reproductive output defined in Equation 5.5, with  $d = 2$ :

$$\begin{cases} \rho = p^2 \\ \sigma = (1 - p)^2 \end{cases}$$

As a consequence, collective lineages composed solely of soma ( $p = 0$ ) are immortal and do not produce propagules ( $\sigma = 1, \rho = 0$ ). Conversely, collective lineages composed solely of germ ( $p = 1$ ) certainly die ( $\sigma = 0$ ). As mentioned in Proposition 16 and since  $d > 1$ , there is an evolutionary attracting proportion of germs, which is around 0.2 regardless of the internal ecology of the collective. It must be stressed that this is a direct consequence of the current assumption that life-history traits depend on the proportion  $p$  and not, for instance, the population size. However, the germ-to-soma switch rate  $s$  required to achieve this proportion depends on the growth differential of the soma  $a$  and the duration of a growth phase  $T$ .

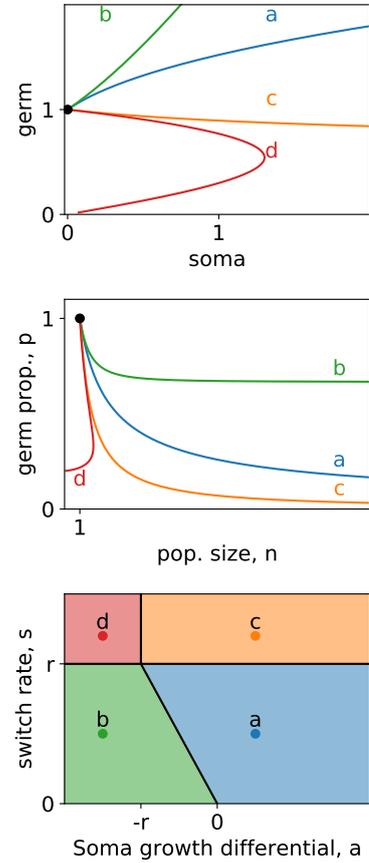


Figure 5.16: **Qualitative Regimes of the germ-soma model.** a. Vanishing germ proportion b. Asymptotic germ proportion, c. Vanishing germ. d. Extinction. Black dots are the initial conditions:  $S, G = (0, 1)$  or  $n, p = (1, 0)$ .

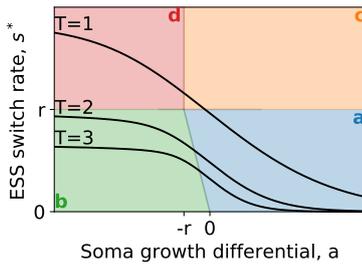


Figure 5.17: **The ESS on the switch rate  $s^*$  decreases with the growth differential of soma  $a$ .** Color code is identical to Figure 5.16.

Figure 5.17 shows the position of the evolutionary singular value of  $s$  as a function of the growth advantage of the soma  $a$ , for three values of  $T$ , numerically solved from the invasion fitness gradient in Equation 5.2, Proposition 14. Note that the ESS  $s^*$  decreases when the growth differential of the soma increases. This means that the faster the soma grows, the smaller the switch rate required to reach the optimal proportion of germ to soma. Moreover, increasing the duration of the growth phase  $T$  decreases the value of the evolutionary stable switch rate.

If the growth differential  $a$  and growth rate duration  $T$  are both small enough, the ESS of  $s$  lies in a region where the population is not viable (Figure 5.16 *c* and *d*). In this case, the model predicts an “evolutionary suicide”: successive invasions of mutants lead to a population with smaller and smaller collective size (Figure 5.16 *d*), or an absence of germ cells (Figure 5.16 *c*), leading to a probable extinction of the whole population.

Overall, this section illustrates how the dependence of life-history parameters ( $\rho$  and  $\sigma$ ) on the ecological dynamics of the particles translates conditions on the evolutionary outcome imposed by the survival-reproduction trade-off into relationships among particle-level parameters (birth rate in the logistic model, switch rate in the soma-germ model). If solving the ESS for the general trade-off (as in section 5.2) allows drawing general conclusion on the long term evolutionary fate of the system, including the internal ecology of the collectives reveals new insights, as well as new challenges. The obvious advantage is to give a more mechanistic interpretation of the survival and reproduction parameters. However, the ESS might exist in terms of  $\rho$  and  $\sigma$  in the general trade-off, but there might not be any trait value that allow to reach it, or this trait value might not be viable in the long run. Overall, not all values of collective life-history parameters and their trade-off may be attained by tuning the microscopic parameters.

In this section, the case where the collective population might not be viable arose (for instance when the number of germ cells decreases with each new generation), but the current model is ill-equipped to tackle this question: collective population size is fixed. In the next section, the problem of the viability of the collective-level population is addressed with different model that includes the collective-level population dynamics.

## 5.4 Collective Population Dynamics

When designing a time-consuming experiment such as that presented in chapter 4, great care is accorded to avoid both the premature extinction of the population and its unchecked growth. Thus, it has been decided that the population size of collective would be kept constant. Such a simplification is necessary for the sake of the experiment, but simplifies considerably the collective-population dynamics that could arise from a real niche colonization process. For instance, propagules may fail altogether to reach new niches or collectives may not reproduce fast enough. The system would therefore miss a necessary —if slightly obvious— condition for any transition in individuality by ecological scaffolding: that the collectives do not go extinct.

This section presents a modified version of the model discussed so far (Figure 5.2) to take into account a fixed population of niches (for instance the reeds in the pond scum thought experiment) that can be empty or occupied by a collective. The key parameter  $\delta$  controls the density of niches, and allow deriving conditions for the extinction of the population of collectives. Unlike the previous models, this is characterised by overlapping generations, so that some collectives survive longer than others. However, its solution relies on a separation of time scales between collective and particle dynamics.

### Niche occupancy equilibria and extinction of the collective population

Consider that the environment contains a fixed, large, number  $D$  of niches. Among those niches, a proportion  $E$  is empty, and a proportion  $N$  is occupied by collectives. At all time  $E + N = 1$ . Each collective may go extinct and disappear from the population at a rate  $\mu$ .

Each collective produces a quantity  $\rho$  of propagules, that are dispersed in the environment and form a well-mixed germ-bank. Empty niches are colonised by propagules from the germ-bank. As a consequence, each collective colonises new niches at a per-capita rate of  $\rho\delta E$  (Figure 5.18), where  $\delta$  is a parameter that controls niche density. If  $\delta$  is high, niches are close to each other and the colonisation events are more likely, whereas if  $\delta$  is low the niches are far apart and the colonisation events are rare. In the limit where  $\delta = 0$ , niches are so isolated from each other that there can be no dispersal.

As in the previous sections, let  $\theta$  be the set of adaptive traits of the resident population of collectives. To simplify notation we omit the argument " $\theta$ " when possible, and note  $\mu = \mu(\theta)$ ,  $\rho = \rho(\theta)$ . The dynamics of niche occupancy is governed by the following ordinary differential equation:

$$\begin{cases} \frac{dE}{dt} = \mu N - \delta\rho NE \\ \frac{dN}{dt} = \delta\rho NE - \mu N \end{cases} \quad (5.12)$$

Collective demography depends on life history parameters ( $\mu$  and  $\rho$ ) that in turn derive from collective traits. As long as those traits are inherited across collective generations, the proportion of niches occupied by collectives can be obtained by solving Equation 5.12 (Figure 5.19).

#### Proposition 19 (Equilibria of collective populations):

The ecological equilibrium  $(E^*, N^*)$  of the population governed by Equation 5.12 with  $E(t) + N(t) = 1$ ,  $\forall t$  and  $E(0) \in (0, 1)$  is:

$$(E^*, N^*) = \left( \frac{\mu}{\delta\rho}, 1 - \frac{\mu}{\delta\rho} \right) \quad (5.13)$$

The population goes to extinction if  $E^* \geq 1$ , i.e. if  $\mu > \delta\rho$ .

An immediate consequence of Proposition 19 is that there is a critical density of empty niches  $\delta^* = \frac{\mu}{\rho}$  under which the population of collectives

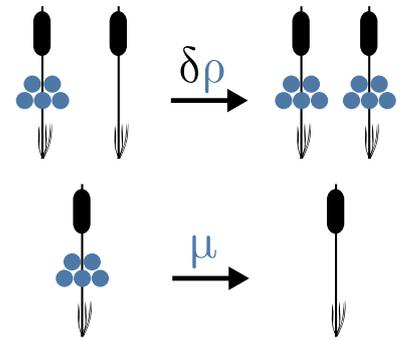


Figure 5.18: **Birth-death events.** Birth events (rate  $\delta\rho$ ) occur when cells from a collective colonise an empty niche. Death events (rate  $\mu$ ) occur when a bacterial mat collapses, resulting in the emptying of the niche.

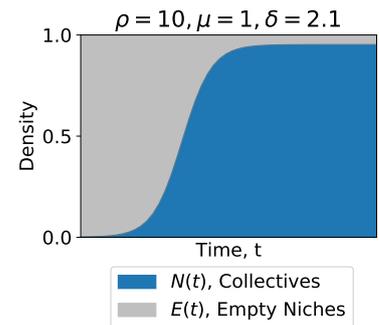


Figure 5.19: **Niches occupancy dynamics.** The proportion of occupied niches in the population tends toward the equilibrium  $1 - \frac{\mu}{\delta\rho}$

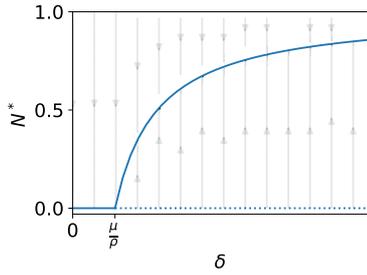


Figure 5.20: **Ecological equilibrium.** The proportion of occupied niches increases with the niche density  $\delta$ . The population goes extinct if  $\delta < \frac{\mu}{\rho}$

goes extinct (Figure 5.20).  $\frac{\mu}{\rho}$  is the death-to-reproduction ratio, meaning that increased death rate or decreased propagule production increases the minimal density threshold necessary to support the population. Since  $\mu$  and  $\rho$  are positive, for any trait values there is always a minimum value of niche density  $\delta$  below which the collective population is not viable. Conversely, it is only if collectives are immortal ( $\mu = 0$ ), not isolated ( $\delta > 0$ ) and produce propagules ( $\rho > 0$ ) that they colonise the entirety of the niches ( $N^* = 1$ ).

In other words, a first condition for a viable population of collectives in an ecological scaffolding population structure, is that the density of niches is high enough to support the population. A population of fixed reproductive output could be driven to extinction if the density of niches in the environment diminished. In addition, since the minimal density depends on the life-history parameters of the collectives, a viable population could in principle be driven to extinction if its death-to-reproduction ratio became, along an evolutionary trajectory, smaller than the critical niche density. The next paragraph applies adaptive dynamics to the evolution of a collective trait that determines life-history parameters, and aims to establish how this affects niche colonisation.

### Evolutionary dynamics of niche occupancy

Assuming that life-history parameters depend on an underlying set  $\theta \in \mathbb{R}^p$  of collective traits, the invasion fitness gradient can be derived using the same method and assumptions as in Proposition 14:

#### Proposition 20 (Invasion fitness gradient):

*The invasion fitness gradient for the trait  $\theta$  within a population whose niche dynamics is governed by Equation 5.12 is:*

$$g(\theta) = \frac{\nabla \rho(\theta)}{\rho(\theta)} \mu(\theta) - \nabla \mu(\theta) \quad (5.14)$$

(Proof page 114.)

The structure of this equation is strikingly similar to that obtained for the previously discussed model with discrete generations and fixed collective population size (Proposition 14, Equation 5.2). Note that in that model, niches were never empty (except in the limit case of vanishingly small size of particle population), as birth and death happened at the end of collective generations. All niches left empty by the death of collectives were simultaneously filled by propagules. Immortal collective lineages had a survival probability of one in the previous model ( $\sigma = 1$ ), and they have a null death rate in the current model ( $\mu = 0$ ). However, the two parameters are not exactly equivalent in the two cases, since a null survival probability in the previous model ( $\sigma = 0$ ) corresponds to the limit case of infinite death rate ( $\mu \rightarrow \infty$ ) in the current model. Therefore, the two expressions for the fitness gradient are not equivalent.

The invasion fitness gradient Equation 5.14 is a sum of two terms. The first concerns the relative change in propagule output, weighted by the mortality within the resident population. The second term is the change in collective death rate. Contrary to Equation 14, this second term is not normalised by the resident death rate: a given increase in death rate has the same effect on the gradient regardless of the current mortality—a consequence of the variable size of the collective population in this model. Finally, note that the invasion fitness gradient is independent of  $\delta$ , meaning that the evolutionary trajectory only depends on the niches via the life-history parameters of the collectives, and not directly via their density.

As in section 5.1, consider that collectives have two traits that map directly to death rate and reproductive output. That is,  $\theta = (\theta_1, \theta_2)$ ,  $\mu(\theta) = \theta_1$  and  $\rho(\theta) = \theta_2$ . In this case survival and reproduction evolve independently of one another without trade-offs. Figure 5.21 shows the evolutionary flow resulting from Proposition 20 in this situation. As expected from the similarities in equations, the structure of the evolutionary flow is akin to that in the discrete generation model (Figure 5.5). Both propagule output and survival increase until immortal lineages invade the population ( $\mu = 0$ ), which is a *de facto* endpoint for the eco-evolutionary dynamics. Super-imposing to the flow the area where the collective population is viable (black in Figure 5.21, derived from Equation 5.13), we can see that some trait regions explored by evolution in the absence of collective demography should in fact be excluded, because the collective population can not be sustained. The structure of the flow moreover prevents the so-called evolutionary suicide: the evolutionary trajectory only points out of the extinction area. Thus, there is no path of successive invasions that leads to the extinction of the population.

Just as discussed in the previous sections, the assumption that survival and reproductive output can evolve independently is unrealistic. The next paragraph explores a simple functional trade-off between those two life history parameters.

### Simple trade-off and population dynamics

As discussed in section 5.3, here we specify the particle-level ecology that shapes the collective state, and link both life-history parameters to such collective state, in such a way that a trade-off exists between survival and reproduction. Now, though, collective generations are asynchronous and collectives of different age coexist, hence the state of the collective population at a given time is distributed. Such collective phenotypic diversity can however be neglected if we suppose that the particle ecological dynamics within collectives is fast compared to collective demographic events. In this case, the state of most collectives will be close to the ecological equilibrium, that is parametrized by particle-level parameters. Functions of particle traits that provide the rates of propagule production  $\rho(\theta)$  and collective death  $\mu(\theta)$  can thus be inverted, so as to find what particle ecology corresponds to a collective ESS.

Let us illustrate this concept by using the logistic model, Equation 5.7. The evolvable trait is the maximum birth rate of particles  $b$ , and the ecological equilibrium is  $n^* = b$ . The collective life-history parameters must be

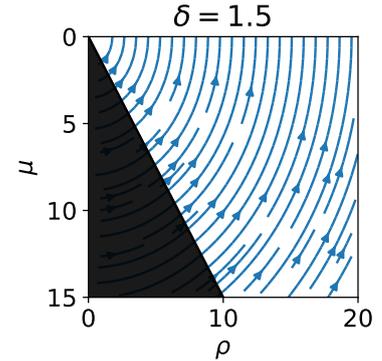


Figure 5.21: **Evolutionary trajectories** when the death rate  $\mu$  and propagule production  $\rho$  evolve independently. The  $\mu$  axis is inverted to keep the convention from the previous section that immortal collectives are at the top. The black area represents a region where the population of collectives goes to extinction.

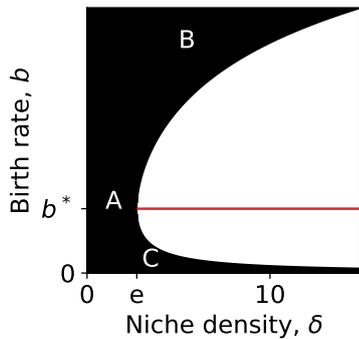


Figure 5.22: **Evolutionary stable birth rate as a function of niche density** when death rate is exponential  $\mu = e^b$  and propagule production is linear  $\rho = b$ . The ESS (red line, solution of Equation 5.14) is constant for all values of  $\delta$ . The black area represents a region where the population of collectives goes extinct (Proposition 19). *A*: No collective population is viable when the density of niches is lower than  $e$ . *B*: large collectives have a high  $\mu$  and are not viable. *C*: small collectives have a low  $\rho$  and are not viable.

computed at this equilibrium. Consider a simple exponential trade-off akin to that used in Proposition 16: production of propagules is proportional to the density of particles  $\rho(b) = n(b) = b$  and collective death rate increases exponentially with the density of particles  $\mu(b) = e^n(b) = e^b$ . Replacing these in the formula for the invasion fitness Equation 5.14 and solving the resulting expression ( $g(b) = e^b (\frac{1}{b} - 1) = 0$ ), one obtains that the evolutionary stable birth rate  $b^*$  is equal to 1.

Figure 5.22 illustrates the ESS for various niche density  $\delta$ , with superimposition of the non-extinction condition (Proposition 19). If the niche density is lower than a threshold ( $\delta < e$ ), the collective population goes extinct for any value of  $b$  (Figure 5.22 A). Above this threshold, there is a range of values of  $b$  for which the population is viable, this range gets wider as the density of niches  $\delta$  increases. If the collectives are too large (large  $b$ , Figure 5.22 B) their death rate is too high, and the population goes extinct. Conversely, if the collectives are too small (small  $b$ , Figure 5.22 C) they do not produce enough propagule to compensate the death and the population goes extinct. In the white region where collective populations persist, the ESS value  $b = 1$  is constant with respect to the niche density  $\delta$ .

As in section 5.3, the death-reproduction trade-off imposes an optimal value for life history parameters, although this value might not be reachable by the population if it lies in a region of the trait space where the population goes extinct. What is adjusted as  $\delta$  varies is not collective size, but the occupancy of the niches: if the niches are denser, they also have a higher occupancy at evolutionary equilibrium (Equation 5.13, with  $\frac{\mu}{\rho}$  constant). This is only possible in this model (and not in the model of section 5.3) because the collective population size is allowed to vary.

Overall, by adding an explicit collective population dynamics, this model distances itself farther from the experimental setup. This allows deriving necessary conditions for ecological scaffolding in a more general scenario. This paragraph showed that if the “scaffolding” structure of niches is too sparse, the population of collectives cannot be maintained. The next paragraph explores a consequence of a high density of niches: secondary colonisation events on an occupied niche.

## 5.5 Migration between collectives

At the centre of the debate on the existence of collectives as units of evolution on their own right, lays the question of the individuality of collectives. A widespread argument against the relevance of collective-level Darwinian processes is that, in nature, groups are not isolated enough to be treated as units of evolution (Williams, 2018; Maynard Smith, 1987). This difficulty is sidestepped in the growing body of work on ecological scaffolding by imposing single-cell bottlenecks and niche isolation. Clonal growth ensures that there is no mixing occurring during the lifetime of a collective, so that collectives have been treated as units of evolution both in experiments (Hammerschmidt et al. (2014) and Chapter 4) and model design (Black et al. (2019) and previous sections of the present chapter). As a consequence, collective lineages remain composed of a single type of particles.

There are two ways of relaxing this assumption. First, considering that collective birth events involve multiple particles coming from a single (as in Chapter 6) or several parental collectives. Second, considering that, even though colonisation of empty niches happens through dispersal of a single cell, established collectives can be invaded by propagules from other collectives. This paragraph explores the second option.

**Ecological model**

Again, let us focus on the collective demography given a large number of possible niches. Niches can be occupied either by resident or mutant collectives, that are characterised by different reproductive outputs and mortality rates. However, now collectives also vary in their ability to take over a niche already colonised by a collective of different type (Figure 5.23). As previously assumed, the probability of successfully colonising an empty niche is taken proportional to the reproductive output via the niche density parameter  $\delta$ . The ability of a propagule to invade and replace another extant collective is measured by an additional collective parameter  $\gamma$ . This invasion rate is expressed as a function of the trait  $\theta$  of the resident and  $\theta'$  of the mutant, that separately control life-history parameters of the resident and mutant collectives, respectively. Contrary to the previous model, collective demography needs to take into account both resident and mutant populations, since substitution events involve pairs of mutant and resident. Let  $E$  be the proportion of empty niches,  $N$  the proportion of niches occupied by collectives with resident traits  $\theta$  (and derived rates  $\sigma(\theta)$ ,  $\rho(\theta)$  and  $\gamma(\theta, \theta')$ ) and  $M = 1 - N - E$  the proportion of niches occupied by collectives with mutant traits  $\theta'$  (and derived rates  $\sigma'(\theta')$ ,  $\rho'(\theta')$  and  $\gamma'(\theta, \theta')$ ). The collective population dynamics is now described by the following differential equations:

$$\begin{cases} \frac{dE}{dt} = \mu N + \mu' M - \delta E(\rho N + \rho' M) \\ \frac{dN}{dt} = -\mu N + \delta \rho N E + \delta(\rho \gamma - \rho' \gamma') N M \\ \frac{dM}{dt} = -\mu' M + \delta \rho' M E + \delta(\rho' \gamma' - \rho \gamma) N M \end{cases} \quad (5.15)$$

First, consider the ecological equilibrium of such a system. Proposition 21 shows that adding migration events does not change the condition for extinction of the whole population.

**Proposition 21 (Extinction conditions):**

*The population governed by Equation 5.12 is not extinct,  $\forall t$ , if:*

$$\delta < \frac{\mu}{\rho} \text{ or } \delta < \frac{\mu'}{\rho'}$$

*(Proof page 114.)*

This indicates that a population cannot be rescued by addition of the capacity to colonise niches that are already occupied.

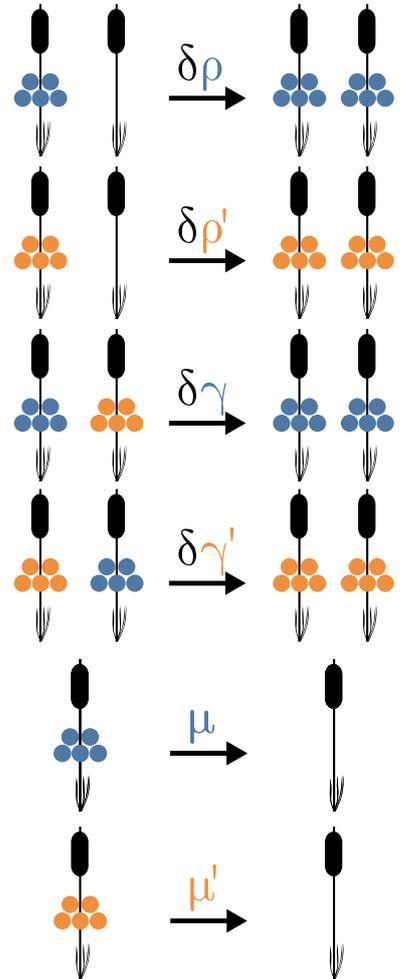


Figure 5.23: **Birth-death events** in a collective population with two types (resident, blue and mutant, orange). In order to compute the rates of change of the respective proportions, it is necessary to multiply the rates illustrated here by the probability of encountering an empty niche (top two cases) or of producing a propagule (third and fourth case).

### Evolution of the collective trait

The invasion fitness gradient of this model can be derived using the same assumptions as in Proposition 14 and Proposition 20:

**Proposition 22 (Invasion fitness gradient):**

*The invasion fitness gradient for trait  $\theta$  in the population governed by Equation 5.15 is:*

$$g(\theta) = \frac{\nabla \rho}{\rho} [\mu - \gamma(\delta\rho - \mu)] - \nabla \gamma(\delta\rho - \mu) - \nabla \mu'$$

*(Proof page 114.)*

The new invasion fitness gradient contains two additional terms in  $\gamma$  compared to that without secondary invasion (Proposition 20, Equation 5.14). First, there is a term proportional to the increase in  $\gamma$ , that is weighted by  $(\delta\rho - \mu)$ . This coefficient is positive whenever the collective population does not go extinct. Then, there is an additional contribution to the coefficient multiplying the relative increases in propagule production. Even if collective lineages are immortal ( $\mu = 0$ ), propagule production is now under selection because propagules can subvert established collectives.

To proceed, this model should be studied in the context of a three-way trade-off between death rate  $\mu$ , propagule production  $\rho$ , and the ability to perform secondary invasion  $\gamma$ . For instance, propagules that are able to invade existing mats more easily might have a higher death rate. In the case of the pond-scum scenario, one can imagine a collective with a decreased switch rate from germ-like to soma-like cells, that could easily take over existing mats (high  $\gamma$ ), but would produce less resilient mats when alone (high  $\mu$ ). The parallel with the evolution of virulence is striking (see Discussion), suggesting a path to recruit theoretical and conceptual tools for future inquiries.

## 5.6 Discussion

The present chapter aimed at building a theoretical understanding of the long-term evolution within collective populations such as the experimental collective selection regime (Hammerschmidt et al., 2014) or the more general pond scum experiment (Rainey and Kerr, 2010). This was achieved by building deterministic models of the collective ecological dynamics and study them within the framework of adaptive dynamics. Collectives were characterised by two life history parameters: survival, and propagule production. The following assumptions were made: the population of collectives is large, mutations are rare enough so the ecological and evolutionary timescales can be readily separated (evolution is modelled as a succession of invasion by initially rare mutants). The mutations are small enough in their phenotypic effects that we only consider continuous evolutionary trajectories (the only evolutionary

paths considered are of small incremental changes). Under these conditions, several conclusions can be drawn.

First, if we consider a fixed-size population of collectives, as in experimental biology, natural selection favours collectives of increasing survival probability and reproductive output. If there is no underlying physical or ecological constraint, the population is inevitably invaded by immortal collective lineages. Immortal collectives do not provide the opportunity for collective birth-death anymore, thus selection on other traits (such as propagule production) grinds to a halt (and might even be counteracted by intra-collective selection). More precisely, the advantage of a mutation on propagule production (measured by its invasion fitness) is weighted by the current mortality rate of the population. As a consequence, when designing experimental protocols to select for propagule production one should keep the mortality rate as high as possible, while allowing the population to not go extinct with a reasonable chance, and in any way avoid having immortal collective lineages.

In order to observe evolutionarily stable phenotypes that are not immortal, it is necessary to take into account trade-offs between survival and reproduction. More precisely, the trade-off between survival probability and propagule production must be negative and non-linear. Two examples of non-linear trade-offs were considered. In the case of exponential decay of the survival as a function of the propagule output, an evolutionarily stable phenotype exists, moreover it is lower than the maximal survival probability allowed by the trade-off (except if the maximal probability is one). In the case of the power-law trade-off, a non-immortal evolutionary stable phenotype exists if the trade-off is concave. In this situation, immortal collectives are predicted to be evolutionary unstable: they cannot be reached by a succession of selective sweeps. The concavity of the trade-off means that when collectives lineages are close to immortal, the production of a small quantity of propagule has (relatively) large effects on the survival probability. The curvature of this kind of trade-off has already been extensively studied in the context of the evolution of multicellularity. Michod (2007) predicts that their shape should be concave in the early stage of the transition —such as our model.

Those geometric constraints are independent of the underlying ecology of the particles: the collective process imposes the evolutionary outcome in terms of survival probability and propagule production. However, mutations on collective-level life history parameter are ultimately the result of mutations on the particle traits. Studying the evolutionary dynamics of particle traits in the context of ecological scaffolding requires to explicitly model their ecology. Indeed, collectives are dynamical objects. If the optimal collective life history parameters values are known and depend only on the survival-reproduction trade-off, the path to reach them can be long and windy (or even non-existent) depending on the behaviour of the particles. To illustrate this, two ecological models were presented and discussed. First, in a simple logistic ecology where all particles were identical the evolutionary stable birth rate is dependent on the duration of the generations. The trade-off ensures the existence of a stable evolutionary collective size, and other parameters constrain the actual particle trait value. Second, in a germ-soma ecology where there exist two phenotypically distinguished particle types, an

evolutionary stable proportion of germ might exists—even though it can lie in an unreachable region of parameter space. Depending on other ecological traits of the particle population such as the growth advantage of the soma, coexistence between the two phenotype may or may not evolve.

Finally, the model was modified to relax the assumption, required by experimental work, that the population size of collectives was fixed and that the collective generations were synchronised. The new model has a structure remarkably close to the previous one. This allowed to characterise the conditions upon which one might hope to observe evolution by ecological scaffolding.

The first condition was of course that the collective population does not go to extinction. The model pointed out that there is a lower limit in the density of niches under which the population would go extinct. This threshold generally depends on the life history parameter values. However, when neglecting collective-to-collective migrations, the evolutionary stable trait value is independent on the density of the niches provided that the population is viable. Conversely, for any fixed trait-value, there is a density below which the population is not viable. The model was applied to a situation where particle ecology was logistic ecology.

The second obstacle to ecological scaffolding studied here was the existence of identity-muddying migrations events between collectives. Migrations were modelled as single cell events, but migrations to existing collectives resulting in a replacement of one collective by another were allowed. As a consequence, evolutionary equilibria may be modified.

It turns out that the last two models have a structure very close to classical epidemiological models of susceptible and infected individuals. Empty niches can be seen as susceptible individuals, while colonised niches are infected. Collective birth is similar to the inoculation of a new host from pathogens living within an already infected host, collective death is similar to the recovery or death of the host. The migration between collective is treated in the same way as superinfection (Nowak and May (1994), see Gandon et al. (2001) for an adaptive dynamics treatment). This similarity has both practical and conceptual consequences. First, drawing rigorous theoretical links between the two domains should be possible, and as a consequence there is hope to make rapid progress by applying results from the vast body of work in epidemiology to the question of transition in individuality. More conceptually, this illustrates how infectious diseases may be subject to ecological scaffolding. Pathogens are undeniably selected within their host, but also on the basis of survival and reproduction of the infection among individuals. For instance, HIV infections are genetically diverse, even if they typically are founded by a single genotype. Diversification process leads to lineages that are specific to tissues of the host, and some genotypes are more likely to be transmitted than others (Joseph et al., 2015). It is tempting to treat infectious diseases as loose multicellular organisms in light of these observations. Now, diseases are usually not considered as individuals on their own, and they certainly do not present the same level of integration than multicellular metazoan for instance. This begs the question whether the conditions for further integration via ecological scaffolding are not met, or defeated by opposing physical or evolutionary forces. Multicellular pathogens do exist,

but to my knowledge, their transition to multicellularity is not the product of ecological scaffolding within their current host.

Approximations introduced here for modelling purpose neglect some of the richness of multi-level eco-evolutionary dynamics.

The adaptive dynamics limit of evolutionary process only informs us about possible evolutionary paths that integrate successive invasions of small-effect mutants whereas experiments show (in Chapter 4) that multiple mutations coexists simultaneously within the population.

A farther reaching consequence of focusing on long-term evolution comes into light when assuming single-cell bottleneck as well. Indeed, in the models from this chapter, any small mutant in the particle trait can always appear and be selected to found a new collective. As a consequence, any direction in the particle trait space may be explored (within the limits of physics and genetics). Any direction that can be explored will be explored because there is a small chance that the particle mutation occurs just before the end of the generation, and an even smaller chance (but strictly positive) that the mutant particle would be selected as the propagule for the next generation. This relies on the fact that drift will supersede selection within collectives: it is the central assumption behind the shifting balance theory (Wright, 1948). However, if the single-cell bottleneck ensures that no mutant trait value may be excluded, the relative time to reach it will vary greatly if the mutant is able or not to invade the underlying particle population. Overall, it means that the canonical equation of adaptive dynamics would yield erroneous results in terms of the speed of transient dynamics. The jump rate from one collective trait value to the next would require modelling of the underlying events using results from stochastic approaches (as in Chapter 3 or in Wahl and Gerrish (2001)) to correct for this effect. A good method to build intuition on this problem would be to perform stochastic simulations at the particle level, as presented in Chapter 6.

A burning question that is left unanswered is the origin of the division of labour between particles. The germ-soma model that was discussed in this chapter shows an evolutionary path promoting an increasing production of somatic cells by the germ line even starting from switching rates that are close to zero. However, this models makes a really strong assumption, namely that two different soma and germ phenotype exist a priori. An alternative model of this phenomenon could suppose the preexistence of a phenotypic switch that would be effectively neutral with respect to the investment in survival and reproduction of both types. One could then focus on the evolution of a trait that would govern their actual differentiation. In this hypothesis, the transition from germ to soma may be purely phenotypical, or rely on an easily reversible mutation (as seen in Chapter 4).

This chapter gives some insight in what we might expect to be general properties of ecological scaffolding process, such has the effect of the geometry in the survival trade-off, or the conditions on the scaffolding structure. In addition, models point out where the particle eco-evolutionary dynamics – that is a characteristic of the biological system at hand – might constrain the collective eco-evolutionary dynamics, and in return be driven by it.

Overall, this part tentatively lifts the curtain on the evolutionary dynamics of nested Darwinian populations and how collective-level selection can, under

some conditions, promote the emergence of collective level-adaptation. So far, the only collectives considered were mostly genetically clonal (thanks to single cell bottlenecks). In the last part of this manuscript we will relax this assumption. Having several types of particles within the same collective for any meaningful amount of time raises an immediate question: what kind of evolutionary mechanism can maintain the partners within the collective? The final part of this manuscript is dedicated to this problem of collective-level heredity.

## Appendix: Proofs

### Proof (Proposition 13 - Invasion fitness):

Consider a monomorphic population of  $D$  collectives bearing trait  $\theta$ , with  $k$  mutants collectives bearing trait  $\theta'$ .

Each collective generations, there are  $B$  free spots:

$$B = \underbrace{(D - k)(1 - \sigma(\theta))}_{\text{Resident death}} + \underbrace{k(1 - \sigma(\theta'))}_{\text{Mutant death}}$$

The relative weight in the gamete pool of a phenotype  $i$  represented by  $n_i$  collectives is  $\frac{n_i \sigma_i \rho_i}{\sum_j n_j \sigma_j \rho_j} = \frac{n_i \sigma_i \rho_i}{W}$ .

In this case:  $W = (D - k)\rho(\theta)\sigma(\theta) + k\rho(\theta')\sigma(\theta')$ .

After one collective generation, there will be  $k'$  collectives with trait  $\theta'$ .

$$\begin{aligned} k' &= \sigma(\theta')k + \frac{k\sigma(\theta')\rho(\theta')}{W}B \\ &= k \left[ \sigma(\theta') + \frac{\sigma(\theta')\rho(\theta')}{W}B \right] \\ &= k\sigma(\theta') \left[ 1 + \rho(\theta')\frac{B}{W} \right] \end{aligned}$$

Factoring in the hypothesis of adaptive dynamics, we consider that the mutant is rare  $k \approx 0$  and the resident is at equilibrium  $D - k \approx D$ .

Hence,  $B \approx D(1 - \sigma(\theta))$  and  $W \approx D\rho(\theta)\sigma(\theta)$ .

We get the following invasion fitness:

$$f(s, s') = \ln \left( \frac{k'}{k} \right) = \ln \left( \sigma(\theta') \left[ 1 + \rho(\theta') \frac{1 - \sigma(\theta)}{\sigma(\theta)\rho(\theta)} \right] \right)$$


---

### Proof (Proposition 14 - Invasion fitness gradient):

The proof is presented for a scalar trait  $\theta$ , but can easily be generalised to a vector-trait. Let us define  $h$  such that  $f(\theta', \theta) = \ln(h(\theta', \theta))$ .

$$\begin{aligned} g(\theta) &= \left. \frac{d}{d\theta'} f(\theta', \theta) \right|_{\theta'=\theta} = \left. \frac{\frac{d}{d\theta'} h(\theta', \theta)}{h(\theta', \theta)} \right|_{\theta'=\theta} \\ &= \left. \frac{d}{d\theta'} h(\theta', \theta) \right|_{\theta'=\theta} \quad (h(\theta, \theta) = 1) \\ &= \left. \frac{d}{d\theta'} \sigma(\theta') \right|_{\theta'=\theta} + \left. \frac{d}{d\theta'} \sigma(\theta') \rho(\theta') \right|_{\theta'=\theta} \frac{1 - \sigma(\theta)}{\rho(\theta)\sigma(\theta)} \\ &= \sigma'(\theta) + [\sigma'(\theta)\rho(\theta) + \rho'(\theta)\sigma(\theta)] \frac{1 - \sigma(\theta)}{\rho(\theta)\sigma(\theta)} \\ &= \frac{\sigma'(\theta)}{\sigma(\theta)} + \frac{\rho'(\theta)}{\rho(\theta)}(1 - \sigma(\theta)) \end{aligned}$$


---

**Proof (Proposition 15 - Linear reproduction):**

Let  $\rho(\theta) = \alpha\theta$  with  $\alpha \in \mathbb{R}_+^*$ . Since  $\mu$  has to be positive to be meaningful, the trait  $\theta$  is necessary positive.

Let  $\theta^*$  be an evolutionary singular strategy. By definition,  $g(\theta^*) = 0$ .

$$g(\theta^*) = 0 \Leftrightarrow \frac{d\sigma(\theta^*)}{d\theta} \frac{1}{\sigma(\theta)} + \frac{\alpha}{\alpha\theta} (1 - \sigma(\theta)) \quad (5.16)$$

$$\Leftrightarrow \frac{d\sigma(\theta^*)}{d\theta} = -\frac{\sigma(\theta^*)(1 - \sigma(\theta^*))}{\theta^*} \quad (5.17)$$

Since  $\sigma \in [0, 1]$  and  $\theta > 0$ , the slope of  $\sigma$  in the ESS  $\partial_\theta \sigma(\theta^*)$  is negative.

**Linear.** Suppose that  $\sigma = \theta \mapsto 1 - \beta\theta$ . For  $\theta^*$  to exist,  $\beta$  must be positive. Then  $\sigma' = \theta \mapsto -\beta$ . Then Equation 5.17 becomes:

$$-\beta = -\frac{\beta\theta^*(1 - \beta\theta^*)}{\theta^*} \Leftrightarrow \beta\theta^* = 0 \quad (5.18)$$

Which corresponds to  $\theta = 0$ , since  $\beta > 0$ . This corresponds to immortal collective lineages, and thus is not an internal ESS.

**Inverse.** Suppose that  $\sigma = \theta \mapsto \frac{1}{1+\theta}$ . Then, for all values of  $\theta > 0$ ,  $\sigma \in [0, 1]$  and  $\sigma'(\theta) = -(1+\theta)^{-2} < 0$ . Then Equation 5.17 becomes:

$$-\frac{1}{(\theta+1)^2} = -\frac{1}{\theta} \frac{1}{1+\theta} \left(1 - \frac{1}{1+\theta}\right) \Leftrightarrow 1 = 1 \quad (5.19)$$

Meaning that all values of  $\theta > 0$  are ESS.

**Exponential.** Suppose that  $\sigma = \theta \mapsto \beta e^{-\theta}$ ,  $0 < \beta \leq 1$ . Then, for all value of  $\theta > 0$ ,  $\sigma \in [0, 1]$  and  $\sigma'(\theta) = -(1+\theta)^{-2} < 0$ . Then Equation 5.17 becomes:

$$-\beta e^{-\theta} = -\frac{\beta e^{-\theta}(1 - \beta e^{-\theta})}{\theta} \quad (5.20)$$

$$\theta + \beta e^{-\theta} - 1 = 0\theta = W(-\beta e^{-1}) + 1 \quad (5.21)$$

Where the Lambert  $W$  Function is such that  $W(ze^z) = z$ .

**Proof (Proposition 16 - Non linear trade-off):**

Let  $p$  be a continuously derivable function of  $\theta$ , with values in  $[0, 1]$ . Let  $\rho(\theta) = p(\theta)^d$  and  $\sigma(\theta) = (1 - p(\theta))^d$ . With  $d > 0$ . Then  $\rho'(\theta) = dp(\theta)^{d-1}p'(\theta)$  and  $\sigma'(\theta) = -d(1 - p(\theta))^{d-1}p'(\theta)$  and:

$$\begin{aligned} g(\theta) &= \frac{dp(\theta)^{d-1}p'(\theta)}{p(\theta)^d} [1 - (1 - p(\theta))^d] - \frac{-d(1 - p(\theta))^{d-1}p'(\theta)}{(1 - p(\theta))^d} \\ &= d\frac{p'(\theta)}{p(\theta)} [1 - (1 - p(\theta))^d] - d\frac{p'(\theta)}{1 - p(\theta)} \\ &= dp'(\theta) \left[ \frac{1 - (1 - p(\theta))^d}{p(\theta)} - \frac{1}{(1 - p(\theta))} \right] \\ &= dp'(\theta) \frac{1 - 2p(\theta) - (1 - p(\theta))^{d+1}}{(1 - p(\theta))p(\theta)} \end{aligned}$$

Thus, if there is  $\theta^*$  such that  $g(\theta^*) = 0$  then:

$$\begin{cases} \partial_{\theta}^* p(\theta^*) = 0 \\ 1 - 2p(\theta^*) - (1 - p(\theta^*))^{d+1} = 0 \end{cases} \quad \text{or}$$

Let  $f = x \mapsto 1 - 2x - (1 - x)^{d+1}$  with  $x \in [0, 1]$ . Note that  $f(0) = 0$  and  $f(1) = -1$ . Then  $f'(x) = d + 1(1 - x)^d - 2$

$$\begin{aligned} 0 &< x < 1 \\ 0 &< (1 - x) < 1 \\ 0 &< (1 - x)^d < 1 && \text{because } d > 0 \\ 0 &< (d + 1)(1 - x)^d < d + 1 \\ -2 &< f'(x) < d - 1 \end{aligned}$$

Hence, if  $d < 1$ ,  $f'(x) < 0$  for all values of  $x \in [0, 1]$ . Thus, if  $d < 1$ ,  $f$  is decreasing on  $[0, 1]$ . Since  $f(0) = 0$ , if  $d < 1$ , then  $\forall x \in [0, 1]$ ,  $f(x) < 0$ , meaning that there is no  $\theta^*$  such that  $g(\theta^*) = 0$ .

---

**Proof (Proposition 17 - Logistic dynamics):**

Introducing a mutant  $m$  with birth rate  $b'$ :

$$\begin{cases} \frac{dn}{dt} = n(b - n - m) \\ \frac{dm}{dt} = m(b' - n - m) \end{cases} \quad (5.22)$$

$$\frac{1}{m} \frac{dm}{dt} = b' - n - m \quad (5.23)$$

When the resident is at equilibrium ( $n = b$ ) and the mutant rare ( $m \approx 0$ ), the invasion fitness is:

$$f(b, b') = b' - b \quad (5.24)$$


---

**Proof (Proposition 18 - Proportion of Germ Cells):**

We can diagonalise  $A$ :

$$A = P^{-1}DP = \begin{bmatrix} -\frac{1}{\frac{a}{s}+1} & 0 \\ \frac{1}{\frac{a}{s}+1} & 1 \end{bmatrix} \begin{bmatrix} r - s & 0 \\ 0 & r + a \end{bmatrix} \begin{bmatrix} -1 - \frac{a}{s} & 0 \\ 1 & 1 \end{bmatrix}$$

Thus, since  $x_0 = (1, 0)$ :

$$x(t, s, a) = P^{-1}e^{Dt}Px_0 = \begin{bmatrix} e^{t(r-s)} \\ \frac{s(e^{t(a+r)} - e^{t(r-s)})}{a+s} \end{bmatrix}$$

Let  $p(t, s, a)$  be the proportion of germ line cells after a time  $t$ .

$$p(t, s, a) = \frac{\langle [1, 0], x(t, s, a) \rangle}{\langle [1, 1], x(t, s, a) \rangle}$$

**Proof (Proposition 20 - Invasion Fitness gradient with overlapping generations):**

Introducing a mutant  $M$  with traits  $s'$ , we note  $\rho' = \rho(s')$ ,  $\mu' = \mu(s')$ .

$$\begin{cases} \frac{dE}{dt} = B - \delta E(\rho N + \rho' M) \\ \frac{dN}{dt} = -\mu N + \delta \rho N E \\ \frac{dM}{dt} = -\mu' M + \delta \rho' M E \end{cases} \quad (5.25)$$

Now we consider a constant niche population size such that  $B = \mu N + \mu' M$  and  $N + M + E = 1$ . The Invasion Fitness of the mutant is thus:

$$\frac{1}{M} \frac{dM}{dt} = \delta \rho' E - \mu' \quad (5.26)$$

If the mutant is rare in a monomorphic population at equilibrium,  $(E, N, M) = (E^*, N^*, 0) = (\frac{\mu}{\delta \rho}, 1 - \frac{\mu}{\delta \rho}, 0)$ , thus:

$$f(\theta', \theta) = \frac{1}{M} \frac{dM}{dt} \Big|_{E=E^*} = \delta \rho' \frac{\mu}{\delta \rho} - \mu' \quad (5.27)$$

$$= \frac{\rho'}{\rho} \mu - \mu' \quad (5.28)$$

Fitness gradient:

We note  $\nabla \rho = \nabla \rho(\theta)$

$$g(\theta) = \frac{\partial f(\theta', \theta)}{\partial \theta'} \Big|_{\theta'=\theta} = \frac{\nabla \rho}{\rho} \mu - \nabla \mu' \quad (5.29)$$

**Proof (Proposition 21 - Extinction condition with secondary invasions):**

The Jacobian of the two-dimensional reduction of Equation 5.15 is:

$$\begin{bmatrix} M(\delta \gamma + \rho) - 2N\delta \rho + \delta \rho - \mu & N(\delta \gamma + \rho) \\ M(\delta \gamma' + \rho') & -2M\delta \rho' + N(\delta \gamma' + \rho') + \delta \rho' - \mu' \end{bmatrix} \quad (5.30)$$

The population is extinct if the equilibrium  $e_1 = (1, 0, 0)$  is stable. The Jacobian in  $e_1$  is:

$$\begin{bmatrix} \delta \rho - \mu & 0 \\ 0 & \delta \rho' - \mu' \end{bmatrix} \quad (5.31)$$

Whose eigenvalues are  $\delta \rho - \mu$  and  $\delta \rho' - \mu'$ . Thus,  $e_1$  is stable if both  $\delta \rho < \mu$  and  $\delta \rho' < \mu'$ .

**Proof (Proposition 22 - Invasion fitness gradient with secondary invasions):**

Let us take a constant population size i.e  $B = \mu N + \mu' M$  and  $E + N + M = 1$ .

Let us take a rare mutant in a monomorphic population at equilibrium,  $(E, N, M) = (E^*, N^*, 0) = (\frac{\mu}{\delta \rho}, 1 - \frac{\mu}{\delta \rho}, 0)$ .

Thus:

$$f(\theta', \theta) = \frac{1}{M} \frac{dM}{dt} = \rho' \frac{\mu}{\rho} - \mu' + (\rho' \gamma' - \rho \gamma) \left( \delta - \frac{\mu}{\rho} \right) \quad (5.32)$$

$$= \rho' \left[ \frac{\mu}{\rho} - \gamma' \left( \delta - \frac{\mu}{\rho} \right) \right] - \mu' - \rho \gamma \left( \delta - \frac{\mu}{\rho} \right) \quad (5.33)$$

Fitness gradient:

We note  $\nabla \rho = \frac{\partial \rho(\theta')}{\partial \theta'} \Big|_{(\theta'=\theta)}$

$$\frac{\partial f(\theta', \theta)}{\partial \theta'} \Big|_{\theta'=\theta} = \nabla \rho \left[ \frac{\mu}{\rho} - \gamma \left( \delta - \frac{\mu}{\rho} \right) \right] - \nabla \gamma (\delta \rho - \mu) - \nabla \mu' \quad (5.34)$$

$$= \frac{\nabla \rho}{\rho} [\mu - \gamma (\delta \rho - \mu)] - \nabla \gamma (\delta \rho - \mu) - \nabla \mu' \quad (5.35)$$


---



## Chapter 6

# An Ecological Recipe for the Evolution of Collective-level Heredity

“Reproduire un être vivant, au contraire, ce n’est pas recopier le parent tel qu’il est au moment de la procréation. C’est créer un nouvel être. C’est mettre en route, à partir d’un état initial, une série d’évènements qui conduisent à l’état des parents.”

— FRANÇOIS JACOB, *La Logique du Vivant: Une histoire de l’hérédité* (1970)

THIRTY YEARS AGO, in an article arguing the importance of the “super-organism”, Wilson and Sober expressed surprise that biologists had not recognised that communities — in the laboratory — “*could be treated as entities with heritable variation and selected accordingly*” (Wilson and Sober, 1989). That they might be treated as such stemmed from recognition that the eukaryotic cell is a tight-knit community of two once free-living microbes (Margulis, 1970), but also from observations in nature of social insect colonies (Wilson, 1985), cellular slime molds (Bonner, 1982; Buss, 1982), and especially of phoretic insect communities (Wilson and Knollenberg, 1987).

Phoretic insect communities comprise a focal organism — often an insect such as a beetle — that moves between patchily distributed ephemeral resources carrying with it a myriad of associated organisms, including mites, nematodes and microbes. Communities associated with each insect differ by virtue of the composite members, with the conceivable possibility that some associations may harm the carrier insect, while others may benefit the carrier. Given that the role of dispersal is somewhat analogous to a community-level reproductive event, Wilson and Sober argued that selection at the level of insect communities was likely to trump within-community selection leading to communities “*becoming organised into an elaborate mutualistic network that protects the insect from its natural enemies, gathers food, and so on.*”

If this might happen in nature, then why might this not be realised even more potently in the laboratory? Indeed, the logic of Darwinism says it should. Provided there exists heritable variance in fitness at the level of communities, then communities will participate as units in the process of evolution

by natural selection in their own right. In nature, the necessary conditions are likely rare (Goodnight and Stevens, 1997), but ecological circumstances can sometimes conspire to ensure that variation among communities is discrete, that communities replicate and that offspring communities show some resemblance to parental communities. Phoretic insect communities are an apparent case in point. In the laboratory, however, the experimenter can readily construct conditions that ensure communities (or any collective of cells) are units of selection in their own right (Johnson and Boerlijst, 2002; Day et al., 2011; Xie and Shou, 2018; Xie et al., 2019). A critical requirement is a birth-death process operating over a time scale longer than the doubling time of individual cells (Hammerschmidt et al., 2014; Rainey et al., 2017; Black et al., 2019).

Empirical proof that selection really can shape communities was provided by Swenson and colleagues who performed two studies in which artificial selection was imposed on microbial communities from soil (Swenson et al., 2000a,b). In the first they imposed selection for communities that affected plant growth. In the second they selected communities for ability to degrade the environmental pollutant 3-chloroaniline. In both instances communities at extreme values of community function were repeatedly propagated. In both studies a significant response was measured at the level of the community.

Although the finding was a surprise to some (Goodnight, 2000), it is consistent with expectations that communities of entities — no matter their identity — will participate in the process of evolution by natural selection provided communities are discrete, they replicate and that offspring communities resemble parental communities (Godfrey-Smith, 2009). Discreteness is conferred by simply compartmentalising communities via their placement in independent vessels. Replication is achieved by taking a sample of the selected communities with transfer to a new vessel. Heredity, however, is less tangible, especially in the Swenson experiments, where the selected communities were pooled before redistributing into fresh vessels. Nonetheless, intuition says that heredity becomes established through interactions (Wilson and Sober, 1989; Goodnight, 2000). Understanding the mechanistic bases of community-level heredity and its emergence motivates our study.

We begin by posing a thought experiment realisable via ever improving capacity to manipulate small volumes of liquid (Baraban et al., 2011; Sackmann et al., 2014; Cottinet et al., 2016). Consider a millifluidic device that controls the composition of emulsions. Consider thousands of microlitre-sized droplets each harbouring communities comprised of two types of microbes that differ solely in the colour of a fluorescent protein: one type encodes a red fluorescent protein and the other a blue fluorescent protein. Interest is in the evolution of communities that are of the colour purple (an equal ratio of red to blue cells). Within each droplet red and blue microbes replicate with growth rate and interaction rates being subject to evolutionary change. In the mean time the experimenter, via lasers installed on the device, has determined the precise colour of each droplet and *a priori* decided that half of the droplets furthest from an equal ratio of red-to-blue will be eliminated, whilst the fraction closest to the colour purple will be allowed to replicate. Replication involves a dilution step during which nutrients are replenished. A further round of growth then ensues along with a further round of droplet-

level selection. The protocol continues thereafter with selection taking place at the level of communities via a birth-death process. Figure 6.1 depicts the schema.

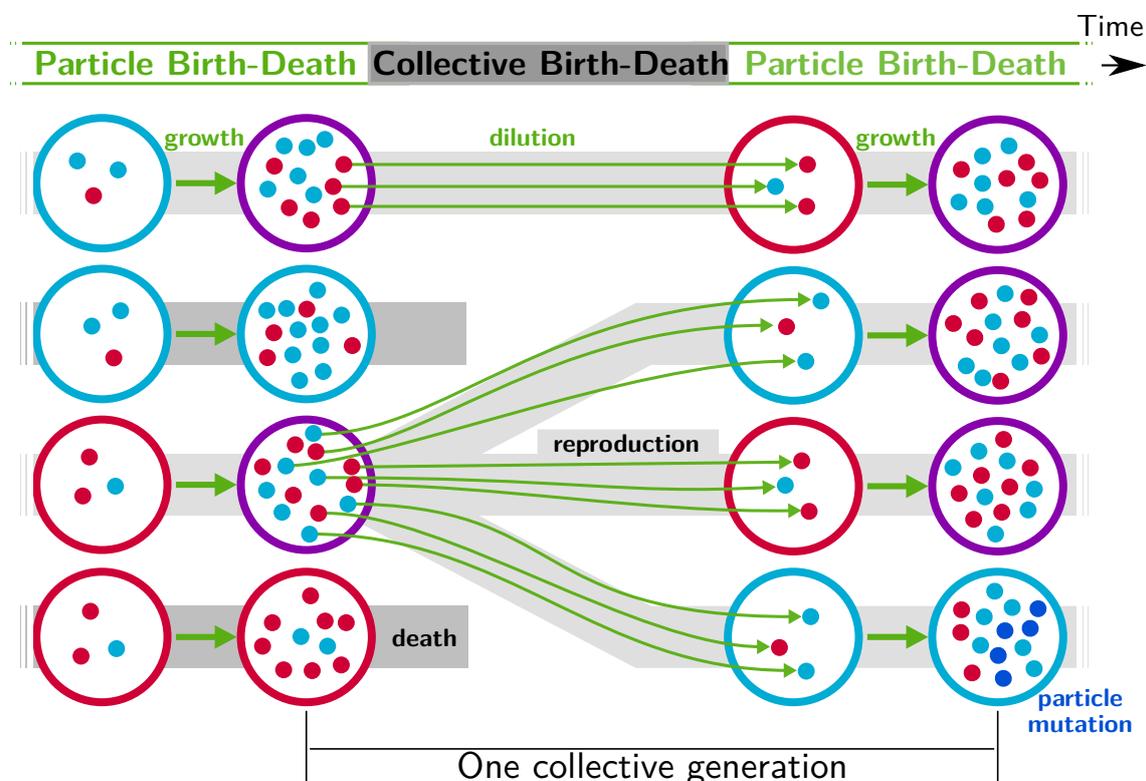


Figure 6.1: **Nested model of evolution.** Collectives (large circles) follow a birth-death process (grey) with non-overlapping generations. Collectives are composed of particles (small spheres) that also follow a birth-death process (*growth*, represented by thick green arrows). Offspring collectives are founded by sampling particles from parent collectives (*dilution*, represented by thin green arrows). Survival of collectives depends on colour. Collectives that contain too many blue or red particles are marked for extinction. The number of collectives is kept constant. Mutation affects particle traits (see main text for details).

Under this regime communities within droplets are endowed with Darwinian-like properties (we use this term to convey the fact that removal of the scaffold leads to, at least initially, complete loss of community-level individuality). Collective-level variation is discretised by virtue of the bounds provided by the immiscibility of oil and water (communities are thus confined to droplets). Additionally, the device ensures that droplets engage in a birth-death process: droplets furthest from the collective-level trait are extinguished, whereas those closest to the colour purple are diluted and split, thus effecting collective-level reproduction. Not determined by the device however is the relationship between parent and offspring droplets. Because the trait of the parent community depends on properties of the cellular constituents, there is — in the absence of interactions between red and blue cells

— little chance that purple-coloured communities will reliably give rise to purple-coloured offspring. This is in part due to the stochastic nature of the dilution phase (a droplet with an equal ratio of red to blue is unlikely to give rise to offspring droplets founded with the same equal ratio of types) but also to within-droplet selection favouring fast growing types. Nonetheless, purple coloured droplets can in principle be maintained simply by imposing strong community-level selection in a manner envisioned by the "stochastic corrector" model (Maynard Smith and Szathmary, 1995; Grey et al., 1995; Johnston and Jones, 2015). But it is conceivable that repeated imposition of droplet-level selection might lead to the evolution of interactions among the constituent cells akin to a developmental process. Such a process would ensure that the ratio of red to blue cells, no matter their ratio in newly "born" communities, would, after replication of cells and achievement of "adult" stage, resemble previous parental community phenotypes.

## 6.1 Results

### 6.1.1 A nested model of collective evolution

Our model, based on the schema outlined above and depicted in Figure 6.1, supposes a population of collectives subject to a birth-death process. Each collective is comprised of two kinds of self-replicating particles (red and blue) that together determine collective colour. Colour is important because it is the phenotype upon which collectives succeed or fail. Collectives that are too far from an optimal colour face extinction, whereas those within acceptable bounds persist with the possibility of reproduction. Birth-death at the level of collectives affects the eco-evolutionary dynamics of particles. We capture this scenario with a nested model and begin with numerical simulations of the stochastic, multi-level demographic process. In the next section a deterministic approximation is presented.

In the numerical model (Appendix 6.3), the ecological dynamic of particles is expressed by a stochastic birth-death process (Champagnat et al., 2006; Doebeli et al., 2017). Each particle of type  $i \in \{0, 1\}$  is characterised by four traits (hereafter *particle traits*): colour ( $c_i$ , red or blue), net maximum growth rate  $r_i$ , and two competition parameters ( $a_i^{\text{intra}}$  and  $a_i^{\text{inter}}$ ). At any particular instant particles either reproduce or die. Particles of type  $i$  reproduce with a constant birth rate  $r_i$  and die as a consequence of competition. The rate of death is density-dependent such that each particle of type  $j$  increases the death rate of  $i$ -type particles by  $r_i a_j^{\text{intra}}$  if they share the same colour ( $c_j = c_i$ ), or by  $r_i a_j^{\text{inter}}$  when colours are different ( $c_j \neq c_i$ ). All transition rates can be found in Appendix 6.3 Table 6.3.4.

Mutations are introduced at the level of particles. Mutation affects either particle maximum growth rate ( $r$ ) or the inter-colour competition parameter ( $a^{\text{inter}}$ ) by a small random quantity. In the spirit of adaptive dynamics (Geritz et al., 1998), the particle type carrying the new set of traits is referred to as a mutant, and the existing type is designated the resident. Mutations are assumed to be rare. In order to accelerate numerical simulations one mutant type is introduced at every extinction event, with the mutant bearing the same colour as the extinct type.

Collectives also undergo a birth-death process. The number of collectives  $D$  is constant and collective generations are discrete and non-overlapping. Each collective generation begins at time  $t = 0$  with offspring collectives containing  $B$  founding particles. Particles replicate, interact and evolve according to the particle traits. After duration  $T$ , collectives attain “adult” stage, and a fixed proportion of collectives  $\rho$  is marked for extinction. This allows the possibility of selection on collectives based on their properties (the *collective phenotype*), which is derived from the composing particles. Our focus is collective colour, which is defined as the proportion  $\phi$  of red particles.

Initially, collectives contain red and blue particles in uniformly distributed ratios. Collectives are subject to evolution under two contrasting regimes: one neutral and the other selective. Under the neutral regime, the pool of collectives marked for extinction is sampled at random, whereas under the selective regime, collectives marked for extinction are those whose adult colour departs most from an arbitrarily fixed optimal colour  $\hat{\phi}$ . Extinguished collectives are replaced by offspring from uniformly sampled extant collectives (Figure 6.1). All other collectives are replaced by their own offspring. Reproduction involves uniformly sampling  $B$  particles from the parent collective. Particles from one collective never mix with particles from any other. This establishes an unambiguous parent-offspring relationship (De Monte and Rainey, 2014). The adult colour of offspring collectives depends on the founding frequencies of particles (whose variance is negatively related to bottleneck size  $B$ ), and on ensuing particle-level population dynamics.

### 6.1.2 Selection on collectives drives the evolution of particle traits

In the absence of collective-level selection (neutral regime), collectives rapidly converge to a monochromatic phenotype (Figure 6.2A). Once collectives are composed of either all-red or all-blue particles, the contrasting colour cannot be rescued, since colour change by mutation or migration is not possible. The distribution of collective colour is thus quickly biased toward faster growing particle types with selection driving particle growth rate (Figure 6.2C). The inter-colour competition trait (Figure 6.2E) is irrelevant once collectives become monochromatic (it then evolves by pure drift).

The dynamic is very different once selection is imposed at the level of collectives. By rewarding collectives closest to the colour purple (a fixed  $\hat{\phi} = 0.5$  ratio of red to blue particles) it is possible to prevent fixation of either colour (Figure 6.2B). Starting, as above, from collectives containing red and blue particles in uniformly distributed ratios, mean collective colour shifts rapidly toward red as a consequence of the faster initial growth rate of red particles, but after a few tens of generations mean collective colour approaches purple. From generation 1,000, variance of the colour distribution progressively decreases. This is indicative of improvement in the ability of purple parent collectives to give rise to offspring collectives that resemble parental types. This is associated with escalating particle growth rate (Figure 6.2D) and a saturating increase in between-colour competition. The latter reflects directional selection that moves the average phenotype in the population of collectives towards an optimal trait value (reached by generation 7,000 see

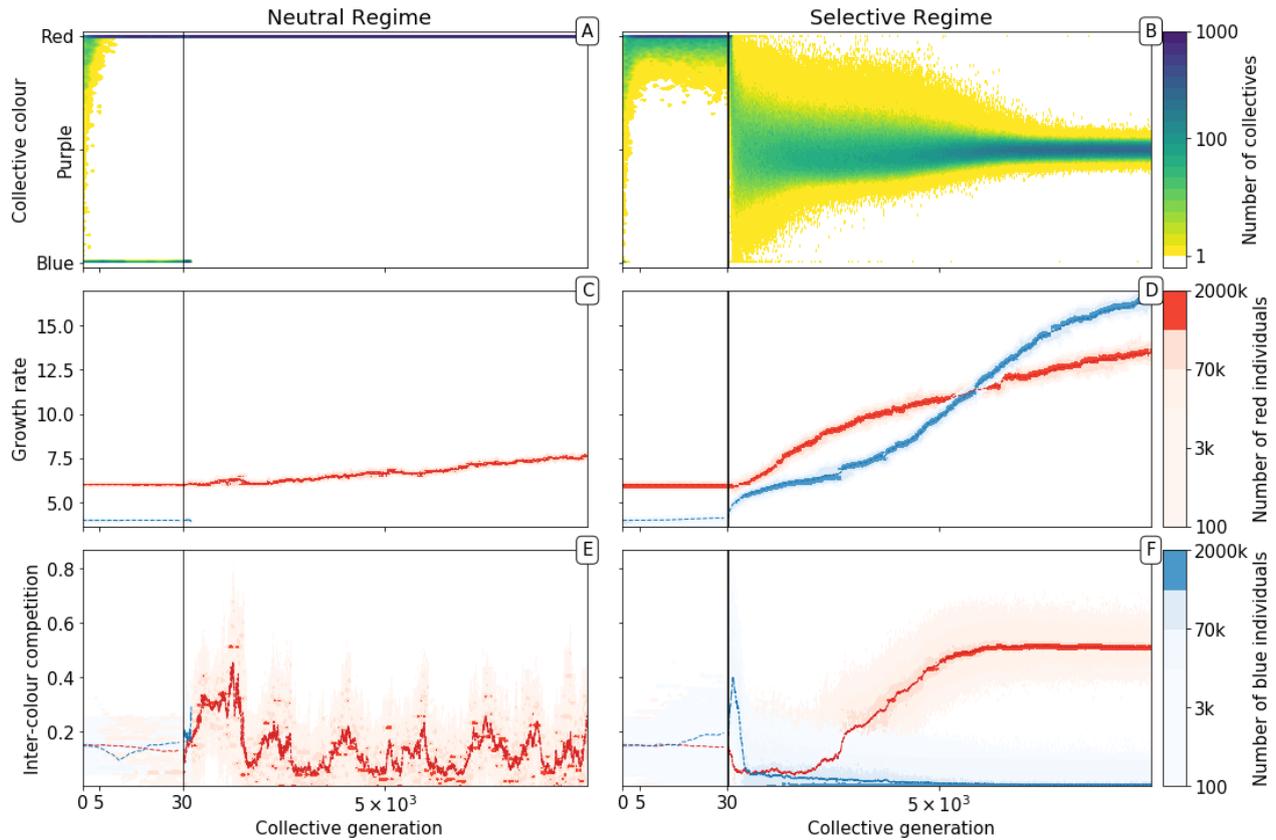


Figure 6.2: **Evolutionary dynamics of collectives and particles.** A population of  $D=1,000$  collectives was allowed to evolve for  $M = 10,000$  generations under the stochastic birth-death model described in the main text (see Appendix 6.3 for details on the algorithm used for the numerical simulations). Initially, each collective was composed of  $B = 15$  particles of two types: red ( $r_0 = 6, a_0^{\text{intra}} = 0.8/K, a_0^{\text{inter}} = 0.15/K, c_0 = \text{red}$ ) and blue ( $r_1 = 4, a_1^{\text{intra}} = 0.3/K, a_1^{\text{inter}} = 0.15/K, c_1 = \text{blue}$ ), with  $K = 1,500$ . The proportions at generation 0 were randomly drawn from a uniform distribution. At the beginning of every successive collective generation, each offspring collective was seeded with founding particles sampled from its parent. Particles were then grown for a duration of  $T = 1$ . When the adult stage was attained, 200 collectives ( $\rho = 20\%$ ) were extinguished, allowing opportunity for extant collectives to reproduce. Collectives were marked for extinction either uniformly at random (*neutral regime*, panels A, C, E, as well as Appendix 6.3 Figures 6.9A and 6.10A), or based on departure of the adult colour from the optimal purple colour ( $\hat{\phi} = 0.5$ ) (*selective regime*, panels B, D, F, as well as Appendix 6.3 Figures 6.9B and 6.10B). Panels A and B, respectively, show how the distribution of the collective phenotype changes in the absence and presence of selection on collective colour. The first 30 collective generations (before the grey line) are magnified in order to make apparent early rapid changes. In the absence of collective-level selection purple collectives are lost in fewer than 10 generations leaving only red collectives (A) whereas purple collectives are maintained in the selective regime (B). Selection for purple-coloured collectives drives evolutionary increase in particle growth rate (D) compared to the neutral regime (C). In the neutral regime, inter-colour evolution of competition traits are driven by drift (E), whereas with collective-level selection density-dependent interaction rates between particles of different colours rapidly achieve evolutionarily stable values, with one colour losing its density-dependence on the other (F).

Figure 6.2F).

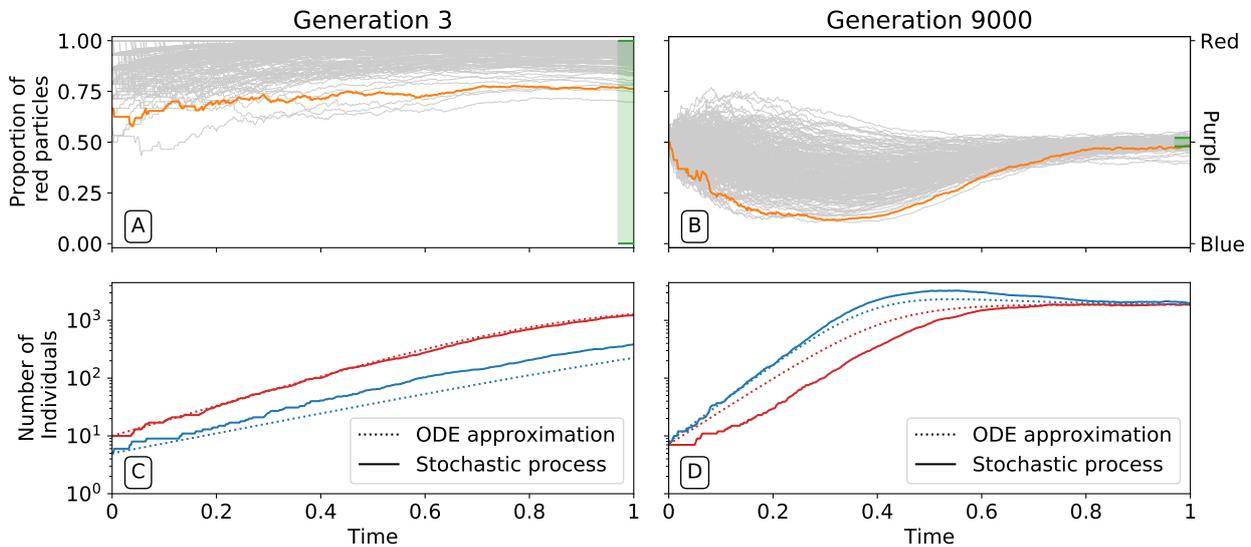


Figure 6.3: **Ecological dynamics of particles.** A sample of 300 (from a total of 1,000) collectives were taken from each of generations 3 (A,C) and 9,000 (B,D) in the evolutionary trajectory of Figure 6.2. The dynamic of particles was simulated through a single collective generation ( $0 \leq t \leq T = 1$ ), based on the particle traits of each collective. Each grey line denotes a single collective. The green area indicates the  $1 - \rho = 80\%$  fraction of collectives in the whole population whose adult colour is closest to  $\hat{\phi} = 0.5$ . Single orange lines indicate collectives whose growth dynamic — number of individual particles — is shown in C and D, respectively. Dotted lines show the deterministic approximation of the particle numbers during growth (Appendix 6.4 Equation 6.1). Initial trait values result in exponential growth of particles (C), leading to a systematic bias in collective colour toward fast growing types (A). Derived trait values after selection yield a saturating growth toward an equilibrium (B) leading to the re-establishment of the purple colour by the end of the generation, despite initial departure (A).

By affecting particle parameters, selection on colour also modifies dynamics within collectives. Figure 6.3 shows variation of colour within a single collective growth phase at generation 3 and generation 9,000. Prior to selection shaping particle traits, both red and blue particle types follow approximately exponential growth (Figure 6.3C). The resulting adult collective colour is thus biased towards the faster-growing red type. In contrast, at generation 9,000 (Figure 6.3B), both particle types reach a saturating steady state that ensures that adult colour is purple. Initial departures from a 1:1 ratio — caused by the stochasticity of collective reproduction and / or particle growth dynamics — are compensated for during the latter part of the growth phase (Figure 6.3D). Compensation is a consequence of the evolution of inter-colour competition traits (Figure 6.2F). Population expansion is in turn dependent upon earlier increases in particle growth rate (Figure 6.2D). Moreover, selection favours competition traits values for which blue types have no effect on red types:  $a^{inter}$  of blue types is close to zero by generation 5,000 (Figure 6.2F).

The ability of offspring collectives to correct departures from the optimal

colour during the course of growth is akin to a developmental, or canalising process: irrespective of phenotype of the newborn (which will likely be different to that of the adult) the child — as it grows to adulthood — develops a phenotype that closely resembles that of the parent. Evidence of this apparent canalising process can be seen upon removal of collective-level selection (Figure 6.4). Collectives founded by particles with ancestral traits become composed of a single (red or blue) colour in less than 10 generations (Figure 6.4A). In contrast, derived collectives are comprised of particles whose traits ensure that collectives continue to express phenotypes narrowly distributed around the optimal (purple) phenotype (as long as mutation is turned off (Figure 6.4B)). Even when mutation is allowed to drive within- and between-collective dynamics, stability of phenotype holds for more than 200 generations (Appendix 6.3 Figure 6.11).

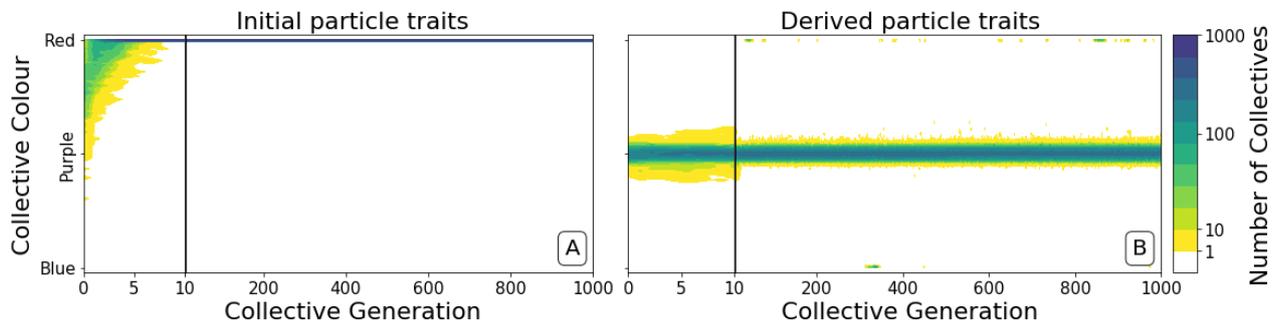


Figure 6.4: **Dynamics of ancestral and derived collectives in the neutral regime.** Comparison of the dynamics of the colour distribution after removing selection (neutral regime). The population of 1,000 collectives is initially composed of collectives with a colour distribution identical to that at generation 10,000 in Figure 6.2B. Particle traits are: (A) as in generation 1 of Figure 6.2; (B) derived after 10,000 generations of collective-level selection for purple. In both instances, particle mutation was turned off in order to focus on ecological dynamics, otherwise parameters are the same as in Figure 6.2 A. Appendix 6.3 Figure 6.11 shows the outcome with particle mutation turned on. The first 10 collective generations are magnified in order to make apparent the initial rapid changes. The particle traits derived after evolution are such that the majority of collectives maintains a composition close to the optimum  $\hat{\phi}$  even when the selective pressure is removed. This feature is instead rapidly lost in populations of collectives with the same initial colour, but with particle traits not tuned by evolution.

### 6.1.3 From particle ecology to collective phenotype

To understand the mechanistic basis of the canalising process, particle traits must be linked to the evolutionary emergence of collective-level inheritance, which we define as the capacity of collectives to re-establish the parental collective colour. Figure 6.5 shows the relationship between the initial colour of collectives at the moment of birth (the moment immediately following dilution,  $t = 0$  (the newborn colour)), and collective colour after a single particle growth cycle (the moment immediately preceding dilution,  $t = T$  (the adult colour)). Figure 6.5A shows this relationship at generation 3 while Figure 6.5B shows this relationship at generation 9,000 of the selective regime.

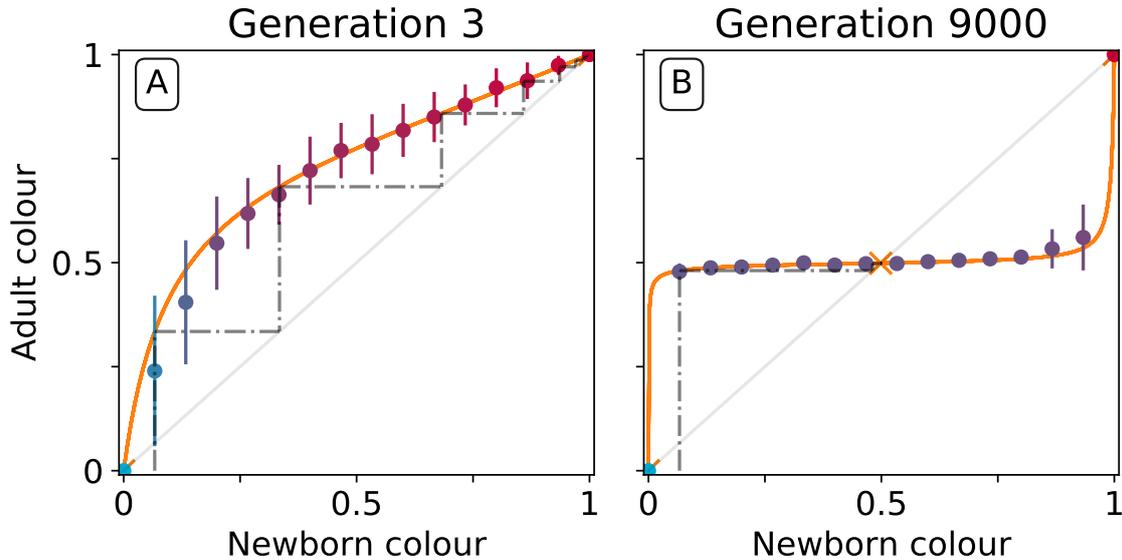


Figure 6.5: **Effect of collective-level selection on newborn-to-adult colour.** The adult colour of collectives as a function of their newborn colour is displayed for collectives of uniformly distributed initial colour. Stochastic simulations are realized by using particle traits representative of: A) generation 3 and B) generation 9000 (as in Figure 6.3). Dots indicate the mean adult colour from 50 simulations and its standard deviation. The orange line depicts the growth function  $G$  for the corresponding deterministic approximation (see main text and Appendix 6.4). The dashed line traces the discrete-time deterministic dynamics of the collective colour, starting from  $\phi = \frac{1}{B}$ , and across cycles of growth and noise-less dilution. For ancestral particle traits (A), collective colour converges towards the red monochromatic fixed point. After selection for collective colour (B), the growth function is such that the optimum colour ( $\hat{\phi}$ ) is reliably produced within a single generation for virtually the whole range of possible founding colour ratios. The latter mechanism ensures efficient correction of alea occurring at birth and during development.

At generation 3, the proportion of red particles increases (within a collective generation), irrespective of the initial proportion. This is because red particles grow faster than blue and the primary determinant of particle success is growth rate (interactions are negligible in exponential growth). Thus, the only way that purple collectives can be maintained is if the collective phenotype is sufficiently noisy, to ensure that some collectives happen by chance to be purple, due to, for example, stochastic effects at dilution. Even if offspring collectives do not resemble their parents, purple colour is maintained via strong purifying selection that purges collectives that are either too red or too blue. This mechanism has been referred to as stochastic correction (Maynard Smith and Szathmary, 1995; Grey et al., 1995; Johnston and Jones, 2015).

This is in marked contrast to the situation at generation 9,000. After a single growth cycle, the proportion of red particles increases when the initial proportion is below, and decreases when it is above the optimal proportion 0.5. Thus at generation 9,000, irrespective of initial conditions, the adult colour of any given collective will be closer to  $\hat{\phi} = 0.5$  than it was on founding.

Accordingly, extreme purifying selection is no longer required to maintain the parent-offspring relationship. Indeed, offspring collectives return to the parent phenotype even when the phenotype at birth departs significantly from the parent (adult) phenotype. This “correction” stems from the ecological dynamics of the particles and resembles a developmental process. Hereafter we refer to this correction process as the *developmental corrector*.

The relationship between newborn and adult colour of collectives shown in Figure 6.5 can be used to follow the fate of collectives over several cycles of growth and reproduction, provided the stochastic effects associated with the dilution phase are momentarily ignored. The iteration using particle trait values from generation 3 is shown by the dotted line in Figure 6.5A (the adult colour of a collective is used as the newborn colour for the next cycle, following a “staircase” geometric procedure). Because red particles grow faster than blue, it takes just six collective generations for red particles to fix within collectives. Conversely, after particle trait evolution (Figure 6.5B), the same staircase approach applied to newborn collectives of any colour shows rapid convergence to the colour purple (0.5) irrespective of the starting point. The difference in the relationship between initial and final colour at generation 3 and 9,000 is evidence of the emergence of a mechanism for developmental correction.

In order to systematically explore the possible newborn-to-adult colour map and to understand how it changes through the evolution of particle traits, we use a deterministic approximation (orange line in Figure 6.5). This approximation is denoted  $G$  or *growth function* (Appendix 6.4 Definition 1) and stems from an ordinary differential equation model often referred to as the competitive Lotka-Volterra system (Appendix 6.4 Equation 6.1). This model is the limit for vanishing noise of the stochastic particle ecology, and provides a good approximation of the simulations (Dotted lines in Figure 6.3C-D). The growth function  $G$  captures the outcome of the ecological dynamics (i.e., the fraction of red particles) after founding populations are allowed to grow for a finite time interval  $T$ . We note similarity between the  $G$  function and the recently proposed “community-function landscape” (Xie and Shou, 2018). The shape of  $G$  depends on the value of particle traits  $\theta$  (growth rates  $r_0$  and  $r_1$ , and competition parameters  $a_{00} = a_0^{\text{intra}}$ ,  $a_{10} = a_0^{\text{inter}}$ ,  $a_{01} = a_1^{\text{inter}}$ ,  $a_{11} = a_1^{\text{intra}}$ ), but also on the bottleneck size at dilution  $B$  and the collective generation duration  $T$ . The fixed points of  $G$  (i.e.,  $\phi$  such that  $G(\phi) = \phi$ ) are of particular interest: in the deterministic model these represent colours that are left unchanged during a generation. A fixed point is said stable if the colours of collectives starting in its neighbourhood all converge to it ( $\phi = 1$  in Figure 6.5A,  $\phi = 0.5$  in Figure 6.5B), and unstable otherwise ( $\phi = 0$  in Figure 6.5A,  $\phi = 0$  and  $\phi = 1$  in Figure 6.5B).

Under collective-level selection for colour,  $T$  and  $B$  are constant and particle traits evolve so that  $G$  eventually has a stable fixed point, corresponding to the target colour  $\hat{\phi}$ . Progressive change in shape of the  $G$  function across collective generations in a simulated lineage (Figure 6.2B) is illustrated in Figure 6.6. Note that these changes are continuous: small mutations in particle traits reflect as small changes in the shape of  $G$ .

The evolutionary trajectory of Figure 6.2B can now be understood in terms of the progressive evolution of particle traits (see Appendix 6.4 for a

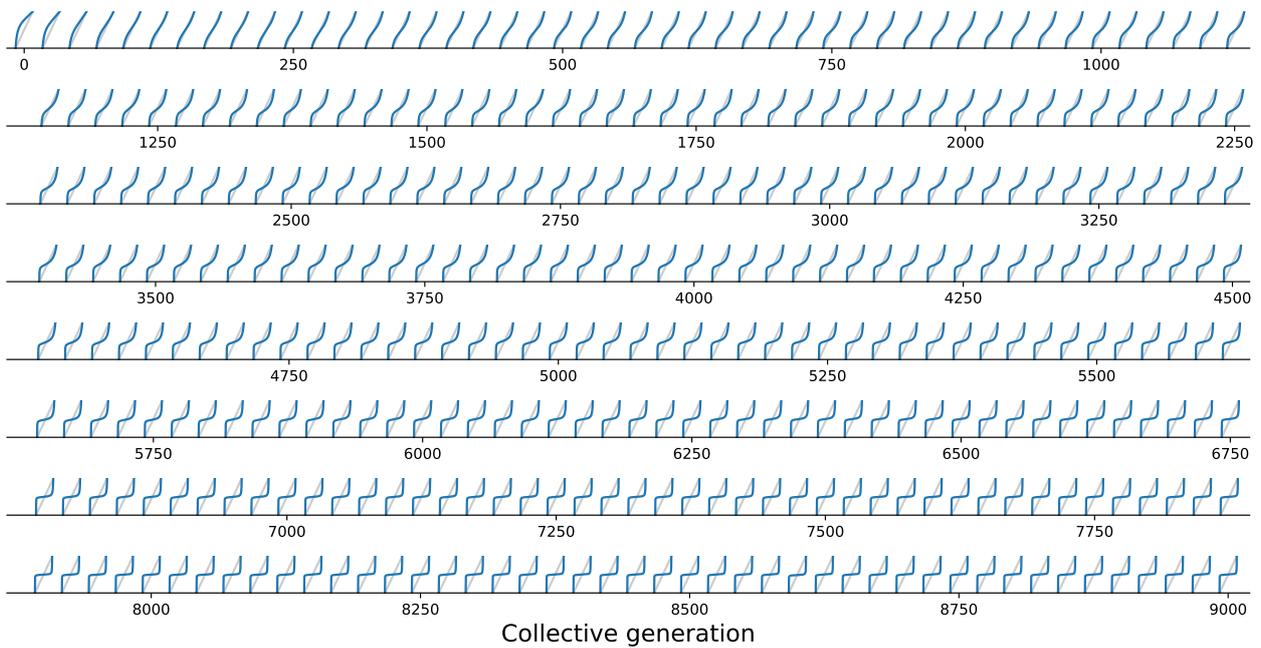


Figure 6.6: **Evolutionary variation of the growth function under collective selection.**  $G$  function associated with the resident types for a single lineage of collectives from the simulation of Figure 6.2B, plotted every 20 collective generations from 0 to 9000. The result of iterations of  $G$  gradually changes from fixation of the fast growing particle (Figure 6.7-1) to convergence toward the colour purple (Figure 6.7-4).

detailed description). At the beginning, particles compete mostly during exponential phase, so that adult colour is biased towards the fast-growing type. Initial improvement in transmission of colour from parent to offspring arises as exponential growth rates  $r_i$  of the particles align. Correspondingly, the  $G$  function approaches linearity. A successive increase in maximal growth rate separates particle and collective time scales, allowing particles to experience density-dependent interactions. Eventually, such interactions evolve towards a regime where the  $G$  function is nonlinear and fluctuations are readily compensated, thus developmental correction ensures a reliable colour inheritance.

The  $G$  function, which allows characterisation of particle ecology, can now be used as a guidance to optimize the “life cycle” of growth and dilution that acts as a scaffold to the evolutionary process. In a typical experiment of community-level evolution, collective generation duration  $T$  and bottleneck size  $B$  are fixed. Some choices of these collective-level parameters are however likely to lead to collective phenotypes that are so far from the optimum that collective lineages go extinct under a given selection rule. For instance, if in the first generation competitive exclusion occurs rapidly, then distinguishing collectives based on collective colour may be impossible. Intuition suggests that the closer the fixed point of the  $G$  function is to the target colour, the more efficient will collective-level selection be, and the faster the evolutionary dynamics. It is thus possible to use the distance between the fixed point of

$G$  and the target composition  $\hat{\phi}$  as a proxy for the probability that collective lineages may go extinct before attaining the desired colour. Below, we examine how the position of the fixed point of  $G$  changes as a function of collective generation duration  $T$  and bottleneck size  $B$ .

#### 6.1.4 Effect of collective generation duration and bottleneck size

The growth function  $G$  is readily computed from the particle traits and collective parameters even though it has in general no analytic expression (but see Appendix 6.4 for limit cases of exponential and saturating particle growth). There are four possible qualitative shapes of  $G$ , that differ in the position and stability of the fixed points (illustrated in Appendix 6.4 Figure 6.14-1 to 6.14-4).

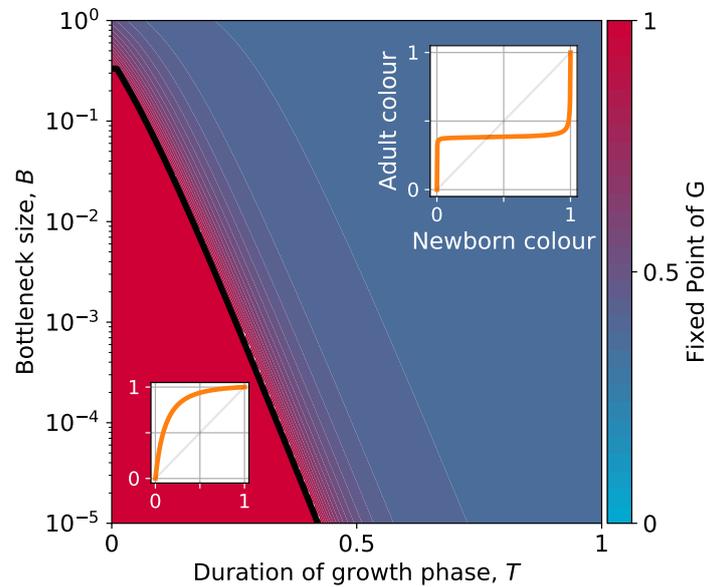


Figure 6.7: **Stable fixed point of  $G$  as a function of collective-level parameters.** Classification of the qualitative shape of the growth function and dependence on collective parameters  $B$  (bottleneck size) and  $T$  (growth phase duration), for the case when particle traits allow coexistence ( $a_{01} < a_{11}$  and  $a_{10} < a_{00}$ ,  $r_0 > r_1$ , see Appendix 6.4 Figure 6.14 for the other possible cases). The black line represents the limit of the region of stability of the fixed point of  $G$ , separating the two qualitatively different scenarios illustrated in the inset (see Appendix 6.4 Proposition 26): for short collective generations and small bottleneck size, the faster growing red type competitively excludes the blue type over multiple collective generations. In order for particle types to coexist over the long term, growth rate and the initial number of particles must both be large enough for density-dependent effects to manifest at the time when selection is applied.

The qualitative dependence of  $G$  and its fixed points on collective-level

parameters varies with the underpinning particle ecology, making it easier for some communities to be starting points for the successful evolution of inheritance. Particle traits can be classified in four broad classes, based of the nature of the corresponding ecological equilibrium. For each of these classes, and when red particles grow faster than blue  $r_0 > r_1$ , the fixed points of  $G$  are illustrated in Appendix 6.4 Figure 6.14-A to 6.14-D as a function of the collective-level parameters  $B$  and  $T$ . Figure 6.7 refers to the relevant class where inter-colour interaction traits are smaller than the intra-colour interaction traits. Here, the particle populations converge in the long term to a coexistence equilibrium, where collective colour is  $\phi^* = \frac{a_{11}-a_{01}}{a_{11}-a_{01}+a_{00}-a_{10}}$  (in general, different from the optimum). This equilibrium can be approached within a single collective generation if  $T$  and  $B$  are large (top right corner). On the other hand, when  $T$  and  $B$  are small (red region), the only stable fixed point are collectives composed solely of fast-growing particles. This corresponds to cases where individual and collective time scales (quantified by  $r^{-1}$  and  $T$ , respectively) are insufficiently separated, or newborn size is too small, so that particle demography is essentially exponential and interactions cannot provide sufficient correction. In order to speed up evolution of purple colour, thus, the best is for collective generation duration and bottleneck size to have intermediate values. Knowledge of the exact values requires however some preliminary measure of the ecological dynamics. Even in the absence of such information, the diagram in Figure 6.7 can be used to orient experimental design by revealing intrinsic trade-offs. A decrease in generation time, necessary for practical reasons, may for instance be compensated by an increase in bottleneck size, without affecting the average collective phenotype.

Even when collective-level parameters are optimized so that the attractor of the  $G$  function is initially close to the target colour, collective-level selection will keep acting on the particle traits, and affect phenotypic variability within the population of collectives. As stability of the fixed point increases, so does fidelity of phenotype transmission from parent to offspring collectives. Once collective-level processes are set as to minimize the probability of collective extinction, the main obstacles to evolving higher inheritance come from constraints acting on particle traits, that may limit the range of attainable  $G$  functions. Trade-offs on particle ecology indeed may prevent the  $G$  function to attain an internal fixed point. We will discuss two examples on constrained evolution in the following paragraph.

### 6.1.5 Constrained trajectories

Thus far we have considered evolution within a four-dimensional parameter space defined by maximum growth rates and inter-color competition parameters. In real systems, however, constraints and trade-offs may limit the range of achievable variations in particle traits. For instance, even though faster-growing particles will always experience positive selection, cell replication rate cannot increase boundlessly. Here we consider two instances of constrained evolution, where only a subset of particle traits are allowed to mutate.

First, we consider the case where competition parameters are vanishingly small, so that  $G$  has no internal fixed point. Under such conditions, particle growth rates evolve to be identical (Figure 6.8A). In the absence of interac-

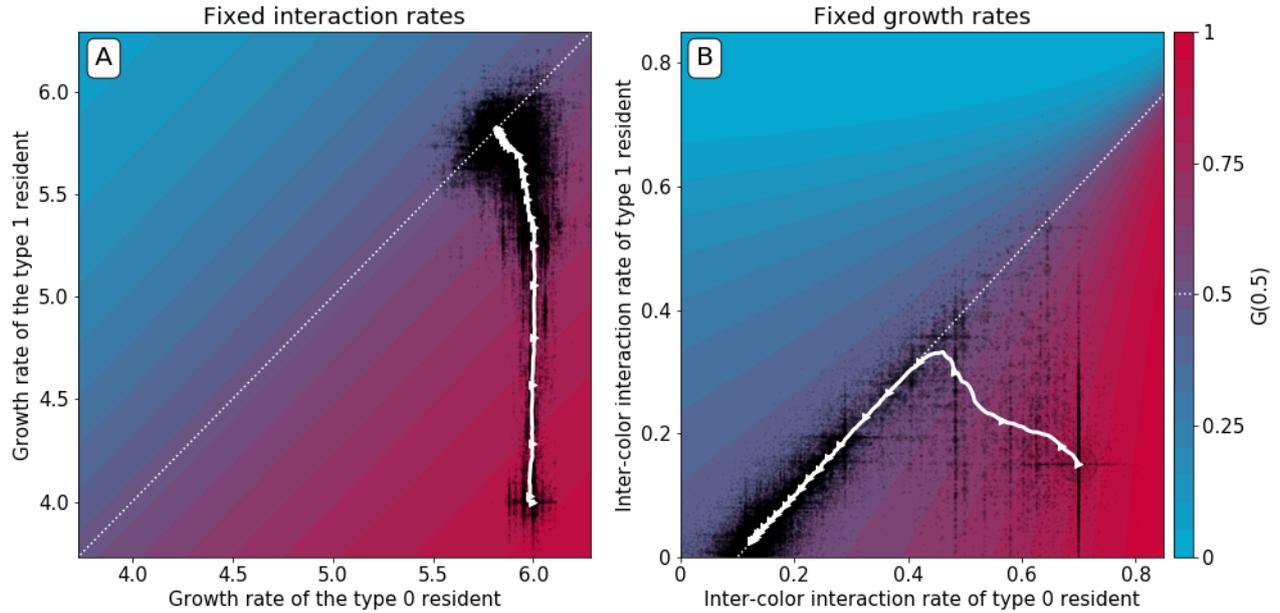


Figure 6.8: **Constrained evolutionary trajectories.** Dynamics through time of resident particle traits (black dots, whose size measures their abundance in the collective population) along simulated evolutionary trajectories of 300 generations, when particle-level parameters are constrained. For both panels  $D = 1000$ ,  $\hat{\phi} = 0.5$ ,  $\rho = 20\%$ ,  $B = 15$ , and  $T = 1$ . The trajectory of the average resident traits is shown in white. The heatmap represents the value of  $G(0.5)$  as a function of the evolvable traits, and the white dotted line indicates where collective colour is optimum. A) Particle growth rates evolve and particles do not compete ( $a_1^{\text{inter}} = a_1^{\text{intra}} = 0$ ). The evolutionary dynamics leads to alignment of growth rates ( $r_0 = r_1$ ). B) Inter-colour competition traits evolve and particle growth rates are constant ( $r_0 = r_1 = 25$ ). The evolutionary dynamics first converge toward the optimality line. In a second step, asymmetric competition evolves:  $a_1^{\text{inter}} \rightarrow 0$  and  $a_0^{\text{inter}} \rightarrow a_0^{\text{intra}} - a_1^{\text{intra}}$ . This results in a flatter  $G$  function around the fixed point, providing a faster convergence to optimum colour across collective generations (Appendix 6.4 Figure 6.17). Similar results are obtained for non-identical, but sufficiently high, growth rates.

tions, this is the only available solution to maintain collectives with an equal number of red and blue type particles. Under these circumstances,  $G$  converges to the identity function. In the deterministic approximation, collective composition remains constant in time, but stochastic fluctuations that cause colour to deviate from the optimum will be amplified across collective generations. These deviations are nonetheless corrected in the collective population by propagating only those collectives whose colour is closest to the optimum. Such stochastic correction (Maynard Smith and Szathmary, 1995), however, has a high risk of failure if selection is strong and collective population size is small.

Second, we consider the case when mutations only affect the two inter-type competition parameters, while growth rates are held constant (to sufficiently high values, so that particles experience density-dependent effects in the growth phase). The evolutionary trajectory can be visualized in the plane of the interaction parameters ( $a_{01}, a_{10}$ ). Figure 6.8B shows the result

of a stochastic simulation superimposed to the value of  $G(0.5)$ . Independent of the initial values of the interaction parameters, evolution draws the system to the manifold associated with the optimal proportion  $\hat{\phi}$  (white dashed line). Evolution within this manifold is neutral in the deterministic approximation, but the presence of stochastic fluctuations drives further improvement of the fitness landscape. Correction is indeed more efficient and the distribution of collective phenotypic diversity narrower when the gradient of  $G$  in the fixed point is smaller. The condition on particle traits for the latter to vanish only depends on the carrying capacities of the two particle types, and corresponds to the type with smallest carrying capacity having zero interaction rate (see Appendix 6.4). A similar outcome is observed when, along an evolutionary trajectory, growth rates stop to influence adult colour 6.2. Developmental correction thus selects for maximal asymmetry in interactions, whereby one particle type drives the ecological dynamics of the other type, but is itself only affected by its own type .

## 6.2 Discussion

In nature, communities rarely ever qualify as units of selection in the traditional sense (Lewontin, 1970; Godfrey-Smith, 2009), because communities in nature rarely manifest heritable variance in fitness. In the laboratory however, experimenters can exogenously impose (scaffold) Darwinian-like properties on communities such that they have no choice, but to become units of selection (Wilson and Sober, 1989; Xie et al., 2019; Black et al., 2019). This typically involves placement of communities in some kind of container (pot, test-tube, flask, droplet, etc.) so they are bounded and variation at the community level is thus discrete. Communities are then allowed time for individual members to replicate and interact. At the end of the “growth” period, community function is assessed based on pre-determined criteria. The experimenter then effects replication of successful communities while discarding those that under-perform. Replication typically involves transferring a sample of individuals from a successful community to a new container replete with a fresh supply of nutrients.

Experimental and theoretical studies demonstrate that artificial selection on microbial communities results in rapid functional improvement (Swenson et al., 2000a,b; Goodnight, 2000; Wade, 2016; Xie et al., 2019). This is not unexpected given that experimental manipulations ensure that communities engage directly in the process of evolution by (artificial) selection as units in their own right. However, for such effects to manifest there must exist a mechanism of community-level inheritance.

Consideration of both the effectiveness of artificial selection and the problem of heredity led to early recognition that the answer likely lies in interactions (Wilson and Sober, 1989; Swenson et al., 2000a,b; Goodnight, 2000; Rainey et al., 2017). The intuition stems from the fact that in the absence of interactions, communities selected to reproduce because of their beneficial phenotype will likely fail to produce offspring communities with similar functionality. If so, then these communities will be eliminated at the next round. Consider however, an optimal community in which interactions emerge among individuals that increase the chance that offspring communities resemble the

parental type. Such an offspring community will then likely avoid extinction at the next round: selection at the level of communities is thus expected to favour the evolution of interactions because inheritance of phenotype is now the primary determinant of the success of communities (at the community level). Indeed, simulations of multi-species assemblages have shown that evolution of interaction rates not only improves diversity-dependent fitness, but also increases collective “heritability”, defined as the capacity of randomly seeded offspring communities to reach the same dynamical state as their parents (Ikegami and Hashimoto, 2002; Penn, 2003). Further studies have stressed the role of the extracellular environment and of specific interaction networks, pointing out that microscopic constraints can affect the capacity of communities to participate in evolutionary dynamics at the higher level (Williams and Lenton, 2007; Xie and Shou, 2018; Xie et al., 2019).

Here, inspired by advances in millifluidics, we have developed a minimal mechanistic model containing essential ingredients of multi-scale evolution and within-community competition. We considered collectives composed of two types of particles (red and blue) that interact by density-dependent competition. By explicitly modelling demographic processes at two levels of organization, we have obtained mechanistic understanding of how selection on collective character affects evolution of composing particle traits. Between-collective selection fuels changes in particle-level traits that feed back to affect collective phenotype. Selection at the level of communities thus drives the evolution of interactions among particles to the point where derived communities, despite stochastic effects associated with sampling at the moment of birth, give rise to offspring communities that reliably recapitulate the parental community phenotype. Such is the basis of community-level inheritance. Significantly, it has arisen from the simplest of ingredients and marks an important initial step in the endogenisation of Darwinian properties: properties externally imposed stand to become endogenous features of the evolving system (Black et al., 2019).

The mechanism by which particles interact to establish community phenotype is reminiscent of a development process. We refer to it as the “developmental corrector”. In essence, it is akin to canalisation, a central feature of development in complex living systems (Buss, 1987), and the basis of inheritance (Griesemer, 2002). Developmental correction solves the problem of implementing specific protocols for mitigating non-heritable variations in community function (Xie et al., 2019).

The developmental correction can be viewed as an evolutionary refinement of the stochastic corrector mechanism (Maynard Smith and Szathmary, 1995; Grey et al., 1995; Johnston and Jones, 2015). Both the stochastic and developmental correctors solve the problem of producing enough well-formed collectives at each successive generation to prevent community-level extinction. The stochastic corrector mechanism relies on a low-fidelity reproduction process coupled to high population sizes. Deviations from successful collective states are corrected by purging collectives that depart significantly from the optimal collective phenotype. However, in the absence of strong collective-level selection the optimal community phenotype is rapidly lost. In contrast, the developmental corrector mechanism ensures that the optimal community phenotype is maintained without need for hard selection. Regardless of per-

turbations introduced by demography or low initial particle number, most collectives reliably reach a successful adult state. In our simulations, we show that community phenotype is maintained even in the absence of community-level selection, although ultimately mutational processes affecting particle dynamics result in eventual loss of the developmental corrector mechanism.

An operationally relevant question concerns the conditions (the initial state of the population, the nature of the scaffold and of particle-level interactions) for selection on a collective character to result in evolution of developmental correction. While we did not detail the probability of collective lineage extinction, it is possible that collectives become monochromatic before evolution has had time to act on particle traits. In such cases, which are more likely if particle-level parameters are far from the coexistence equilibrium, and if time-scales of particle and collective generations are not well separated, then collective-level evolution will grind to a halt. In all other cases, provided there are no other evolutionary constraints, selection will eventually lead the system toward regions of particle parameter space where the collective phenotype becomes reliably re-established. The efficiency of this selective process and its transient behaviour depends on collective-level parameters that control growth and reproduction.

From our models and simulations it is clear that once density-dependent interactions govern the adult state, then collective-level selection for colour is promptly effected. This happens provided the intra-collective ecology lasts long enough for nonlinear effects to curb particle growth. When this is not the case, for example when the bottleneck at birth is small, or collective-level generation time too short, evolution of developmental correction will be impeded. The latter favours rapidly growing particles (Abreu et al., 2019) and offers little possibility for the evolution of developmental correction. When the ecological attractor within collectives leads to the extinction of one of the two types, long collective-level generation times are incompatible with the maintenance of diversity (van Vliet and Doebeli, 2019). However, in our model, particle-level evolution changes the nature of the attractor from extinction of one of the two types to stable coexistence, and concomitantly particle and collective time-scales become separated. Even before developmental correction becomes established, evolution can transiently rely on stochastic correction to ensure the maintenance of particle co-existence.

There are two aspects to heredity: resemblance — the extent to which reproduction and development maintain the average offspring phenotype — and fidelity (or determination) — a measure of its variance (Jacquard, 1983; Bourrat, 2017). In our model, resemblance is established once density-dependent interactions counter the bias toward fast replicating particles: when the  $G$  function has an internal fixed point in  $\hat{\phi}$ , systematic drift of average collective colour is prevented. The increase in resemblance is associated with progressive divergence of particle and collective demographic time scales. As a consequence, the collective phenotype is placed under the control of particle traits rather than demographic stochasticity. On a longer time scale, fidelity improves by subsequent changes in interaction parameters under the constraint that they do not affect average adult colour. The variance of the phenotype around the optimum is reduced by increasing canalization (flattening of the  $G$  function). This is best achieved by a strong asymmetry in the

competition traits, whereby one type has a logistic, uncoupled, dynamic, and the second type adjusts its growth to the former's density. Interestingly, it is always the type with the lower carrying capacity, regardless of its relative growth rate, that acts as the driver.

The mechanism of developmental correction is broadly relevant and extends beyond cells and communities to particles of any kind that happen to be nested within higher-level self-replicating structures. As such, the mechanism of developmental correction may be relevant to the early stages in each of the major (egalitarian) evolutionary transitions in individuality (Queller, 1997; Maynard Smith and Szathmary, 1995), where maintenance of particle types in optimal proportions was likely an essential requirement. For example, it is hard to see how protocells evolved from lower level components (Takeuchi and Hogeweg, 2009; Baum and Vetsigian, 2017), chromosomes from genes (Maynard Smith and Szathmary, 1993), and the eukaryotic cell from independent bacterial entities (Martin and Muller, 1998) without some kind of self-correcting mechanism acting at the collective level.

Beyond these fundamental considerations the mechanism of developmental correction and the ecological factors underpinning its evolution have important implications for top-down engineering of microbial communities for discovery of new chemistries, new functions, and even new organisms. The minimal recipe involves partitioning communities into discrete packages, provision of a period of time for cell growth, selective criteria that lead to purging of sub-optimal collectives and reproduction of optimal collectives to establish the next generation of collectives. These manipulations are readily achieved using millifluidic devices that can be engineered to operate in a Turing-like manner allowing artificial selection on community-level traits across thousands of independent communities. As mentioned above, critical tuneable parameters beyond number of communities, mode of selection and population size, are duration of collective generation time and bottleneck size at the moment of birth.

The extent to which the conclusions based on our simple abstract model are generally applicable to the evolution of more complex associations, such as symbioses leading to new forms of life, will require future exploration of a broader range of particle-level ecologies. Possibilities to make community dynamics more realistic by complexifying mathematical descriptions of particle-level processes are plentiful (Williams and Lenton, 2007; Zomorodi and Segre, 2016). Of particular interest for the evolution of efficient developmental correction are cases when community ecology has multiple attractors (Penn and Harvey, 2004), is highly sensitive to initial conditions (Swenson et al., 2000b), or presents finite-effect mutations sustaining "eco-evolutionary tunnelling" (Kotil and Vetsigian, 2018). Besides enlarging the spectrum of possible within-collective interactions, future relevant extensions may explore the role of physical coupling among particles and of horizontal transmission between collectives (van Vliet and Doebeli, 2019) in enhancing or hampering efficient inheritance of collective-level characters.

## 6.3 Stochastic Model

This appendix presents an outline for the approximated stochastic simulation of nested population dynamics described in the main text. We first attend to the case where particle populations are monomorphic without mutation and then move to consider the role of particle-level mutation. Full implementation of the model is available as a Python package.

### 6.3.1 Parameters

First, we list the parameters of the numerical model introduced in the main text. The collective selection regime is specified by a set of collective-level parameters that are kept constant along the evolutionary trajectory.

```
// Parameters
D ← 1000                                ▷ Number of collectives
M ← 10,000                              ▷ Number of collective generations
B ← 15                                  ▷ Bottleneck size
T ← 1.0                                ▷ Duration of a collective generation
regime ← selective                       ▷ Selective or Neutral regime
ρ ← 0.2                                ▷ Fraction of extinguished collectives
φ* ← 0.5                                ▷ Optimal collective colour (for selective regime)
```

Collectives are comprised of two kinds of self-replicating particle (red and blue) that carry a different set of traits. Traits can mutate (see the mutation section below), and are represented as global variables.

```
// Particle traits
// Carried by red particles
r0                                       ▷ Maximum growth rate
a0intra                                 ▷ Competition with red particles
a0inter                                ▷ Competition with blue particles
// Carried by blue particles
r1                                       ▷ Maximum growth rate
a1intra                                 ▷ Competition with blue particles
a1inter                                ▷ Competition with red particles
```

The state variables ( $D \times M$  matrices) store the adult state of collectives along a trajectory.

```
// State variables
N0                                       ▷ number of blue individuals in each collective at each generation
N1                                       ▷ number of red individuals in each collective at each generation
Φ                                       ▷ proportion of red individuals in each collective at each generation
```

### 6.3.2 Initial conditions

Initial conditions consist in defining the number of red and blue particles in each collective at the beginning of generation 0.

```
procedure INITIAL CONDITIONS
  x0 ← 0.5                                ▷ Initial red-blue ratio
  for d from 1 to D do
    N0[d, 0] ← RandomBinomial(B, x0)
```

$$\mathbf{N}_1[d, 0] \leftarrow B - N_0[d, 0]$$

### 6.3.3 Outline of the main loop

The main loop of the algorithm applies the sequence growth-selection-reproduction for each generation.

```

procedure MAIN LOOP
  for  $m$  from 1 to  $M$  do
    // Particle Growth
    for  $d$  from 1 to  $D$  do
       $\mathbf{N}_0[d, m], \mathbf{N}_1[d, m] \leftarrow \text{GROWTH}(\mathbf{N}_0[d, m], \mathbf{N}_1[d, m], r_0, r_1, a_0^{\text{intra}} a_0^{\text{inter}}, a_1^{\text{intra}} a_1^{\text{inter}})$ 
       $\Phi[d, m] \leftarrow \mathbf{N}_1[d, m] / (\mathbf{N}_0[d, m] + \mathbf{N}_1[d, m])$ 
    // Collective-level selection
    parents  $\leftarrow \text{SELECT\_COLLECTIVES}(\Phi[d, m], \rho, \phi^*)$ 
    // Collective-level reproduction
    for  $d$  from 1 to  $D$  do
       $N_0[d, m + 1] \leftarrow \text{RandomBinomial}(B, \Phi[\text{parents}[d], m])$ 
       $N_1[d, m + 1] \leftarrow B - N_0[d, m + 1]$ 

```

### 6.3.4 Particle-level ecology

The ecological dynamics of particles is expressed by a multi-type birth-death process with a linear birth rate and a linearly density-dependent death rate. Each type of particle  $i$  is characterised by four traits: colour ( $c_i$ , red or blue), maximum growth rate  $r_i$ , and two density-dependent interaction parameters. Particles of the same colour interact according to  $a_i^{\text{intra}}$  whereas particles of different colour interact according to  $a_i^{\text{inter}}$ . Interaction terms are in the order of 0.1 and scaled by a carrying capacity term  $K$ . The dynamic is modelled by a continuous-time Markov jump process with rates:

Each particle of type $i$ ...	With rate...
Reproduces (add a particle of type $i$ )	$r_i$
Dies (remove a particle of type $i$ )	$r_i \left( \sum_j \delta_{c_i=c_j} x_j a_j^{\text{intra}} K^{-1} + \sum_j \delta_{c_i \neq c_j} x_j a_j^{\text{inter}} K^{-1} \right)$

Where  $\delta_{i=j} = 1$  if  $i = j$  or 0 if  $i \neq j$ . Additionally  $\delta_{i \neq j} = 1$  if  $i \neq j$  or 0 if  $i = j$ .

The stochastic trajectory of the system is simulated using a Poissonian approximation used in the basic  $\tau$ -leap algorithm (Gillespie, 2001),  $dt$  is chosen to be small enough so that population size never becomes negative.

```

function GROWTH( $n_0, n_1, r_0, r_1, a_0^{\text{intra}} a_0^{\text{inter}}, a_1^{\text{intra}} a_1^{\text{inter}}$ )
  // Stochastic simulation of the population dynamics
  while  $t < T$  do
    birth0  $\leftarrow \text{RandomPoisson}(dt \times n_0 r_0)$ 
    birth1  $\leftarrow \text{RandomPoisson}(dt \times n_1 r_1)$ 
    death0  $\leftarrow \text{RandomPoisson}(dt \times n_0 r_0 (n_0 a_0^{\text{intra}} + n_1 a_1^{\text{inter}}))$ 
    death1  $\leftarrow \text{RandomPoisson}(dt \times n_1 r_1 (n_0 a_0^{\text{inter}} + n_1 a_1^{\text{intra}}))$ 
     $n_1 \leftarrow n_1 + \text{birth1} - \text{death1}$ 
     $n_0 \leftarrow n_0 + \text{birth0} - \text{death0}$ 
     $t \leftarrow t + dt$ 
  Return  $n_0, n_1$ 

```

Early tests using an exact stochastic simulation algorithm (Doob-Gillespie SSA, Gillespie (1976)) did not exhibit qualitative changes in the trajectory, but greatly increased the computation duration.

### 6.3.5 Collective-level selection

The collective-level selection phase consists in associating each of the  $D$  new collectives from generation  $m + 1$  with a single parent at generation  $m$ . In the main text we contrast two regimes: colour-neutral and colour-selective. In both cases a fixed proportion  $\rho$  of the collective population at generation  $m$  is marked for extinction. In the neutral regime, collectives to be eliminated are selected uniformly at random (Appendix 6.3 Figure 6.9A), whereas in the selective regime they are those that rank the highest in their distance to the optimal colour  $\hat{\phi}$  (Appendix 6.3 Figure 6.9B). Each surviving collective produces offspring. Moreover, the remaining collectives from generation  $m + 1$  are generated by a parent chosen uniformly at random from the set of surviving collectives.

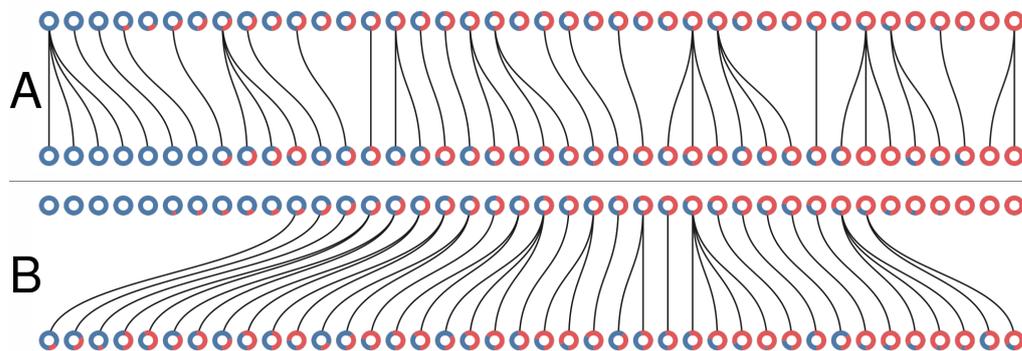


Figure 6.9: **Collective-level selection regimes.** Each ring represents a collective. The blue section of the ring represents the proportion of blue particles at the adult stage of the collective. Parent and offspring are linked by a black line. A: One generation of the neutral regime. B: one generation of the selective regime. In both cases  $M = 1$ ,  $D = 40$ ,  $\rho = 0.4$ .

---

```

function SELECT_COLLECTIVES( $\phi, \rho, \phi^*, \text{regime}$ )
  // Return the indices of the new collectives' parents
  surviving  $\leftarrow$  empty list                                 $\triangleright$  indices of the surviving collectives
  reproducing  $\leftarrow$  empty list                             $\triangleright$  indices of the reproducing collectives
  parents  $\leftarrow$  empty list                                 $\triangleright$  indices of the parent collectives
  if regime is selective then
    threshold  $\leftarrow$  Percentile( $(\phi - \phi^*)^2, \rho$ )
    for  $d$  from 1 to  $D$  do
      if  $(\phi[d] - \phi^*)^2 <$  threshold then
        Add  $d$  to surviving
    else if regime is neutral then
      surviving = RandomMultinomialWithoutReplacement( $1\dots d, n = (1 - \rho)D$ )
  // Extinct collectives are replaced by the offspring of a randomly drawn surviving collective
  for 1 to  $D - \text{Length}(\text{surviving})$  do
    Add RandomChoice(surviving) to reproducing
  // Surviving collectives have at least one offspring, and the population size is kept constant
  // by additional reproduction events.
  parents  $\leftarrow$  Concatenate(surviving, reproducing)
  Return parents

```

Other selection procedures can be implemented, such as randomly sampling  $\rho D$  collectives with weight based on the colour-deviation to the optimal colour. A non-exhaustive exploration of other selection rules indicates that the qualitative results of the model are robust to changes in the selective regime, as long as collectives with an optimal colour are favoured and the collective population does not go extinct.

Collective reproduction is implemented by seeding an offspring collective with a sample of  $B$  particles drawn according to the proportion of the parent collective. We assume that the final particle population sizes are big enough so that each reproduction event can be modelled as an independent binomial sample. Smaller population sizes might require simultaneous multinomial sampling of all offspring.

### 6.3.6 Mutation of particle traits

The complete model adds the possibility for particle trait  $(r, a^{inter})$  to mutate. Each collective contains two variants of each color. Whenever a variant goes extinct, the remaining type is called the ‘resident’, and a mutant type is created as follows. First, traits of the resident are copied in one newborn particle, then one of the mutable traits — either the growth rate ( $r$ ) or the inter-colour competition trait ( $a_{inter}$ ) — is chosen at random, and finally a random value is added that is taken from a uniform distribution over  $[-\varepsilon, \varepsilon]$  (in Figure 6.2,  $\varepsilon = 0.1$ ). Traits are kept positive by taking the absolute value of the result. This process is in the spirit of adaptive dynamics in which invasion of a single new mutant is repeatedly assessed in a monomorphic population. We checked that relaxing this assumption (i.e., allowing more than two types of each colour in each collective), or waiting for rare mutations to appear did not change the qualitative results. The pseudo-code outlined above was modified in order to track the trait value of both resident and mutant types in each collective, rather than having the trait values as global variables.

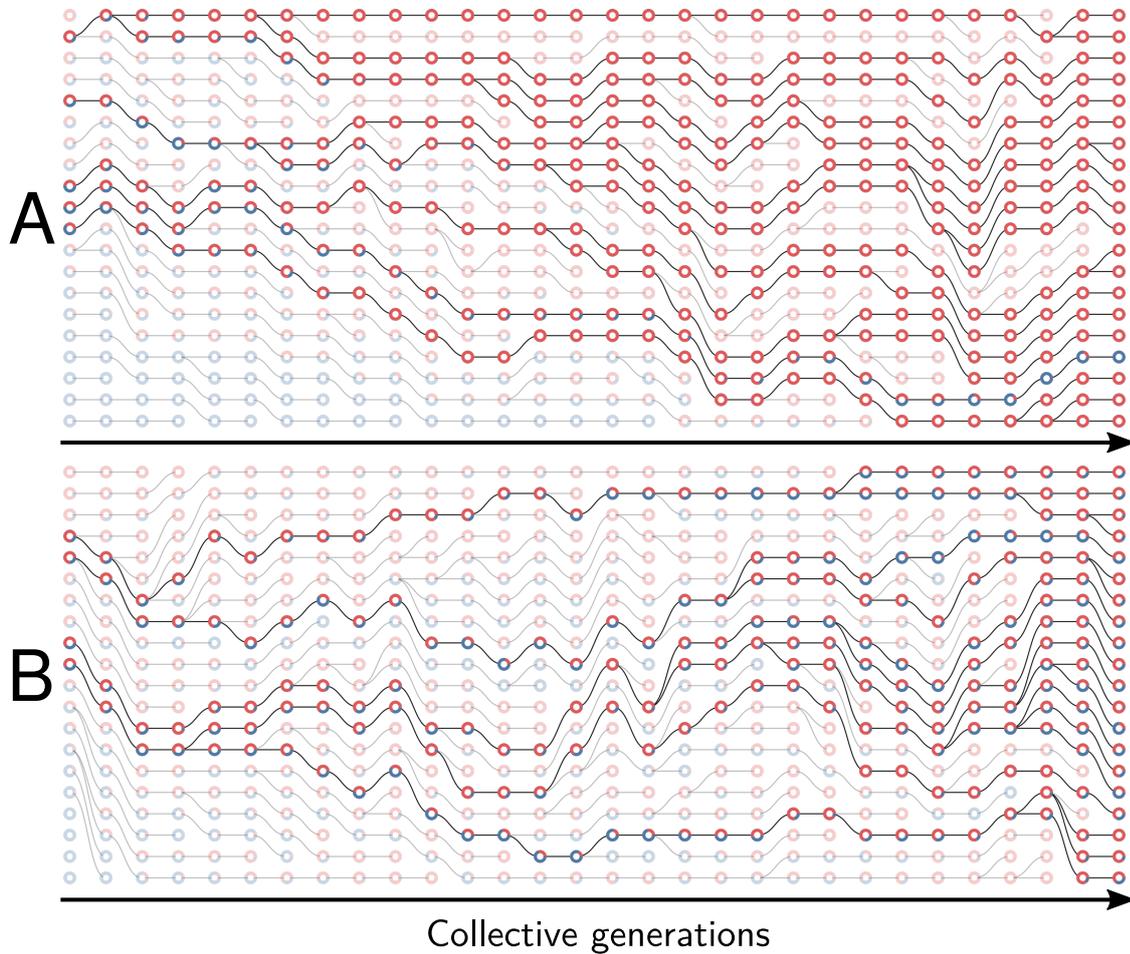


Figure 6.10: **Example of collective genealogy** (Supplement of Figure 6.2) Symbols and colours are as in Appendix 6.3 Figure 6.9 and extinct lineages are marked transparent. Collective-level parameters in this simulation are  $M = 30$ ,  $D = 20$ ,  $\rho = 0.1$ . *A*. Neutral regime: at the final generation, collectives are monochromatic and most likely composed of the faster-growing type. *B*. Selective regime: at the final generation, collectives contain both red and blue particles.

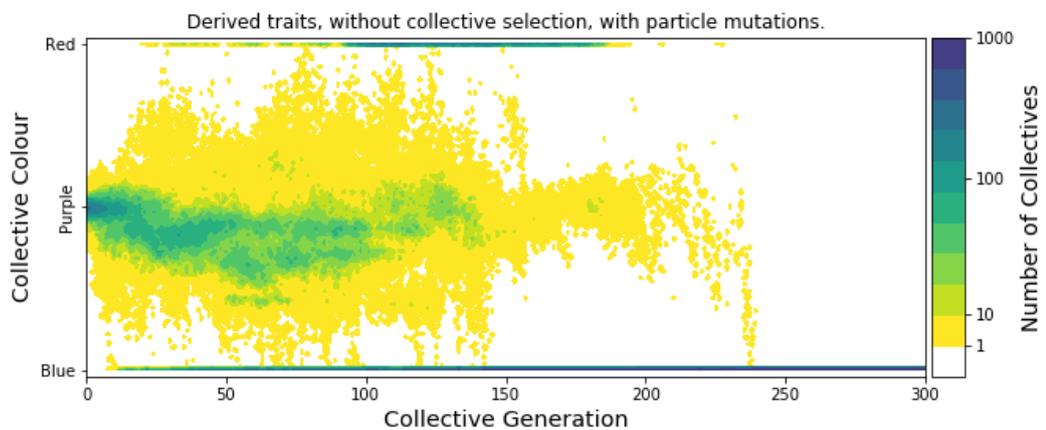


Figure 6.11: **Particle trait mutations lead to slow loss of optimal collective colour after removal of collective-level selection.** (Supplement of Figure 6.4) Modification of the collective phenotype distribution when particle traits mutate and no selection for colour is applied, starting from the particle traits after 10,000 generations of selection for purple colour (as in Figure 6.4B). Collectives continue to produce purple offspring for more than 200 generations, before drift of particle traits erodes developmental correction. In contrast, for the ancestral particle traits lineages become monochromatic in less than 10 generations (Figure 6.4A).

## 6.4 Lotka-Volterra deterministic particle ecology

### 6.4.1 Model for intra-collective dynamics

The stochastic ecological dynamics of red ( $N_0$ ) and blue ( $N_1$ ) particles within a collective simulated by the algorithm described in Appendix 6.3 is approximated (see Figure 6.3) by the deterministic competitive Lotka-Volterra Ordinary Differential Equation:

$$\begin{cases} \frac{dN_0}{dt} = r_0 N_0 \left(1 - \frac{a_{00}}{K} N_0 - \frac{a_{01}}{K} N_1\right) \\ \frac{dN_1}{dt} = r_1 N_1 \left(1 - \frac{a_{10}}{K} N_0 - \frac{a_{11}}{K} N_1\right) \end{cases} \quad (6.1)$$

Here,  $\mathbf{r} = (r_0, r_1)$  is the pair of maximal growth rates for red and blue particles. Since the system is symmetric, we consider only the case where red particles grow faster than blue particles ( $r_0 > r_1$ ). The effect of pairwise competitive interactions between cells are encoded in the matrix  $\mathbf{A} = (a_{ij})_{i,j \in \{0,1\}^2}$ . All competitive interactions are considered harmful or neutral ( $0 \leq a_{ij}$ ). No specific mechanism of interaction is assumed so as to explore the space of all possible interaction intensities, and thus qualitatively different ecological dynamics of the particle populations. Evolutionary trajectories constrained by particle traits are briefly discussed in the last section of the Results (main text).

The carrying capacity of a monochromatic collective is  $\frac{K}{a_{00}}$  for red particles and  $\frac{K}{a_{11}}$  for blue particles.  $K$  is a scaling factor for the intensity of pairwise interactions that can be used to rescale the deterministic system to match the stochastic trajectories. Without loss of generality, we thus assume that  $K = 1$ .

A natural set of alternate coordinates for the system in Equation 6.1 are total population size  $N := N_0 + N_1$  and collective colour, defined as the frequency of red individuals  $x := \frac{N_0}{N}$ . In these coordinates the deterministic dynamics are the solution to the following ODE.

$$\begin{cases} \frac{dN}{dt} = Ng(x, N) \\ \frac{dx}{dt} = x(1-x)h(x, N) \end{cases} \quad (6.2)$$

The functions  $g$  and  $h$  are polynomials in  $x$  and  $N$ , of coefficients:

Monomial	Coefficient in $h$	Monomial	Coefficient in $g$
1	$r_0 - r_1$	1	$r_1$
$N$	$a_{11}r_1 - a_{01}r_0$	$N$	$-a_{11}r_1$
$Nx$	$r_1(a_{10} - a_{11}) + r_0(a_{01} - a_{00})$	$x$	$r_0 - r_1$
		$xN$	$r_1(2a_{11} - a_{10}) - r_0a_{01}$
		$x^2N$	$r_0(a_{01} - a_{00}) + r_1(a_{10} - a_{11})$

Appendix 6.4 Figure 6.12 shows an example of the ODE flow in both coordinate systems. Two unstable trivial equilibria and a stable coexistence equilibrium are located at the intersection of the isoclines. Within one collective generation, the dynamics follow such flows for duration  $T$ , starting from initial conditions on the line  $N_0 = B - N_1$  ( $N = B$ ).

Even though Equation 6.1 is more directly related to the individual-based stochastic simulation, the dynamics of collective colour are understood more easily using Equation 6.2. Therefore, in the following we will use the latter formulation.

The dynamics of particle types across collective generations are modelled as a piecewise continuous time change (Appendix 6.4 Figure 6.13), where  $x^m(t)$  is the fraction of red particles at time  $t \in [0, T]$  during collective generation  $m$ .

We focus in particular on the succession  $N^m(T), x^m(T)$  of collective ‘‘adult’’ states at the end of each successive generation  $m$ . In the following we note  $\phi^m = x^m(T)$ , which is the adult colour of the collective at the end of the growth phase. At the beginning of each collective generation, we impose that, regardless of the number of cells in the parent, every collective contains the same number of cells  $N^m(0) = B \in \mathbb{R} \forall m$ . In contrast, the newborn collective colour  $x^m(0)$  depends on the colour of the parent. In this deterministic model, we consider that there is no stochastic

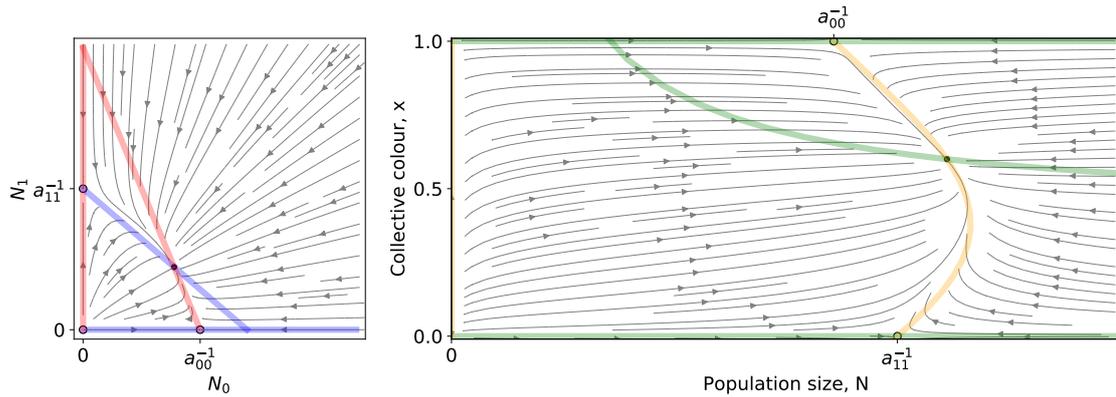


Figure 6.12: **Left:** A Lotka-Volterra flow (Equation 6.1 ) in  $(N_0, N_1)$  coordinates. **Right:** The same flow in  $(N, x)$  coordinates. Coloured lines are the null isoclines. Empty (resp. filled) circles mark unstable (resp. stable) equilibria.

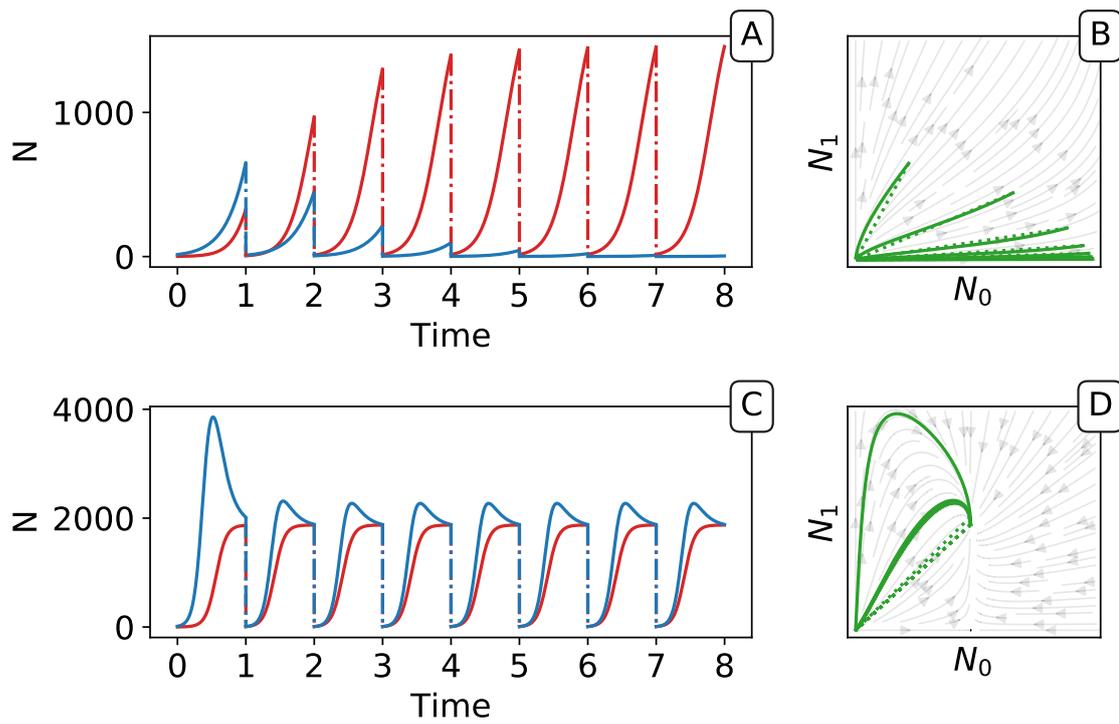


Figure 6.13: **Piecewise continuous trajectory.** Iterating the deterministic model yields a piecewise continuous trajectory in the  $(N_0, N_1)$  space. The growth phase (continuous lines) alternates with the dilution (dotted lines). In (A, B) traits are taken from generation 3 and in (C,D) from generation 9,000 of the main simulation with selection (Figure 6.2). Successive adult states can be computed using the  $G_\theta$  function as a recurrence map (see Figure 6.5 in the main text).

variation or bias due to sampling at birth, so the collective colour of the newborn is equal to that of its parent:  $x^m(0) = x^{m-1}(T) = \phi^{m-1}$ . In the first generation ( $m = 0$ ), the population size is  $B$  in every collective and the fraction of red particles is chosen uniformly at random between 0 and 1.

The proportion of red particles at adult stage  $\phi^m$  is obtained from these initial conditions  $(B, \phi^{m-1})$  by integrating Equation 6.2. Since these equations are not explicitly solvable, there is no analytic expression for the result of the transient dynamics. However, having constrained the initial conditions to a single dimension, the adult colour is a single-valued function of the initial composition of the collective, defined as follows.

**Definition 1 (Growth function):**

Given the set of particle traits  $\theta := (\mathbf{r}, \mathbf{A}) \in E := [0, \infty)^2 \times [0, \infty)^4$ , the bottleneck size  $B \in (0, \infty)$  and the duration of the growth phase  $T \in (0, \infty)$ , the growth function  $G_\theta$  is defined as the application that maps an initial proportion of red particles  $\phi$  to the proportion of red particles after duration  $T$ . Thus,  $G_\theta(\phi, T, B) = x(T)$ , with  $x(t)$  is the unique solution to the following Cauchy problem:

$$\begin{cases} \frac{dN}{dt} &= Ng(x, N) \\ \frac{dx}{dt} &= x(1-x)h(x, N) \\ N(0) &= B \\ x(0) &= \phi \end{cases} \quad (6.3)$$

To simplify notations in the main text, we set  $G(\phi) = G_\theta(\phi, B, T)$ .

### 6.4.2 Relation with the stochastic evolutionary model

As explained in the main text, the growth function  $G_\theta$  defines the recurrence relation between the colour of the newborn offspring and its colour at adulthood. If the reproduction process entails no stochasticity, the latter also defines the composition of collectives at the next generation. As a consequence, the iterative application of  $G_\theta$  approximates the change in time of the adult colour when a population with a given, fixed, set of parameters and traits is transferred across collective generations.

The fixed points of this discrete-time system allow understanding and classification of behaviours observed along stochastic evolutionary trajectories. When mutation rate of particle traits is sufficiently small and growth rates are not vanishing, collectives approach such fixed points in a few collective generations by iteration of the  $G_\theta$  function (Figure 6.5). The evolutionary trajectory can thus be seen as a succession of fixed points of  $G_\theta$ . The surface of the fixed points of  $G_\theta$  as a function of the particle parameters can be computed numerically, as well as its dependence on the collective generation duration  $T$  and bottleneck size  $B$ . A small number of qualitatively different configurations are possible for the fixed points (illustrated in Appendix 6.4 Figure 6.14 1-4). In particular, the case in which  $G_\theta$  possesses an internal, stable fixed point (Appendix 6.4 Figure 6.14) constitutes the optimal solution to constant selection for collective colour, in that it ensures the highest degree of colour reproducibility, on average, across collective generations.

The parameter values that separate regions with qualitatively different fixed points correspond to transcritical bifurcations, where one of the monochromatic fixed points changes its stability. These lines are analytically computed below, thus allowing generalisation of the conclusions drawn from analysing the representative trajectory of Figure 6.2. In the following, we detail analysis of how the fixed points of the  $G_\theta$  function depend on particle- and collective-level parameters.

It is worth stressing here that the deterministic model provides a good quantitative approximation of the system with particle-level and collective-level stochasticity provided that fluctuations at both levels are small. This is the case if populations of particles are large (as for instance in the case of bacterial populations) and if mutations of particle traits are rare and of small magnitude. However, in the numerical simulations we performed, the conclusions drawn from ensuing analysis held qualitatively also in the case of large fluctuations.

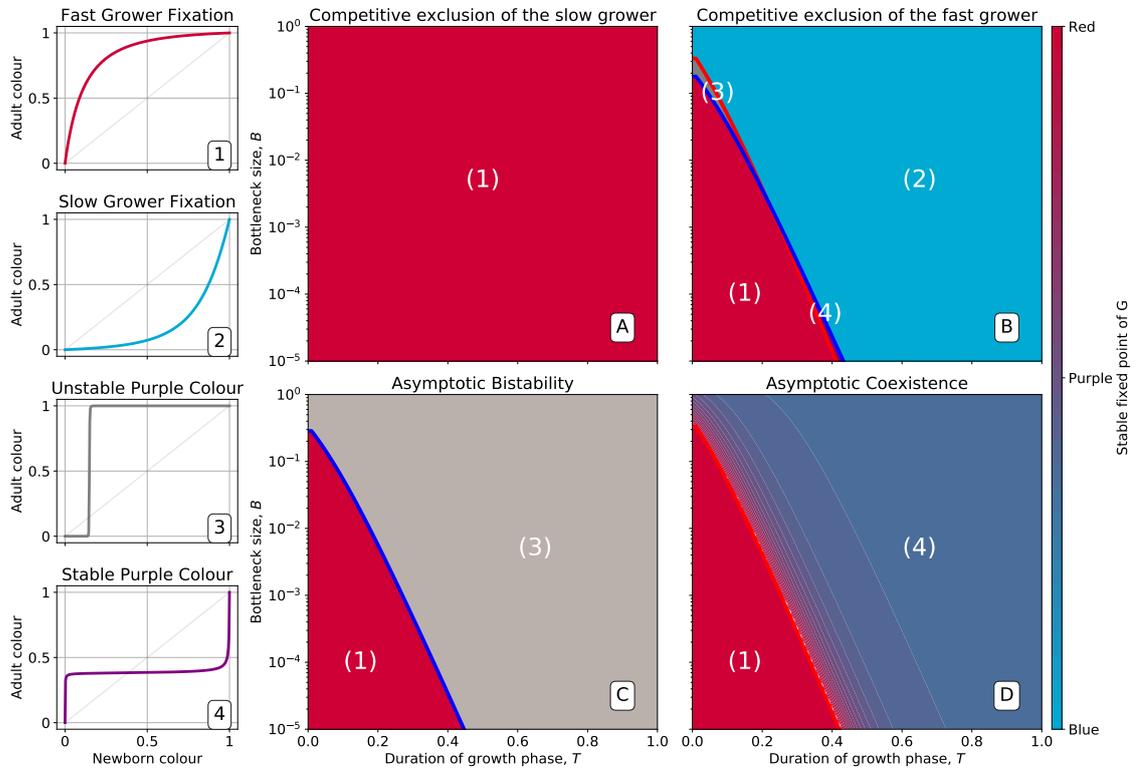


Figure 6.14: **Qualitative behaviours of the growth function  $G_\theta$ .** Panels 1-4 represent the possible qualitative shapes, differing in the position and stability of the fixed points, of the growth function  $G_\theta$  that approximates the within-collective particle dynamics (orange line in Figure 6.5). 1)  $\phi = 1$  is the only stable fixed point (figure 6.5A), and iteration of the red  $G_\theta$  function leads to fixation of red particles. 2)  $\phi = 0$  is the only stable fixed point, and iteration of the blue  $G_\theta$  function leads to fixation of blue particles. 3)  $\phi = 0$  and  $\phi = 1$  are two stable fixed points, and iteration of the grey  $G_\theta$  function leads to fixation of either red or blue particles depending on initial conditions. 4)  $\phi = 0$  and  $\phi = 1$  are unstable and there is a stable fixed point between 0 and 1. Iteration of the grey  $G_\theta$  function leads to coexistence of both particle types. Panels A-D show that when red particles are the fast growing types ( $r_0 > r_1$ ), the shape of  $G_\theta$  and the position of its fixed points depend on the collective-level parameters  $B$  (bottleneck size) and  $T$  (growth phase duration). Particle interaction traits generically belong to one of the four intervals A)  $a_{01} < a_{11}$  and  $a_{00} < a_{10}$ ; B)  $a_{11} < a_{01}$  and  $a_{10} < a_{00}$ ; C)  $a_{11} < a_{01}$  and  $a_{00} < a_{10}$ ; D)  $a_{01} < a_{11}$  and  $a_{10} < a_{00}$  (qualitative nature of the corresponding ecological equilibria is indicated in the titles of the panels, see also Appendix 6.4 Table 6.1). Lines represent the limit of the region of stability of the fixed point of  $G_\theta$ , as derived by Proposition 26: blue lines for the "all blue" state  $\phi = 0$  and red lines for the "all red" state  $\phi = 1$ .

### 6.4.3 Fixed Points of the $G_\theta$ function and their stability

In this paragraph, we list the key properties of the  $G_\theta$  function, that determine how the asymptotic collective colour depends on particle and collective parameters. Proofs of the propositions are provided in the following paragraph.

**Proposition 23 (Fixed points of  $G_\theta$ ):**

Let  $\theta \in E$  be a set of particle traits,  $T \in (0, \infty)$  the duration of the growth phase and  $B \in (0, \infty)$  the bottleneck size.

Then,  $G_\theta(0, T, B) = 0$ , and  $G_\theta(1, T, B) = 1$ , hence  $\phi = 0$  and  $\phi = 1$  are fixed point of  $G_\theta \forall \theta$ .

Moreover, if the stability with respect to  $\phi$  of 0 and 1 is the same, then  $G_\theta$  has at least one fixed point  $\phi \in (0, 1)$ .

The stability of the monochromatic fixed points with respect to the fraction  $\phi$  can be numerically assessed. The number and stability of the fixed points as a function of collective-level parameters  $B$  and  $T$  are illustrated in Appendix 6.4 Figure 6.14. Four different cases are considered, corresponding to the four qualitatively different outcomes of particle-level ecology (competitive exclusion by one or the other type, bistability, coexistence).

The fixed points can be analytically calculated in certain limit cases, corresponding to parameter values when within-collective particle dynamics are exponential or saturating.

**Proposition 24 (Quasi-exponential growth):**

When  $T$  is close to 0 and  $Ba_{max} \ll 1$  with  $a_{max} = \max_{ij} a_{ij}$  the highest element of the competition matrix  $\mathbf{A}$ , then  $G_\theta$  has only two fixed points  $\phi = 0$  and  $\phi = 1$ .

Moreover, the monochromatic fixed point  $\phi = 1$ , corresponding to a population completely composed of the (faster) red type of particle, is stable, whereas the fixed point  $\phi = 0$ , corresponding to a population composed of the slow growing type of particle, is unstable.

**Proposition 25 (Saturating growth):**

As  $T \rightarrow \infty$  the fixed points  $\phi^*$  of  $G_\theta$  and their stability correspond to the equilibria  $x^*$  of Equation 6.1. The following Table 25 lists these equilibria and their stability range.

	Red alone	Blue alone	Coexistence
Equilibrium $x^*$	1	0	$\frac{a_{11} - a_{01}}{Tr(\mathbf{A}) - CoTr(\mathbf{A})}$
Stability range	$a_{00} < a_{10}$	$a_{11} < a_{01}$	$a_{11} > a_{01}$ and $a_{00} > a_{10}$

Here  $Tr(\mathbf{A}) := a_{00} + a_{11}$  is the sum of the diagonal (or trace) of  $\mathbf{A}$ , and  $CoTr(\mathbf{A}) := a_{10} + a_{01}$  denotes the sum of the anti-diagonal elements of  $\mathbf{A}$ .

By linearising the system in proximity of the fixed points it is possible to find exactly the bifurcation parameters where one equilibrium changes stability, thus the limit of the region where there exists an interior fixed point  $\phi^*$ . The bifurcation values in the space of the collective

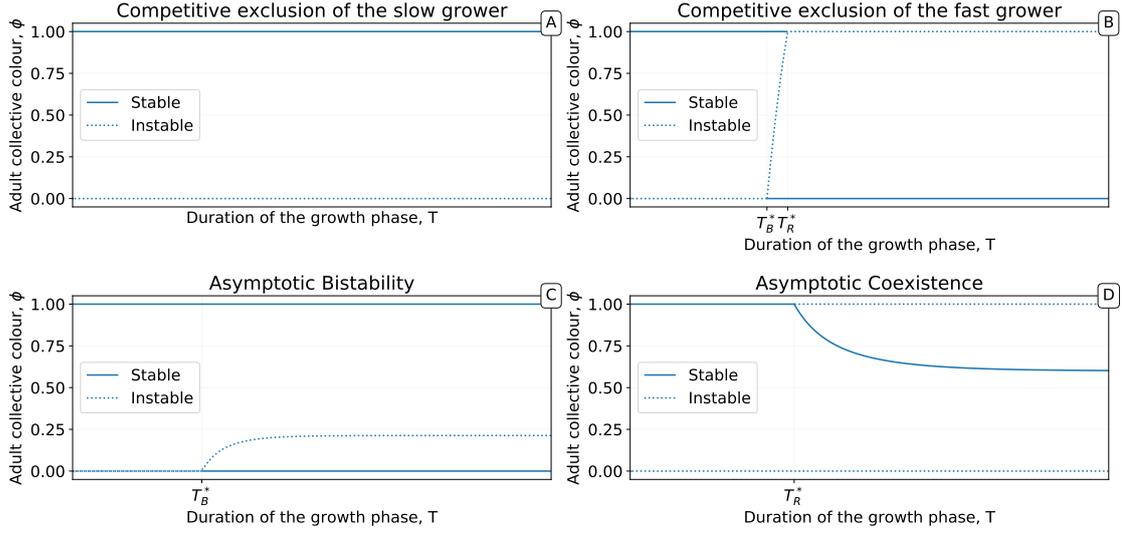


Figure 6.15: **Bifurcation diagrams showing the position and stability of the fixed points of  $G_\theta$  as a function of the duration of the collective generation  $T$ .** A-D particle traits are representative of the scenarios illustrated in Figure 6.7 A-D. Of particular interest is the case illustrated in panel D, where the  $G_\theta$  function acquires — for a sufficiently large separation between the particle maximum division time and the collective generation time — a stable internal fixed point.

parameters  $T$  and  $B$  delimit the region in Figure 6.7 where the  $G_\theta$  function has an internal fixed point.

Even when the fixed point is different from the optimal value  $\hat{\phi}$ , it can nonetheless provide a starting point for evolution to optimize collective colour. Extinction of one of the two colours of particles happens instead very rapidly in the region when the monochromatic fixed points are stable, so that collectives have a higher risk of being extinct before inheritance-increasing mutations appear.

**Proposition 26 (Bifurcations of the monochromatic fixed points of  $G_\theta$ ):**

*The stability of 0 changes at  $(T_B^*, B_B^*)$  and the stability of 1 at  $(T_R^*, B_R^*)$  such that:*

$$B_R^* = \frac{e^{-\alpha_1 T_R^*} - e^{r_0 T_R^*}}{a_{00}(1 - e^{-\alpha_1 T_R^*})} \quad \text{with } \alpha_1 = r_0 r_1 \frac{a_{00} - a_{10}}{r_0 a_{00} - r_1 a_{10}}$$

$$B_B^* = \frac{e^{-\alpha_0 T_B^*} - e^{r_1 T_B^*}}{a_{11}(1 - e^{-\alpha_0 T_B^*})} \quad \text{with } \alpha_0 = r_0 r_1 \frac{a_{11} - a_{01}}{r_1 a_{11} - r_0 a_{01}}$$

These results allow understanding of the interplay between time scales of particle-level ecology and collective reproduction, whose relationship changes along an evolutionary trajectory. Appendix 6.4 Figure 6.15 illustrates the change of the fixed point of  $G_\theta$  with the collective generation duration  $T$  for typical particle traits corresponding to the four qualitative classes of asymptotic equilibria for particle ecology. Of particular relevance for understanding the stochastic trajectory illustrated in Figure 6.2 is panel D.

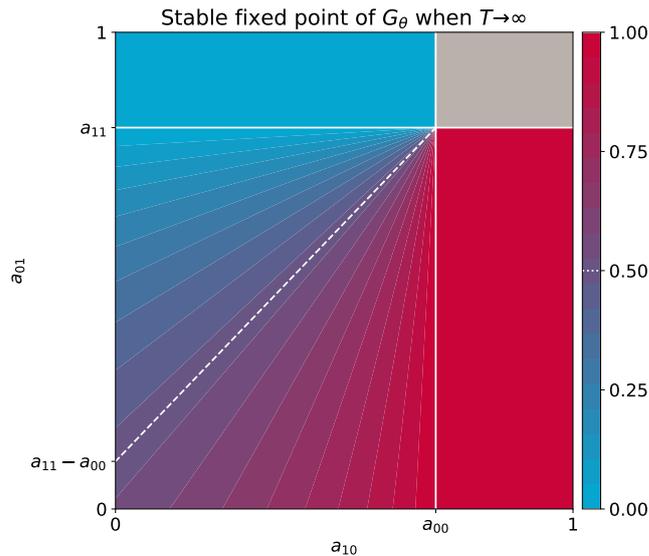


Figure 6.16: **Stable equilibria of the particle ecology as a function of inter-colour competition traits.** These equilibria correspond to the limit of the fixed points of the  $G_\theta$  function when particle-level and collective-level time scales are well separated ( $r \gg \frac{1}{T}$ ), derived in proposition 25. Other interaction parameters are  $a_{00} = 0.7$  and  $a_{11} = 0.8$ , and the result is independent of growth rates. The grey area indicates bistability.

When the time scale of exponential particle growth is comparable to  $T$ , (such as at the beginning of the evolutionary trajectory displayed in Figure 6.2) Proposition 24 indicates that the system is expected to converge to an all-red solution ( $\phi = 1$  is the only stable equilibrium). However, when these fast dynamics occur, the growth rates change by mutation. Selection during the exponential phase generally favours fast growing mutants, which means that particle populations achieve high-density conditions in a shorter time. The system then effectively behaves as if  $T$  had increased, thus leading selection to 'see' interaction traits.

When the time scale of collective reproduction is sufficiently slow with respect to the intra-collective dynamics, the system crosses the bifurcation point  $T_R^*$  (Proposition 26), so that the function  $G_\theta$  now has an internal fixed point (Figure 6.6). In the stochastic simulations, this means that more collectives are reproducibly found close to the optimal colour. It takes a relatively short time to adjust the particle traits so that the fixed point is close to the optimum  $\hat{\phi}$ . In this case, the deterministic approximation produces a close to perfect inheritance of the collective colour. However, fluctuations in particle numbers and in composition at birth still result in a large variance of colours among collectives in the stochastic system.

Here starts the last and slowest phase of the evolutionary trajectory, which results in colour variance reduction through improvement of the ability of particles to correct variations in colour. This is achieved by attaining faster the particle ecological equilibrium, so that fluctuations are more efficiently damped by demographic dynamics. As a consequence, the conditions described by Proposition 25 will be met. This allows identification of a surface in parameter space, where the fixed point of the  $G_\theta$  function  $\phi^*$  identifies with the ecological equilibrium  $x^*$ , that contains evolutionary equilibria. The ecological equilibrium is displayed in Figure 6.16 as a function of the cross-colour interaction parameters  $a_{01}$  and  $a_{10}$ . In the regime where particle and collective time scales are well separated ( $r \gg 1/T$ ), then interaction parameters that correspond to the optimal colour satisfy the following relationship:

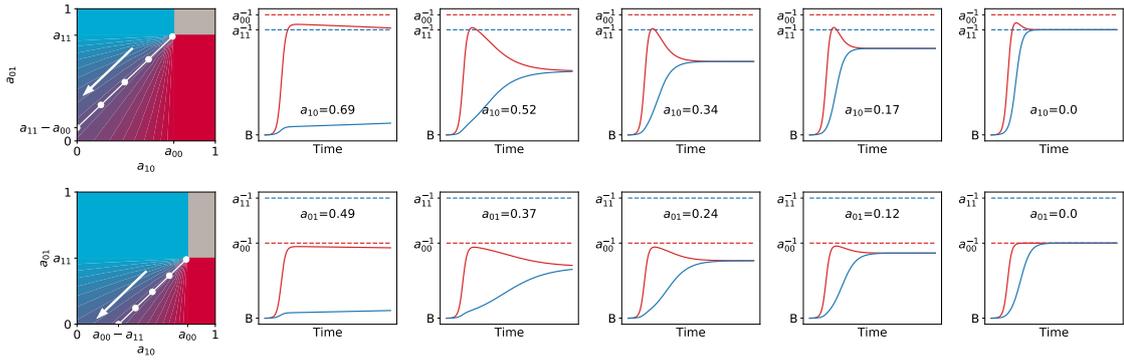


Figure 6.17: **Asymmetric interaction ensures fastest convergence toward the ecological equilibrium.** Particle population dynamics are illustrated for increasingly asymmetric values of  $(a_{01}, a_{10})$ , keeping the ecological equilibrium fix at  $x^* = 0.5$  (along white manifold, in the direction of the arrow). The top panels correspond to cases when the faster-growing particles have a higher carrying capacity  $((a_{00}, a_{11}) = (0.7, 0.8))$ , the bottom panels to the opposite  $((a_{00}, a_{11}) = (0.8, 0.5))$ . In both cases  $r_0 = 60$ ,  $r_1 = 40$ ,  $T = 1$ ,  $B = 0.001$ .

$$a_{00} - a_{10} + \left(1 - \frac{1}{x^*}\right) (a_{11} - a_{01}) = 0.$$

This relation identifies the white dotted line in Appendix 6.4 Figure 6.16 (and in Figure 6.8B). Once it is attained, mutations cause the deterministic system to move neutrally on this surface. As the stochastic simulation shows, particle parameters keep evolving directionally so as to reduce phenotypic variance. This is achieved by making  $G_\theta$  increasingly flatter in the vicinity of the fixed point, so that the target colour is not only more stable, but it is reached in fewer collective generations.

Successive events of mutation and substitution progressively lead to a growing asymmetry in the ecological relationship between the two types of particles: that with smaller carrying capacity becomes insensitive to the other colour; the latter instead experiences competition, so that its growth is curbed and optimal proportion of colours is eventually realized (see Figure 6.3D). As illustrated by Appendix 6.4 Figure 6.17, this conclusion is independent of what is the maximum growth rate, that only affects the advantage of one type at initial stages of growth. Indeed, the position on the intercept between the optimality line and the y axis in Appendix 6.4 Figure 6.16 and Appendix 6.4 Figure 6.17 only depends on the difference between intra-type competition parameters. A consequence of this is that fast-growing types systematically display, when they are also those with a larger carrying capacity, a population overshoot.

#### 6.4.4 Proofs

In this section we present the proof of propositions 24 - 26 above.

##### Proof (Proof of Proposition 23):

In the  $(N, x)$  coordinates we have seen that  $\frac{df(t)}{dt} = g(x(t), N(t))$  and that  $x(t) = 0$  and  $x(t) = 1$  are trivial roots of the polynomial  $g$  (Equation 6.2). Hence 0 and 1 are always fixed points of  $G_\theta$ .

Since  $g$  and  $h$  (from Equation 6.2) are polynomials of  $(N, x)$ , they are smooth (of Differentiability class  $C_\infty$ ) on  $\mathbb{R}^2$ . Thus, the global flow corresponding to the Cauchy problem is also smooth on  $\mathbb{R}^2$ .  $G_\theta$  is the partial application of the global flow to the case where  $N_0 = B$ . Therefore,  $G_\theta$  is continuous on  $[0, 1]$ .

Moreover, let us suppose that 0 and 1 are both unstable. Then  $G'_\theta(0, T, B) > 1$  and  $G'_\theta(1, T, B) > 1$ , with  $G'_\theta$  the derivative of  $G_\theta$  with respect to its first variable. As a consequence there is an  $\varepsilon \in \mathbb{R}$  such that  $G_\theta(\varepsilon, T, B) > \varepsilon$  and  $G_\theta(1 - \varepsilon, T, B) < 1 - \varepsilon$ . Since  $G_\theta$  is continuous, there is at least one  $c \in [0, 1]$  such that  $G_\theta(c, T, B) = c$  by virtue of the intermediate value theorem.

In practice, we never encountered cases when more than one internal fixed point was present. However multistability is expected to occur if the equations describing particle ecology had higher-order nonlinearities.

**Proof (Proof of Proposition 24):**

Since the nonlinear terms in Equation 6.3 are smaller than  $Ba_{\max}$ , and this is negligible with respect to 1, particle ecology is approximated by its linearization as long as the population size remains close to the bottleneck value. Around  $t = 0$ , the Cauchy problem can be written as:

$$\begin{cases} \frac{dN_0}{dt} = r_0 B x \\ \frac{dN_1}{dt} = r_1 B (1 - x) \\ N_0(0) = x B \\ N_1(0) = (1 - x) B \end{cases}$$

In the  $(N, x)$  coordinates, the total population size grows exponentially and is decoupled from the colour. On the other hand,  $x(t)$  follows the logistic differential equation:

$$\frac{dx}{dt} = \frac{d}{dt} \frac{N_0}{N_0 + N_1} = (r_0 - r_1)x(1 - x) \quad (6.4)$$

which can be integrated.

For  $T$  sufficiently small for the population to be in exponential growth phase, the growth map  $G_\theta$  can be approximated by the solution  $\tilde{G}_\theta$  of Equation 6.4:

$$G_\theta(x, T, B) \approx \frac{1}{1 + \left(\frac{1}{x} - 1\right) e^{-(r_0 - r_1)T}} := \tilde{G}_\theta(x, T, B)$$

This function is strictly convex (or concave) on  $(0, 1)$  depending on the sign of  $r_0 - r_1$ :

$$\frac{\partial^2 \tilde{G}_\theta(x, T, B)}{\partial x^2} = \frac{2 [1 - e^{(r_0 - r_1)t}]}{(x e^{(r_0 - r_1)t} + 1 - x)^3}$$

Since  $0 < x < 1$  and  $t > 0$ ,  $\frac{\partial^2 \tilde{G}_\theta(x, T, B)}{\partial x^2}$  is of the same sign as  $1 - e^{t(r_0 - r_1)}$ , that is strictly positive if  $r_1 > r_0$ , or strictly negative if  $r_0 > r_1$ .

Thus,  $\tilde{G}$  is strictly convex on  $(0, 1)$  if  $r_1 > r_0$ , and strictly concave on  $(0, 1)$  if  $r_0 > r_1$ . In the first case, red colour  $x = 0$  is an unstable equilibrium and blue colour  $x = 1$  a stable one, and vice-versa in the second case. Note that the segment  $s = [(0, 0), (1, 1)]$  is a chord of  $\tilde{G}$ . Therefore, the strictly convex (resp. concave)  $\tilde{G}$  do not intersect  $s$  except in  $(0, 0)$  and  $(1, 1)$ .

**Proof (Proof of Proposition 25):**

When the collective generation time  $T$  is much longer than the demographic time scale, the populations within droplets at the adult stage are well approximated by the equilibrium solution of the Lotka-Volterra Equation 6.1. Solving simultaneously equations  $\frac{dN_0}{dt} = 0$  and  $\frac{dN_1}{dt} = 0$  (or equivalently  $\frac{dN}{dt} = 0$ ,  $\frac{dx}{dt} = 0$ ) yields the four equilibria listed in table 6.1 in both coordinate systems. Linear stability analysis allows one to determine the parameter intervals where these are stable, listed in Appendix 6.4 Table 6.1.

Name	Red alone	Blue alone	Coexistence	Extinction
Position in $[N_0, N_1]$	$[\frac{1}{a_{00}}, 0]$	$[0, \frac{1}{a_{11}}]$	$[\frac{a_{11}-a_{01}}{\det(A)}, \frac{a_{00}-a_{10}}{\det(A)}]$	$[0, 0]$
Position in $[N, f]$	$[\frac{1}{a_{00}}, 1]$	$[\frac{1}{a_{11}}, 0]$	$[\frac{\text{Tr}(A)-\text{CoTr}(A)}{\det(A)}, \frac{a_{11}-a_{01}}{\text{Tr}(A)-\text{CoTr}(A)}]$	Undefined
Condition for stability	$a_{00} < a_{10}$	$a_{11} < a_{01}$	$a_{11} > a_{01}$ and $a_{00} > a_{10}$	Never

Table 6.1: **Equilibria of the particle populations and their stability.** Ecological equilibria of the Lotka-Volterra system are listed in the two coordinate systems. Here,  $\text{Tr}(\mathbf{A})$  and  $\text{CoTr}(\mathbf{A})$  are defined as in the caption of Table 25.

**Proof (Proof of Proposition 26):**

We consider the case when the fixed point 0 changes stability. The stability of the fixed point 1 can be studied analogously.

We aim at identifying the values of the collective parameters  $(T, B)$  where a fixed point with  $x = 0$  changes stability through a transcritical bifurcation. The difficulty lies in the fact that one needs to estimate the dynamics of the second variable  $N$  in order to study the stability of the 2-D system. Luckily, this can be done in the limit when the collective contains almost exclusively particles of one single colour (in this case, blue).

Total population size is in this case decoupled from colour and Equation 6.2 can be integrated with initial condition  $(N_0 x_0) = (B, 0)$ , yielding the following trajectory:

$$\tilde{N}(t, B) = \frac{1}{a_{11} - (a_{11} - \frac{1}{B}) e^{-r_1 t}}$$

As long as  $x$  is small, the time derivative is approximated by the non-autonomous system:

$$\frac{dx}{dt} \approx xh(x, \tilde{N}(t, B))$$

with  $h$  as defined in Equation 6.2:

$$h(0, \tilde{N}(t, B)) = (r_0 - r_1) + \tilde{N}(t, B)(a_{11}r_1 - a_{01}r_0)$$

Solving this equation allows computation of the adult colour as a function of the parameters  $T$  and  $B$ . At the bifurcation point  $(T^*, B^*)$ , where stability of the 0 fixed point changes, the newborn colour is the same as the adult colour:  $x(T^*, B^*) = x(0, B^*)$ .  $(T^*, B^*)$  are then solutions of the integral equation:

$$\begin{aligned} 0 &= \int_0^{T^*} h(0, \tilde{N}(s, B^*)) ds \\ 0 &= (r_0 - r_1) + (a_{11}r_1 - a_{01}r_0) \int_0^{T^*} \tilde{N}(s, B^*) ds \\ 0 &= T^* r_0 \left(1 - \frac{a_{01}}{a_{11}}\right) + \left(1 - \frac{r_0 a_{01}}{r_1 a_{11}}\right) \ln \left(\frac{B^* a_{11} + e^{-r_1 T^*}}{B^* a_{11} + 1}\right) \end{aligned}$$

Solving for  $B^*$  we get:

$$B^* = \frac{e^{-\alpha T^*} - e^{r_0 T^*}}{a_{00}(1 - e^{-\alpha T^*})}$$

With  $\alpha = r_0 r_1 \frac{a_{00} - a_{10}}{r_0 a_{00} - r_1 a_{10}}$

Figure 6.7 shows that this approximation retrieves accurately the numerically computed bifurcation lines.

# Conclusion

“This account of Darwinism yields a particular picture of the world. One of the world’s constituents is a great range of Darwinian populations: paradigm cases and marginal ones, some clear and other obscures, some powerful and other suppressed. Some are visible and obvious, other invisible. Some are inside others. They tread through their Darwinian behaviors on a great range of different scales in space and time.”

— PETER GODFREY-SMITH, *Darwinian Populations and Natural Selection* (2009)

THIS MANUSCRIPT AIMED to explore the mechanisms by which collectives of entities might become the subject of evolution by natural selection in their own right. This objective required definition of a minimal set of properties a population of entities must exhibit to partake in evolution by natural selection. These abstract Darwinian properties are the current iteration of a long-going process of formalising that set apart living beings from inert matter (Chapter 1). Life is organised in several nested levels that can exhibit Darwinian properties to a diverse degree. Indeed, Darwinian and non Darwinian are not just binary categories, it is possible to formalise a spectrum of intensity from marginal to paradigmatic Darwinian. Understanding how new Darwinian levels come to be requires mechanisms that cannot just be reduced to a simple transfer of Darwinian qualities from the lower to the higher level: natural selection cannot be invoked as its own cause. One solution to this apparent difficulty is the exogenous imposition of marginal Darwinian properties by certain and particular population structures — ecological scaffolding (Chapter 2).

Ecological scaffolding requires the presence of a nested birth-death process: both particles and collectives must have their own intertwined demography. This population structure has consequences for neutral diversity. Using stochastic models of an experimental setup, I showed that the duration of the collective-level generation as well as the bottleneck size can be tweaked to optimise the total number of mutations, or other measurements of diversity (Chapter 3). Moreover, I developed a Bayesian-network approach to help visualise and interpret collective-level genealogies established in experimental evolution. This approach can also be used as a tool to plan further, iterative, experiments (Chapter 4).

Inspired by experimental and theoretical work conducted on the origin of multicellularity, I used adaptive dynamics to study the trade-off survival-

reproduction of collectives in nested populations. I illustrated how collective-level selection can influence the evolutionary dynamics of particles traits. When I decided to include the demography of collectives within their environment, I obtained equations with equivalents in classical epidemiological models, and derived very simple conditions on the minimal niche density that required by ecological scaffolding (Chapter 5).

Finally, I explored the problem of collective-level heredity when collectives functions are ensured by the presence of two different kinds of organism. The maintenance of a viable collective composition generation after generation can be achieved either by producing numerous noisy offspring, some of which will statistically exhibit the viable composition (stochastic corrector), or by specific mechanisms (such as particle interactions) ensuring the reliable re-establishment of the parental state (developmental corrector). I showed that, within my models, there is an evolutionary path from one to the other. This constitutes a prime example of ecological scaffolding in the sense that population structure imposes a marginal level of heredity (noisy offspring), that can be refined by natural selection into a level-specific developmental mechanism (Chapter 6).

Overall, I argue that the study of Darwinian properties at different organisational levels, and their potential origin by ecological scaffolding, constitutes a fruitful entry point in the study of the evolutionary processes. This framework allows formalisation of a wide range of basic questions in the domain of Major Evolutionary Transitions, from exploring the mechanisms at the origin of complex integrated structures such as multicellular organisms, to the origin of life itself. An undeniable strength of this approach is its close relation to experimental evolution. Observing a major evolutionary transition might be a daunting task. However, being able to impose in a controlled environment the conditions that are thought to be necessary to promote them gives invaluable insight, as illustrated by the many studies cited in this manuscript.

In my opinion, there are many promising directions in which the theoretical work presented here could be extended. I present three here. Firstly, the toy model presented in Chapter 2 could be built upon to try to provide a mechanistic basis for the quantification of Darwinian properties. Indeed, I think that there is a need to provide methods to measure and quantify the dimension of Darwinian space in both experiments and models. Rooting this method in a mechanistic description of population dynamics would constitute a solid base for understanding multi-level phenomenon in their complexity. Secondly, in Chapter 5 I attempted to build an adaptive dynamics treatment of ecological scaffolding. As outlined in the discussion, my approach is limited by not taking into account invasion dynamics at the lower level. However, preliminary computations give me hope that it will be possible to build a “canonical equation of ecological scaffolding”, that would solve this problem by altering transition probability at the higher level. Thirdly, I think that the conditions for ecological scaffolding could be made more precise, both from the point of view of the nature of the scaffold (number of niches, dispersal rate, migration mechanisms) and the nature of the internal ecology of collectives (presence of alternative ecological equilibria, time-scale separation with the higher level). A fascinating example arise when lower-level entities are not Darwinian, as in the first major evolutionary transition: the origin of life.

# Bibliography

“By pleasure drawn from discovery of new truth, the scientist is part poet, and by pleasure drawn from new ways to express old truths, the poet is part scientist.”

— EDWARD O. WILSON, *Letters to a Young Scientist* (2013)

- Abreu, C. I., Friedman, J., Woltz, V. L. A., and Gore, J. (2019). Mortality causes universal changes in microbial community composition. *Nature Communications*, 10(1):1–9.
- Aktipis, A. (2016). Principles of cooperation across systems: from human sharing to multicellularity and cancer. *Evolutionary Applications*, 9(1):17–36.
- Ameisen, J.-C. (2008). *Dans la lumière et les ombres Texte imprimé Darwin et le bouleversement du monde*. Fayard/Seuil, Paris.
- Ardré, M., Dufour, D., and Rainey, P. B. (2019). Causes and Biophysical Consequences of Cellulose Production by *Pseudomonas fluorescens* SBW25 at the Air-Liquid Interface. *Journal of Bacteriology*, 201(18):e00110–19.
- Arias-Sánchez, F. I., Vessman, B., and Mitri, S. (2019). Artificially selecting microbial communities: If we can breed dogs, why not microbiomes? *PLOS Biology*, 17(8):e3000356.
- Aristotle (1887). *Traité de la génération des animaux d’Aristote*. Hachette, Paris.
- Axelrod, R. and Hamilton, W. D. (1981). The evolution of cooperation. *Science*, 211(4489):1390–1396.
- Bantinaki, E., Kassen, R., Knight, C. G., Robinson, Z., Spiers, A. J., and Rainey, P. B. (2007). Adaptive Divergence in Experimental Populations of *Pseudomonas fluorescens*. III. Mutational Origins of Wrinkly Spreader Diversity. *Genetics*, 176(1):441–453.
- Baraban, L., Bertholle, F., Salverda, M. L. M., Bremond, N., Panizza, P., Baudry, J., Visser, J. A. G. M. d., and Bibette, J. (2011). Millifluidic droplet analyser for microbiology. *Lab on a Chip*, 11(23):4057–4062.
- Baum, D. A. and Vetsigian, K. (2017). An Experimental Framework for Generating Evolvable Chemical Systems in the Laboratory. *Origins of Life and Evolution of Biospheres*, 47(4):481–497.
- Black, A. J., Bourrat, P., and Rainey, P. B. (2019). Ecological scaffolding and the evolution of individuality: the transition from cells to multicellular life. *bioRxiv*, page 656660.
- Blokhuis, A., Lacoste, D., Nghe, P., and Peliti, L. (2018). Selection Dynamics in Transient Compartmentalization. *Physical Review Letters*, 120(15):158101.
- Boitard, L., Cottinet, D., Bremond, N., Baudry, J., and Bibette, J. (2015). Growing microbes in millifluidic droplets. *Engineering in Life Sciences*, 15(3):318–326.
- Bonner, J. T. (1982). Evolutionary strategies and developmental constraints in the cellular slime molds. *The American Naturalist*, 119(4):530–552.
- Bonner, J. T. (1998). The origins of multicellularity. *Integrative Biology: Issues, News, and Reviews*, 1(1):27–36.
- Bonner, J. T. (2015). *Life Cycles: Reflections of an Evolutionary Biologist*. Princeton University Press, Princeton.

- Bourrat, P. (2017). Evolutionary Transitions in Heritability and Individuality. *bioRxiv*, page 192443.
- Buffon, G.-L. L. (1829). *De la manière d'étudier et de traiter l'histoire naturelle*. Pillot, Paris.
- Buffon, G.-L. L. (1882). *Histoire des animaux*. Barbou frères, Limoges.
- Buffon, G.-L. L. (1884). *Oeuvres complètes de Buffon. Tome 4*.
- Buss, L. W. (1982). Somatic cell parasitism and the evolution of somatic tissue compatibility. *Proceedings of the National Academy of Sciences*, 79(17):5337–5341.
- Buss, L. W. (1987). *The evolution of individuality*. Princeton University Press, Princeton.
- Champagnat, N., Ferrière, R., and Méléard, S. (2006). Unifying evolutionary dynamics: From individual stochastic processes to macroscopic models. *Theoretical Population Biology*, 69(3):297–321.
- Cottinet, D. (2013). *Diversité phénotypique et adaptation chez Escherichia Coli étudiées en millifluidique digitale*. PhD thesis, Université Pierre et Marie Curie.
- Cottinet, D., Condamine, F., Bremond, N., Griffiths, A. D., Rainey, P. B., Visser, J. A. G. M. d., Baudry, J., and Bibette, J. (2016). Lineage Tracking for Probing Heritable Phenotypes at Single-Cell Resolution. *PLOS ONE*, 11(4):e0152395.
- Darwin, C. (1872). *The origin of species*. John Murray, London, 6th edition.
- Darwin, C. (1968). *The variation of animals and plants under domestication*. John Murray, London.
- Dawkins, R. (1976). *The selfish gene*. Oxford university Press, Oxford.
- Day, M. D., Beck, D., and Foster, J. A. (2011). Microbial Communities as Experimental Units. *BioScience*, 61(5):398–406.
- De Monte, S. and Rainey, P. B. (2014). Nascent multicellular life and the emergence of individuality. *Journal of Biosciences*, 39(2):237–248.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.
- Diderot, D. and d'Alembert, J. L. R., editors (1751). *Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers.*, volume 14. Paris.
- Dieckmann, U. and Law, R. (1996). The dynamical theory of coevolution: a derivation from stochastic ecological processes. *Journal of Mathematical Biology*, 34(5):579–612.
- Doebeli, M. (2011). *Adaptive Diversification*. Princeton University Press, Princeton.
- Doebeli, M., Ispolatov, Y., and Simon, B. (2017). Point of View: Towards a mechanistic foundation of evolutionary theory. *eLife*, 6:e23804.
- Douclier, G. (2019). Dropsignal - Millifluidic droplet trains analysis. *Zenodo*, (10.5281):1164108.
- Dupin, J.-B. (2018). *Cultures multi-parallélisées en millifluidique digitale : diversité et sélection artificielle*. PhD thesis, Université Pierre et Marie Curie.
- Duret, C. (1605). *Histoire admirable des plantes et des herbes esmerveillables & miraculeuses en nature...* Buon, Paris.
- Dutta, C. and Pan, A. (2002). Horizontal gene transfer and bacterial diversity. *Journal of Biosciences*, 27(1):27–33.
- Efron, B. (2013). Bayes' Theorem in the 21st Century. *Science*, 340(6137):1177–1178.
- Fay, J. C. and Wu, C.-I. (2000). Hitchhiking Under Positive Darwinian Selection. *Genetics*, 155(3):1405–1413.
- Feil, E. J. (2010). Linkage, Selection, and the Clonal Complex. In Robinson, D. A., Falush, D., and Feil, E. J., editors, *Bacterial Population Genetics in Infectious Disease*, pages 19–35. John Wiley & Sons.
- Ferriere, R. and Michod, R. E. (1996). The Evolution of Cooperation in Spatially Heterogeneous Populations. *The American Naturalist*, 147(5):692–717.

- Fishelson, M., Dovgolevsky, N., and Geiger, D. (2005). Maximum Likelihood Haplotyping for General Pedigrees. *Human Heredity*, 59(1):41–60.
- Fisher, R. A. (1930). *The genetical theory of natural selection*. Oxford University Press, Oxford.
- Fontenelle, B. d. (1912). Lettre sur la question de l'existence des couleurs et sur les animaux considérés comme des machines. In *Fontenelle / textes choisis et commentés par Émile Faguet*. Plon, Paris.
- Foucault, M. (1966). *Les mots et les choses*. Gallimard, Paris.
- Galton, F. (1871). I. Experiments in Pangenesis, by breeding from rabbits of a pure variety, into whose circulation blood taken from other varieties had previously been largely transfused. *Proceedings of the Royal Society of London*, 19(123-129):393–410.
- Galton, F. (1894). *Natural inheritance*. Macmillan and Company.
- Gandon, S., Jansen, V. A. A., and Baalen, M. V. (2001). Host Life History and the Evolution of Parasite Virulence. *Evolution*, 55(5):1056–1062.
- Garcia, T., Doucier, G., and Monte, S. D. (2015). The evolution of adhesiveness as a social adaptation. *eLife*, page e08595.
- Geritz, S., Gyllenberg, M., Jacobs, F., and Parvinen, K. (2002). Invasion dynamics and attractor inheritance. *Journal of Mathematical Biology*, 44(6):548–560.
- Geritz, S. A. H., Kisdi, E., Meszéna, G., and Metz, J. a. J. (1998). Evolutionarily singular strategies and the adaptive growth and branching of the evolutionary tree. *Evolutionary Ecology*, 12(1):35–57.
- Gillespie, D. T. (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of computational physics*, 22(4):403–434.
- Gillespie, D. T. (2001). Approximate accelerated stochastic simulation of chemically reacting systems. *The Journal of Chemical Physics*, 115(4):1716–1733.
- Godfrey-Smith, P. (2009). *Darwinian Populations and Natural Selection*. Oxford university Press, Oxford.
- Goodnight, C. and Stevens, L. (1997). Experimental Studies of Group Selection: What Do They Tell Us about Group Selection in Nature? *The American Naturalist*, 150(S1):S59–S79.
- Goodnight, C. J. (1990a). Experimental studies of community evolution I: The response to selection at the community level. *Evolution*, 44(6):1614–1624.
- Goodnight, C. J. (1990b). Experimental studies of community evolution II: The ecological basis of the response to community selection. *Evolution*, 44(6):1625–1636.
- Goodnight, C. J. (2000). Heritability at the ecosystem level. *Proceedings of the National Academy of Sciences*, 97(17):9365–9366.
- Grey, D., Hutson, V., and Szathmáry, E. (1995). A re-examination of the stochastic corrector model. *Proc. R. Soc. Lond. B*, 262(1363):29–35.
- Griesemer, J. (2001). The Units of Evolutionary Transition. *Selection*, 1(1-3):67–80.
- Griesemer, J. (2002). What is “epi” about epigenetics? *Annals of the New York Academy of Sciences*, 981(1):97–110.
- Grosberg, R. K. and Strathmann, R. R. (2007). The Evolution of Multicellularity: A Minor Major Transition? *Annual Review of Ecology, Evolution, and Systematics*, 38(1):621–654.
- Hamilton, W. D. (1964). The genetical evolution of social behaviour. II. *Journal of theoretical biology*, 7(1):17–52.
- Hammerschmidt, K., Rose, C. J., Kerr, B., and Rainey, P. B. (2014). Life cycles, fitness decoupling and the evolution of multicellularity. *Nature*, 515(7525):75–79.
- Hanschen, E. R., Marriage, T. N., Ferris, P. J., Hamaji, T., Toyoda, A., Fujiyama, A., Neme, R., Noguchi, H., Minakuchi, Y., Suzuki, M., Kawai-Toyooka, H., Smith, D. R., Sparks, H., Anderson, J., Bakarić, R., Luria, V., Karger, A., Kirschner, M. W., Durand, P. M., Michod, R. E., Nozaki, H., and Olson, B. J. S. C. (2016). The *Gonium pectorale* genome

- demonstrates co-option of cell cycle regulation during the evolution of multicellularity. *Nature Communications*, 7(1):1–10.
- Hardin, G. (1968). The Tragedy of the Commons. *Science*, 162(3859):1243–1248.
- Hesters, P. Y., Muir, W. M., and Craig, J. V. (1996). Group Selection for Adaptation to Multiple-Hen Cages: Humoral Immune Response,. *Poultry Science*, 75(11):1315–1320.
- Hochberg, M. E., Rankin, D. J., and Taborsky, M. (2008). The coevolution of cooperation and dispersal in social groups and its implications for the emergence of multicellularity. *BMC Evolutionary Biology*, 8(1):238.
- Hubbell, S. P. (2001). *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton University Press, Princeton.
- Ikegami, T. and Hashimoto, K. (2002). Dynamical Systems Approach to Higher-level Heritability. *Journal of Biological Physics*, 28(4):799–804.
- Jacob, F. (1970). *La logique du vivant: une histoire de l'hérédité*. Gallimard, Paris.
- Jacob, F. (1977). Evolution and tinkering. *Science*, 196(4295):1161–1166.
- Jacquard, A. (1983). Heritability: One Word, Three Concepts. *Biometrics*, 39(2):465–477.
- Johnson, C. R. and Boerlijst, M. C. (2002). Selection at the level of the community: the importance of spatial structure. *Trends in Ecology & Evolution*, 17(2):83–90.
- Johnston, I. G. and Jones, N. S. (2015). Closed-form stochastic solutions for non-equilibrium dynamics and inheritance of cellular components over many cell divisions. *Proc. R. Soc. A*, 471(2180):20150050.
- Joseph, S. B., Swanstrom, R., Kashuba, A. D. M., and Cohen, M. S. (2015). Bottlenecks in HIV-1 transmission: insights from the study of founder viruses. *Nature reviews. Microbiology*, 13(7):414–425.
- Kawecki, T. J., Lenski, R. E., Ebert, D., Hollis, B., Olivieri, I., and Whitlock, M. C. (2012). Experimental evolution. *Trends in ecology & evolution*, 27(10):547–560.
- Kerr, B., Neuhauser, C., Bohannan, B. J. M., and Dean, A. M. (2006). Local migration promotes competitive restraint in a host–pathogen ‘tragedy of the commons’. *Nature*, 442(7098):75–78.
- Kimura, M. (1983). *The neutral theory of molecular evolution*. Cambridge University Press, Cambridge.
- Kirk, D. L. (1997). *Volvox: A Search for the Molecular and Genetic Origins of Multicellularity and Cellular Differentiation (Developmental and Cell Biology Series)*. Cambridge University Press, Cambridge.
- Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.
- Kotli, S. E. and Vetsigian, K. (2018). Emergence of evolutionarily stable communities through eco-evolutionary tunnelling. *Nature Ecology & Evolution*, 2(10):1644–1653.
- Kuhn, T. S. (1970). *The structure of scientific revolutions*. University of Chicago Press.
- Lamarck, J.-B. d. M. d. -. A. d. t. (1809). *Philosophie zoologique, ou Exposition des considérations relatives à l’histoire naturelle des animaux*. Dentu, Paris.
- Lambert, A. (2008). The allelic partition for coalescent point processes. *arXiv:0804.2572 [math]*.
- Lambert, A. and Stadler, T. (2013). Birth–death models and coalescent point processes: The shape and probability of reconstructed phylogenies. *Theoretical Population Biology*, 90:113–128.
- Lavoisier, A.-L. d. -. A. d. t. (1862). *Oeuvres de Lavoisier*, volume 2. Ministère de l’instruction publique, Paris.
- Legay, J.-M. (1997). *L’expérience et le modèle*. Editions Quæ, Versailles.
- Levin, B. R. (1981). Periodic Selection, Infectious Gene Exchange and the Genetic Structure of E. Coli Populations. *Genetics*, 99(1):1–23.
- Levins, R. and Lewontin, R. C. (1985). *The dialectical biologist*. Harvard University Press, Cambridge.

- Lewontin, R. C. (1970). The Units of Selection. *Annual Review of Ecology and Systematics*, 1:1–18.
- Libby, E. and Ratcliff, W. C. (2014). Ratcheting the evolution of multicellularity. *Science*, 346(6208):426–427.
- Lind, P. A., Farr, A. D., and Rainey, P. B. (2017). Evolutionary convergence in experimental *Pseudomonas* populations. *The ISME Journal*, 11(3):589–600.
- Linné, C. v. (1735). *Systema Naturae*. Haak, Leiden.
- Luria, S. E. and Delbrück, M. (1943). Mutations of bacteria from virus sensitivity to virus resistance. *Genetics*, 28(6):491.
- MacArthur, R. H. and Wilson, E. O. (1967). *Theory of Island Biogeography*, volume 1. Princeton University Press, Princeton.
- Margulis, L. (1970). *Origin of eukaryotic cells: evidence and research implications for a theory of the origin and evolution of microbial, plant, and animal cells on the precambrian earth*. Yale University Press, New Haven, USA.
- Martin, W. and Müller, M. (1998). The hydrogen hypothesis for the first eukaryote. *Nature*, 392(6671):37–41.
- Martin, W. and Russell, M. J. (2003). On the origins of cells: a hypothesis for the evolutionary transitions from abiotic geochemistry to chemoautotrophic prokaryotes, and from prokaryotes to nucleated cells. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 358(1429):59–85.
- Maupertuis, P.-L. M. d. (1754). *Essai sur la formation des corps organisés*. Berlin.
- Maupertuis, P.-L. M. d. (1756). Lettre XIV - Sur la génération des animaux. In *Oeuvres*, volume 2. Bruyset, Lyon, nouvelle édition corrigée et augmentée edition.
- Maynard Smith, J. (1987). How to Model Evolution. In *The Latest on the Best: Essays on Evolution and Optimality*, pages 119–131. MIT Press.
- Maynard Smith, J. and Brookfield, J. F. Y. (1983). Models of Evolution [and Discussion]. *Proceedings of the Royal Society of London B: Biological Sciences*, 219(1216):315–325.
- Maynard Smith, J., Feil, E. J., and Smith, N. H. (2000). Population structure and evolutionary dynamics of pathogenic bacteria. *Bioessays*, 22(12):1115–1122.
- Maynard Smith, J. and Price, G. R. (1973). The Logic of Animal Conflict. *Nature*, 246(5427):15–18.
- Maynard Smith, J., Smith, N. H., O'Rourke, M., and Spratt, B. G. (1993). How clonal are bacteria? *Proceedings of the National Academy of Sciences*, 90(10):4384–4388.
- Maynard Smith, J. and Szathmáry, E. (1995). *The Major Transitions in Evolution*. W.H. Freeman, Oxford.
- Maynard Smith, J. and Szathmáry, E. (1993). The Origin of Chromosomes I. Selection for Linkage. *Journal of Theoretical Biology*, 164(4):437–446.
- Mayr, E. (1991). *Darwin et la pensée moderne de l'évolution*. Odile Jacob, Paris.
- McDonald, M. J., Gehrig, S. M., Meintjes, P. L., Zhang, X.-X., and Rainey, P. B. (2009). Adaptive Divergence in Experimental Populations of *Pseudomonas fluorescens*. IV. Genetic Constraints Guide Evolutionary Trajectories in a Parallel Adaptive Radiation. *Genetics*, 183(3):1041–1053.
- Mendel, G. (1866). Versuche iiber Pflanzen-Hybriden. *Verhandlungen des Naturforschenden Vereins in Brunn 4*.
- Michod, R. E. (2000). *Darwinian Dynamics: Evolutionary Transitions in Fitness and Individuality*. Princeton University Press, Princeton.
- Michod, R. E. (2007). Evolution of individuality during the transition from unicellular to multicellular life. *Proceedings of the National Academy of Sciences*, 104(suppl 1):8613–8618.
- Michod, R. E. and Roze, D. (2001). Cooperation and conflict in the evolution of multicellularity. *Heredity*, 86(1):1–7.

- Monod, J. (1970). *Le hasard et la nécessité: essai sur la philosophie naturelle de la biologie moderne*. Seuil, Paris.
- Morgan, T. H. (1915). *The Mechanism of Mendelian heredity*. New York : H. Holt and company.
- Needham, C. J., Bradford, J. R., Bulpitt, A. J., and Westhead, D. R. (2007). A Primer on Learning in Bayesian Networks for Computational Biology. *PLOS Computational Biology*, 3(8):e129.
- Nowak, M. A. (2014). *Evolutionary Dynamics: Exploring the Equations of Life*. Harvard University Press, Cambridge.
- Nowak, M. A. and May, R. M. (1994). Superinfection and the evolution of parasite virulence. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 255(1342):81–89.
- Nuel, G., Lefebvre, A., and Bouaziz, O. (2017). Computing Individual Risks Based on Family History in Genetic Disease in the Presence of Competing Risks. *Computational and Mathematical Methods in Medicine*.
- Okasha, S. (2006). *Evolution and the levels of selection*, volume 16. Clarendon Press, Oxford.
- Okasha, S. (2012). Emergence, hierarchy and top-down causation in evolutionary biology. *Interface Focus*, 2(1):49–54.
- Panke-Buisse, K., Poole, A. C., Goodrich, J. K., Ley, R. E., and Kao-Kniffin, J. (2015). Selection on soil microbiomes reveals reproducible impacts on plant function. *The ISME Journal*, 9(4):980–989.
- Paré, A. (1641). Des Monstres et Prodiges. In *Les œuvres d'Ambroise Paré*. Prost, Lyon, 10th edition.
- Penn, A. (2003). Modelling artificial ecosystem selection: A preliminary investigation. In *European Conference on Artificial Life*, pages 659–666. Springer.
- Penn, A. and Harvey, I. (2004). The Role of Non-Genetic Change in the Heritability, Variation, and Response to Selection of Artificially Selected Ecosystems. In *Artificial Life IX: Proceedings of the Ninth International Conference on the Simulation and Synthesis of Artificial Life*, volume 9, page 352. MIT Press.
- Popovic, L. (2004). Asymptotic genealogy of a critical branching process. *The Annals of Applied Probability*, 14(4):2120–2148.
- Potvin-Trottier, L., Luro, S., and Paulsson, J. (2018). Microfluidics and single-cell microscopy to study stochastic processes in bacteria. *Current opinion in microbiology*, 43:186–192.
- Queller, D. C. (1997). *Cooperators since life began*. University of Chicago Press.
- Rainey, P. B. and Kerr, B. (2010). Cheats as first propagules: A new hypothesis for the evolution of individuality during the transition from single cells to multicellularity. *BioEssays*, 32(10):872–880.
- Rainey, P. B. and Rainey, K. (2003). Evolution of cooperation and conflict in experimental bacterial populations. *Nature*, 425(6953):72–74.
- Rainey, P. B., Remigi, P., Farr, A. D., and Lind, P. A. (2017). Darwin was right: where now for experimental evolution? *Current Opinion in Genetics & Development*, 47:102–109.
- Rainey, P. B. and Travisano, M. (1998). Adaptive radiation in a heterogeneous environment. *Nature*, 394(6688):69–72.
- Ratcliff, W. C., Denison, R. F., Borrello, M., and Travisano, M. (2012). Experimental evolution of multicellularity. *Proceedings of the National Academy of Sciences*, 109(5):1595–1600.
- Ray, J. (1686). *Historia Plantarum*. Clark, London.
- Sackmann, E. K., Fulton, A. L., and Beebe, D. J. (2014). The present and future role of microfluidics in biomedical research. *Nature*, 507(7491):181–189.
- Schrödinger, E. (1944). *What is life?* Cambridge University Press, Cambridge.
- Sniegowski, P. D., Gerrish, P. J., Johnson, T., and Shaver, A. (2000). The evolution of mutation rates: separating causes from consequences. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, 22(12):1057–1066.

- Sniegowski, P. D., Gerrish, P. J., and Lenski, R. E. (1997). Evolution of high mutation rates in experimental populations of *E. coli*. *Nature*, 387(6634):703–705.
- Spiers, A. J., Kahn, S. G., Bohannon, J., Travisano, M., and Rainey, P. B. (2002). Adaptive Divergence in Experimental Populations of *Pseudomonas fluorescens*. I. Genetic and Phenotypic Bases of Wrinkly Spreader Fitness. *Genetics*, 161(1):33–46.
- Stearns, S. C. (1989). Trade-Offs in Life-History Evolution. *Functional Ecology*, 3(3):259–268.
- Stecher, B., Robbiani, R., Walker, A. W., Westendorf, A. M., Barthel, M., Kremer, M., Chaffron, S., Macpherson, A. J., Buer, J., Parkhill, J., Dougan, G., Mering, C. v., and Hardt, W.-D. (2007). *Salmonella enterica* Serovar Typhimurium Exploits Inflammation to Compete with the Intestinal Microbiota. *PLoS Biol*, 5(10):e244.
- Stevens, C. F. (2009). Darwin and Huxley revisited: the origin of allometry. *Journal of Biology*, 8(2):14.
- Strassmann, J. E. and Queller, D. C. (2011). Evolution of cooperation and control of cheating in a social microbe. *Proceedings of the National Academy of Sciences*, 108(Supplement 2):10855–10862.
- Swenson, W., Arendt, J., and Wilson, D. S. (2000a). Artificial selection of microbial ecosystems for 3-chloroaniline biodegradation. *Environmental Microbiology*, 2(5):564–571.
- Swenson, W., Wilson, D. S., and Elias, R. (2000b). Artificial ecosystem selection. *Proceedings of the National Academy of Sciences*, 97(16):9110–9114.
- Szathmáry, E. and Maynard Smith, J. (1995). The major evolutionary transitions. *Nature*, 374(6519):227–232.
- Takeuchi, N. and Hogeweg, P. (2009). Multilevel Selection in Models of Prebiotic Evolution II: A Direct Comparison of Compartmentalization and Spatial Self-Organization. *PLoS Computational Biology*, 5(10):e1000542.
- Takeuchi, N. and Hogeweg, P. (2012). Evolutionary dynamics of RNA-like replicator systems: A bioinformatic approach to the origin of life. *Physics of Life Reviews*, 9(3):219–263.
- Toprak, E., Veres, A., Yildiz, S., Pedraza, J. M., Chait, R., Paulsson, J., and Kishony, R. (2013). Building a Morbidostat: An automated continuous-culture device for studying bacterial drug resistance under dynamically sustained drug inhibition. *Nature protocols*, 8(3):555–567.
- van Vliet, S. and Doebeli, M. (2019). The role of multilevel selection in host microbiome evolution. *Proceedings of the National Academy of Sciences*, 116(41):20591–20597.
- Vaucanson, J. -. A. d. t. (1738). *Le Mécanisme du flûteur automate, présenté à Messieurs de l'Académie royale des sciences, par M. Vaucanson, auteur de cette machine, avec la description d'un canard artificiel... imitant en diverses manières un canard vivant, inventé par le même, et aussi celle d'une autre figure... jouant du tambourin et de la flûte... [Lettre de M. Vaucanson, à M. l'abbé D. F. ... sur le canard et le joueur de tambourin.]*
- Virchow, R. L. K. (1863). *Cellular pathology as based upon physiological and pathological histology ...* J. B. Lippincott, Philadelphia.
- Wade, M. J. (2016). *Adaptation in Metapopulations*. University of Chicago Press, Chicago.
- Wahl, L. M. and Gerrish, P. J. (2001). The Probability That Beneficial Mutations Are Lost in Populations with Periodic Bottlenecks. *Evolution*, 55(12):2606–2610.
- Wahl, L. M., Gerrish, P. J., and Saika-Voivod, I. (2002). Evaluating the impact of population bottlenecks in experimental evolution. *Genetics*, 162(2):961–971.
- Wahl, L. M. and Krakauer, D. C. (2000). Models of experimental evolution: the role of genetic chance and selective necessity. *Genetics*, 156(3):1437–1448.
- Wahl, L. M. and Zhu, A. D. (2015). Survival Probability of Beneficial Mutations in Bacterial Batch Culture. *Genetics*, 200(1):309–320.
- Watson, J. D. and Crick, F. H. (1953a). A structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738.
- Watson, J. D. and Crick, F. H. C. (1953b). Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, 171(4356):737–738.

- Weismann, A. (1892). *Essais sur l'hérédité et la sélection naturelle*. C. Reinwald, Paris.
- Williams, G. C. (2018). *Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought*. Princeton University Press, Princeton.
- Williams, H. T. P. and Lenton, T. M. (2007). Artificial selection of simulated microbial ecosystems. *Proceedings of the National Academy of Sciences*, 104(21):8918–8923.
- Wilson, D. S. (1975a). A theory of group selection. *Proceedings of the National Academy of Sciences*, 72(1):143–146.
- Wilson, D. S. and Knollenberg, W. G. (1987). Adaptive indirect effects: the fitness of burying beetles with and without their phoretic mites. *Evolutionary Ecology*, 1(2):139–159.
- Wilson, D. S. and Sober, E. (1989). Reviving the superorganism. *Journal of Theoretical Biology*, 136(3):337–356.
- Wilson, D. S. and Wilson, E. O. (2007). Rethinking the theoretical foundation of sociobiology. *The Quarterly review of biology*, 82(4):327–348.
- Wilson, E. O. (1975b). *Sociobiology: the new synthesis*. Belknap Press of Harvard University Press, Cambridge, Mass.
- Wilson, E. O. (1985). The Sociogenesis of Insect Colonies. *Science*, 228(4707):1489–1495.
- Wiser, M. J., Ribeck, N., and Lenski, R. E. (2013). Long-Term Dynamics of Adaptation in Asexual Populations. *Science*, 342(6164):1364–1367.
- Wright, S. (1948). On the Roles of Directed and Random Changes in Gene Frequency in the Genetics of Populations. *Evolution*, 2(4):279–294.
- Wright, S. (1949). Population Structure in Evolution. *Proceedings of the American Philosophical Society*, 93(6):471–478.
- Wright, S. (1982). The shifting balance theory and macroevolution. *Annual review of genetics*, 16(1):1–20.
- Wynne-Edwards, V. C. (1962). *Animal Dispersion in Relation to Social Behaviour*. Hafner Publishing Company.
- Xie, L. and Shou, W. (2018). Community function landscape and steady state species ratio shape the eco-evolutionary dynamics of artificial community selection. *bioRxiv*, page 264697.
- Xie, L., Yuan, A. E., and Shou, W. (2019). Simulations reveal challenges to artificial community selection and possible strategies for success. *PLOS Biology*, 17(6):e3000295.
- Zomorodi, A. R. and Segrè, D. (2016). Synthetic Ecology of Microbes: Mathematical Models and Applications. *Journal of Molecular Biology*, 428(5, Part B):837–861.

# Introduction en Français

Dans ce manuscrit, il est question des mécanismes par lesquels des collectifs d'entités deviennent sujets à l'évolution par sélection naturelle en tant que tels. Au préalable, il est nécessaire de rappeler un ensemble d'idées et de concepts à propos de la structure de ce que l'on appelle le "vivant" et de la nature des processus évolutifs.

Une observation de sens commun, qui remonte au moins à l'antiquité, est que "tout change en permanence et rien ne reste en place". Dans ce contexte, les systèmes vivants diffèrent des systèmes inertes par la *manière* dont ils changent, et par l'*organisation* qui en résulte. En deux mots: par des *mécanismes* et des *motifs*. Ainsi, la biologie de l'évolution est la science qui tente simultanément de reconstruire l'histoire du vivant, et de décrire les règles qui gouvernent sa dynamique sur des temps longs.

Commençons par définir la vie de ce point de vue. Nous considérerons comme "vivant" tout système qui participe à un type de changements particulier — une dynamique évolutive particulière. Pour éviter l'écueil du raisonnement circulaire, il faut définir clairement ce que constitue une dynamique évolutive "biologique", c'est-à-dire l'évolution par sélection naturelle. Établir une "recette pour l'évolution par sélection naturelle" — ou, plus formellement, un ensemble de conditions minimales qui doivent être vérifiée par une collection d'objets afin de participer au processus d'évolution par sélection naturelle — a été au centre de la biologie de l'évolution depuis ses origines (Godfrey-Smith, 2009). Dans le dernier paragraphe de l'*Origine des Espèces*, Darwin (1872) soutient que la diversité du monde vivant est une conséquence d'une "lutte pour l'existence", résumée par trois phénomènes clés: "croissance et reproduction", "héritage" et "variabilité". Une version moderne de ces conditions peut être trouvée dans un influent article de Lewontin (1970), un siècle plus tard. Une population d'entités est dite *Darwinienne*, et ainsi constitue une unité d'évolution (ou une unité de sélection, en fonction des auteurs), si elle présente les *propriétés Darwiniennes* suivantes:

**Variation phénotypique** Les entités dans la population sont différentes les unes des autres.

**Reproduction différentielle** Les entités sont capables de produire une descendance, et le nombre de leurs descendants dépend de leur phénotype.

**Héritabilité phénotypique** Le phénotype des descendants ressemble à celui de leurs parents.

Un point important de ce paradigme est sa nature *abstraite* (Okasha, 2006): il se réfère à une collection d'objets et de propriétés sans préciser leur nature physique, que cela soit en termes de composition chimique ou de balance énergétique. En particulier, cette définition ne dépend pas de la génétique: ces premières formulations pré-datent la découverte des lois de l'hérédité chez les diploïdes (Mendel, 1866) ainsi que de la découverte de l'ADN comme support de l'hérédité (Watson and Crick, 1953a). Cependant, ces seules simples idées ont permis de produire une riche diversité de prédictions falsifiables à propos du devenir des systèmes vivants. Avant tout, elles offrent un mécanisme causal qui explique l'apparent "projet" observé dans l'organisation des êtres vivants, et pourquoi ils semblent adaptés à leur environnement, à leur survie et à leur reproduction: leur caractère téléonomique (Monod, 1970). Par conséquent, elles apportent une importante contribution à l'étude de l'émergence de la vie, et de l'apparition de la complexité biologique dans le système solaire et au-delà. S'il est impossible de prédire la diversité des formes et des métabolismes que la vie pourrait présenter ailleurs dans l'univers, il y a de fortes chances que celle-ci devrait posséder des caractéristiques Darwiniennes pour que nous la reconnaissons comme telle.

Le monde vivant sur Terre est organisé en différents niveaux d'organisation emboîtés: gènes, chromosomes, cellules, organismes, populations, communautés... Toutes ces entités peuvent manifester les propriétés Darwiniennes à divers degrés, et donc participer au processus d'évolution par sélection naturelle. Les mammifères, par exemple, présentent des variations phénotypiques héréditaires: sans ambiguïté, ils forment une population Darwinienne. Cependant, c'est aussi le cas pour leurs cellules. Une difficulté majeure dans l'étude de l'évolution est de décider (implicitement ou explicitement) quelles sont les unités d'évolution pertinentes (c'est-à-dire les populations Darwiniennes) pour expliquer les phénomènes observés, en particulier quand des conflits existent entre niveaux d'organisation (Maynard Smith and Brookfield, 1983). L'apoptose cellulaire ne peut être prédite en considérant la cellule comme l'entité Darwinienne et nécessite de prendre en compte l'individu dans son ensemble pour expliquer son origine évolutive. À l'opposé, les lignées de cellules cancéreuses ne peuvent pas être comprises quand l'individu dans son ensemble est l'unité Darwinienne, mais sont une conséquence naturelle de la nature Darwinienne des cellules. La sélection naturelle agit simultanément sur tous les niveaux d'organisation, Démêler leurs effets respectifs est complexe. Cependant dans la majorité des situations, l'approche fondée sur le gène, qui considère que les gènes sont la cause ultime des changements (et que les autres niveaux d'organisation en sont des produits dérivés) donne des prédictions robustes (Dawkins, 1976).

Une fois que l'on a constaté que la vie était organisée en de nombreux niveaux emboîtés, la question qui se pose naturellement est celle de leur origine. Les biologistes de l'évolution appellent l'émergence de ces nouveaux niveaux des transitions évolutives majeures (Szathmáry and Maynard Smith, 1995) ou transitions évolutives de l'individualité (Buss, 1987; Michod, 2000). L'émergence des chromosomes à partir de gènes, des cellules eukaryotes à partir d'ancêtres procaryotes et d'organismes multicellulaire à partir de cellules individuelles en sont des exemples remarquables.

Le problème de l'émergence de la vie et celui des origines des propriétés

Darwiniennes est essentiellement le même. Bien sûr, une transition de l'individualité est selon toute vraisemblance plus probable quand le matériau de départ est une population Darwinienne plutôt que par abiogénèse. En effet, il y a plus de vingt-cinq exemples connus d'émergence de la multicellularité, mais une seule origine du vivant. Néanmoins, chaque transition majeure de l'évolution consiste en l'émergence d'une nouvelle population "vivante". Pour illustrer ceci, prenons l'exemple d'une entité qui présente certaines qualités d'un système vivant, tel que le métabolisme ou la croissance. Si cette entité est incapable de se reproduire, elle ne peut pas être considérée comme vivante (Jacob, 1970). Un tapis bactérien pourrait s'étendre pour recouvrir la majorité de la surface de la terre, il n'en serait pas pour autant un être vivant en tant que tel. Même si les cellules qui le composent sont bien vivantes, le collectif ne possède pas la capacité de produire de nouveaux collectifs qui lui ressemblent — un *appareil téléonomique*, au sens de Monod (1970). De manière générale, sans transition majeure, la vie n'aurait pas beaucoup changé au-delà des systèmes chimiques auto-réplicatifs primitifs.

Reconnaître que les propriétés Darwiniennes doivent émerger à chaque niveau d'organisation ne constitue pas pour autant un mécanisme qui en expliquerait l'origine. Il peut être tentant de considérer que cette émergence est le fruit d'une adaptation des collectifs. Cependant, la sélection naturelle à l'échelle des collectifs ne peut pas être invoquée comme en étant la cause première. En effet, ces mêmes propriétés Darwiniennes sont nécessaires à l'adaptation par sélection naturelle, (Black et al., 2019). Il est aussi insuffisant de considérer que les propriétés Darwiniennes des collectifs sont le simple produit d'un transfert d'un niveau d'organisation au suivant (Griesemer, 2001). En effet, les processus Darwinien collectifs font appel à des mécanismes qui sont qualitativement différent de leur équivalent aux niveaux inférieurs. Par exemple, la reproduction d'un organisme multicellulaire met en jeu un processus de développement qui ne peut pas être réduit à une simple reproduction de cellules. Pour résumer, il nous faut comprendre "comment des propriétés Darwiniennes peuvent-elles émerger à partir d'entités non-Darwiniennes, et donc par des moyens non Darwinien" (Black et al., 2019).

Pour traiter ce problème, ce manuscrit utilise extensivement les outils pour la description des propriétés Darwiniennes due à Godfrey-Smith (2009), centré autour de l'idée que les propriétés Darwiniennes peuvent être quantifiées. Dans ce cadre, une population *minimalement* Darwinienne est "une collection d'entités individuelles causalement connectées dans lesquelles on observe une variation des caractéristiques, qui mène à une différence en succès reproducteur, et qui est héritable dans une certaine mesure". (Godfrey-Smith, 2009, p. 39). Ce niveau minimal est partagé par toutes les populations Darwiniennes. Un sous-ensemble des populations minimales constitue les populations *paradigmatiques* dans lesquelles la dynamique Darwinienne est la plus claire, et donnant lieu à des structures complexes et adaptées (Godfrey-Smith, 2009, p. 41). De l'autre côté du spectre, aux limites des populations minimales se trouvent les populations *marginales*. Les populations marginales ne possèdent pas exactement les propriétés minimales, mais les approximent dans une certaine mesure. Par conséquent, elles peuvent présenter un comportement qui ressemblent à celui d'une population minimalement Darwinienne (Godfrey-Smith, 2009, p. 42). De manière générale, les populations peuvent

être placées sur un “spectre Darwinien” allant de non-Darwinien, marginal, minimal jusqu’à paradigmatique.

Cette distinction permet de mieux traiter les difficultés présentées plus tôt. S’il existe un mécanisme qui permet de promouvoir l’émergence de propriétés Darwiniennes collectives *marginales*, alors la sélection naturelle peut être envisagée comme étant le mécanisme de leur évolution future, menant potentiellement à leur raffinement vers un niveau *paradigmatique*. S’il n’est pas possible, comme nous l’avons souligné plus tôt, de considérer que les propriétés Darwiniennes soient simplement transférées d’un niveau à l’autre, un premier mécanisme qui permet à des collectifs d’obtenir des propriétés Darwiniennes marginales est la cooptation de traits d’un niveau inférieur. Par exemple, chez les algues vertes volvocines, la formation des groupes par adhésion cellulaire co-opte les mécanismes de régulation du cycle cellulaire ancestral (Hanschen et al., 2016). De même, les mécanismes ancestraux d’apoptose sont co-optés pour la fragmentation des amas cellulaires chez la levure “flocon de neige” issue de la sélection artificielle. (Ratcliff et al., 2012). Cependant, la co-option de trait n’est pas toujours possible, ou du moins pas dès les prémises d’une transition majeure, ce qui complique grandement la tâche.

L’hypothèse principale de ce manuscrit est un second mécanisme qui permet à des collectifs d’acquérir des propriétés Darwiniennes collectives *marginales*, par le résultat d’une contrainte exogène due à l’environnement. Ce mécanisme porte le nom d’échafaudage écologique (Black et al., 2019). Il ne nécessite pas de co-option de traits ancestraux à priori, mais est fondé sur la structure de la population. En effet, sous certaines conditions qui restent à définir, l’existence de ressources localisées en dèmes et de mécanismes de dispersion limités entre les dèmes peuvent être suffisants pour imposer des propriétés Darwiniennes marginales à des collectifs d’individus, sans action particulière de ces individus. Ainsi, la sélection naturelle peut agir au niveau des collectifs, promouvant le raffinement de ces propriétés Darwiniennes vers des niveaux plus paradigmatiques.

Cette thèse constitue ma contribution à la recherche de mécanismes généraux qui sous-tendent l’émergence de nouveaux niveaux d’organisation à partir de trois études de cas: la diversité neutre dans les populations emboîtées, l’émergence de la reproduction dans les organismes multicellulaires primitifs, et l’évolution de l’hérédité au niveau des communautés. Ce travail est théorique et utilise une diversité d’approches de modélisation, dont des processus stochastiques, des réseaux Bayésiens et des systèmes dynamiques. Néanmoins, j’ai tenté autant que possible d’éviter d’en faire un exercice purement formel: tous les modèles présentés ont été développés en s’inspirant de systèmes d’évolution expérimentale.

## RÉSUMÉ

---

Le vivant présente une structure emboîtée où des entités élémentaires sont intégrées dans des collectifs de plus haut niveau (gènes dans chromosomes, organelles dans cellules, cellules dans organismes, organismes dans groupes eusociaux). Tous ces niveaux sont sujet à l'évolution par sélection naturelle. En effet, les entités constitutives de chaque niveau d'organisation sont Darwiniennes, c'est-à-dire qu'elles sont distinctes, qu'elles varient de l'une à l'autre, qu'elles se reproduisent et qu'elles donnent naissance à une progéniture qui ressemble à ses parents. L'émergence d'un nouveau niveau d'organisation est un phénomène relativement rare dans l'histoire du vivant. Elle nécessite l'évolution *de novo* de propriétés spécifiques à ce niveau d'organisation qui lui permettent de participer directement à un processus d'évolution par sélection naturelle. Dans ce manuscrit, j'explore à l'aide de modèles mathématiques l'idée que ces propriétés Darwiniennes peuvent être imposée de manière exogène par l'environnement (on parle d'échafaudage écologique). Je montre comment la sélection naturelle peut partir de ses propriétés exogènes et permettre l'évolution de traits endogènes qui constituent le fondement de la reproduction et de l'hérédité au niveau des collectifs.

## MOTS CLÉS

---

Évolution, Écologie, Transitions majeures de l'évolution, Multicellularité, Origines de la Vie, Modélisation, Dynamique Adaptative, Propriétés Darwiniennes, Propriétés Collectives.

## ABSTRACT

---

Life has a nested structure where lower level entities are embedded in higher level collectives (genes in chromosomes, organelles in cells, cells in organisms, organisms in eusocial groups). All levels are subject to evolution by natural selection. This arises from the fact that at each level the focal entities are Darwinian, that is, they are discrete and vary one to another, they replicate and give rise to offspring that resemble parental types. The emergence of a new level of organisation is a relatively rare event in the history of life, and requires the *de novo* evolution of level-specific properties that allow the new level of organisation to participate directly in the process of evolution by natural selection. In this manuscript I explore, using mathematical models, the idea that Darwinian properties can be exogenously imposed (scaffolded) by the environment. I show how natural selection can build upon those scaffolded properties to promote the emergence of endogenous traits underpinning collective-level reproduction and heredity.

## KEYWORDS

---

Evolution, Ecology, Major Evolutionary Transitions, Origins of Life, Multicellularity, Adaptive Dynamics, Darwinian Properties, Collective properties.