UNIVERSIDADE ESTADUAL DE CAMPINAS
Instituto de Física Gleb Wataghin

Vítor Marquioni Monteiro

Aplicações do modelo de Derrida-Higgs finito em dinâmica de populações

Applications of the finite Derrida-Higgs model to population dynamics

CAMPINAS
2024

Vítor Marquioni Monteiro

Applications of the finite Derrida-Higgs model to population dynamics

Aplicações do modelo de Derrida-Higgs finito em dinâmica de populações

Tese apresentada ao Instituto de Física Gleb Wataghin da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Doutor em Ciências, na Área de Física.

Thesis presented to the Institute of Physics Gleb Wataghin of the University of Campinas in partial fulfillment of the requirements for the degree of Doctor of Science, in the area of Physics.

Supervisor: Prof. Dr. Marcus Aloizio Martinez de Aguiar

ESTE TRABALHO CORRESPONDE À VERSÃO FINAL DA TESE DEFENDIDA PELO ALUNO VÍTOR MARQUIONI MONTEIRO, E ORIENTADO PELO PROF. DR. MARCUS ALOIZIO MARTINEZ DE AGUIAR.

CAMPINAS
2024

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Física Gleb Wataghin
Lucimeire de Oliveira Silva da Rocha - CRB 8/9174

Monteiro, Vítor Marquioni, 1995-
M764a Applications of the finite Derrida-Higgs model to population dynamics / Vítor Marquioni Monteiro. – Campinas, SP : [s.n.], 2024.

Orientador: Marcus Aloizio Martinez de Aguiar.
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Física Gleb Wataghin.

1. Processo estocástico. 2. Evolução (Biologia). 3. Mecânica estatística. I. Aguiar, Marcus Aloizio Martinez de, 1960-. II. Universidade Estadual de Campinas. Instituto de Física Gleb Wataghin. III. Título.

Informações Complementares

Título em outro idioma: Aplicações do modelo de Derrida-Higgs finito em dinâmica de populações
Palavras-chave em inglês:
Stochastic processes
Evolution (Biology)
Statistical mechanics
Área de concentração: Física
Titulação: Doutor em Ciências
Banca examinadora:
Marcus Aloizio Martinez de Aguiar [Orientador]
Alex Antonelli
Jacopo Grilli
Maurice de Koning
Sabrina Borges Lino Araújo
Data de defesa: 23-02-2024
Programa de Pós-Graduação: Física

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0002-8904-0230
- Currículo Lattes do autor: http://lattes.cnpq.br/6972804991967652

INSTITUTO DE FÍSICA
GLEB WATAGHIN

MEMBROS DA COMISSÃO EXAMINADORA DA TESEDE DOUTORADO DO ALUNO VÍTOR MARQUIONI MONTEIRO – RA230332 APRESENTADA E APROVADA AO INSTITUTO DE FÍSICA GLEB WATAGHIN, DA UNIVERSIDADE ESTADUAL DE CAMPINAS, EM 23/02/2024.

COMISSÃO JULGADORA:

- Prof. Dr. Marcus Aloizio Martinez de Aguiar – Presidente e Orientador (IFGW/ UNICAMP)

- Profa. Dra. Sabrina Borges Lino Araújo (Universidade Federal do Paraná)

- Prof. Dr. Alex Antonelli (IFGW/ UNICAMP)

- Prof. Dr. Maurice de Koning (IFGW/ UNICAMP)

- Dr. JacopoGrilli (Abdus Salam International Centre for Theoretical Physics)

OBS.: Ata da defesa com as respectivas assinaturas dos membros encontra-se no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria do Programa da Unidade.

CAMPINAS
2024

To my parents.

Acknowledgments

Gratitude. What a simple word to express the greatness of the feeling I address here, in this single gentle page. Five years ago, when this period of my life started, I could not imagine how many people I would meet, and how many arms would support me and stand with me along this turbulent but beautiful journey. It seems like a lovely and crazy dream to be finally finishing this thesis. It was not easy, but I have no regrets. I did it with passion and, of course, with a little help. Okay, perhaps with a bit more than a little.

I acknowledge here anyone who in some way gave me some lightness during this path. To you, dear friend, my most honest thank you. It was a pleasure to have met you in this life.

But some agents in this process should be named, and, with some luck, I am not going to forget anyone.

I start with the wonderful person who received me with open arms as a student, Prof. Marcus. I have no way to thank you enough for the sheltering place you gave me in your research group. You are a role model of humbleness and kindness. It is not a coincidence that your research group is always full of admirable people. And talking about it, thank you Débora and Flávia, for helping me so many times. I have you as examples of amazing people in science (and out of science too). Thank you, Ana, Tiago, Thiago, Luis, Gabi, Isabelle, Leo, Inês, Joao, and all the other great people who were always in the weekly group meetings.

Thank you, Saron, Motoca, and Edinalra, for making my days in Campinas so much lighter.

Thank you, Krissia and Huyrá, for always rooting for me.

Thank you, Luisa, Carol, Bruno, Malu, Lívia, Iago and André, Juh, Pedro H., Pedro S., Italo, Anic, Nicolas, and all my old scout mates, for your unconditional friendship. You are the family that life gave me.

Thank you to my Triestine family, Gülce, Mahjoobeh, Solana, and Solmaz, to whom I owe so much. I have no vocabulary to express how grateful I am for having met you. Life treated me kindly when they put you all in my way. Friends for life! – and also for fun.

Thank you, Matteo, Jacopo, and Erica, for receiving me in Trieste with so much care and courtesy. You are inspiring. You make QLS a place like no other. Thank you Kyrell, Debarshi, Samuele, Tarek, Emanuele, Richmond, Rita, Mikho, Paria, Debraj, Jose, Emily, Xiaotian, Will, Matteo, Louis, Neama, and all the other QLS and ICTP and Trieste folks I had the pleasure to meet. Yes! Trieste brought me so many amazing friends, who certainly made Trieste my second home. I hope to see you all again.

Thank you, Moacyr and Clara, and all my old aminguinhos mates, for the support and friendship since the very beginning of this journey, more than ten years ago.

Lastly, but very far from least, with my deepest gratitude, thank you Mom and Dad, and all my family, for the invaluable and limitless support. I would never be what I am without you. To you, my unconditional love.

I also acknowledge the "Conselho Nacional de Desenvolvimento Científico e Tecnológico", CNPq, that partially funded this project with the grant 140728/2019-8 and the "Fundação de Amparo à Pesquisa do Estado de São Paulo", FAPESP, which funded this project with the grants 2019/13341-7 (Ph.D.) and 2021/12509-1 (BEPE). Moreover, I thank the University of Campinas, UNICAMP, (Campinas – Brazil), and the Abdus Salam International Centre for Theoretical Physics, ICTP, (Trieste – Italy), for hosting me during these years.

My apologies if I forgot some other important name, or if you believe you should have been named here. Trust me, I agree with you. I thank every single small contribution to this work. I wanted you all to be listed here, but my memory likes to prank me and these pages are quite limited. And with the fear I have already said too much, I thank you again, my dear friend.

My best wishes,
Vítor Marquioni Monteiro.

"Let the wine of friendship never run dry.

Here's to you and here's to me."
(Les Misérables, 2012 film)

"E quando estivermos à espera
que a noite volte outra vez
hei de lê contar histórias
escrever nomes na areia
pro vento brincar de apagar."
Raul Bopp, Cobra Norato, 1931.

Abstract

MARQUIONI, V.M. **Applications of the finite Derrida-Higgs model to population dynamics.** 2024. 210p. Thesis (Doctor of Science) - Institute of Physics Gleb Wataghin, University of Campinas, Campinas, 2024.

Sympatric speciation is a process where species emerge in a community in the presence of gene flow. Although already proposed by Darwin in 1859, it remains a very contentious diversification process to which strong evidence is not easy to find. Notwithstanding, in 1991, Derrida and Higgs showed that haploid individuals evolving in sympatry could form reproductively isolated groups under neutral evolutionary forces. Their model considers the evolution of a panmictic finite population, with sexual reproduction and a fixed mutation rate, in which individuals are described by binary sequences representing their genomes. In the limit of infinitely large genomes, a transition between low and high diversity regimes can be observed if mating restrictions based on genetic similarity are included. However, the same transition shows a different behavior when the genome is finite. This thesis presents a theoretical analysis of the distinct regimes of the finite Derrida-Higgs model displays, i.e., the low and high diversity regimes, including a heuristic approximation of the transition between the two. Furthermore, applications of the model and the theory are subsequently presented. The Princepe-Aguiar model for mitochondrial and nuclear genetic material coevolution is analyzed for the case of sympatric communities and our results corroborate the author's conclusions, stating that the barcode property of the mitochondrial DNA does not emerge in the absence of spatial structures. We finish this text with a model of viral evolution during epidemics in which we have studied the genetic variability in a neutral spread for different contact networks, and also the effects of quarantine regimes in such outbreaks spreading over scalefree networks. We, therefore, have introduced the first complete theory for the Derrida-Higgs dynamics which, albeit including heuristic approximations, can be extended to other studies (e.g. the mito-nuclear DNA coevolution model). Moreover, we advocate that our epidemic model provides a general framework to study the evolutionary patterns of a pathogen if the contact network structure is considered.

Resumo

MARQUIONI, V.M. **Aplicações do modelo de Derrida-Higgs finito em dinâmica de populações.** 2024. 210p. Tese (Doutor em Ciências) - Instituto de Física Gleb Wataghin, Universidade Estadual de Campinas, Campinas, 2024.

Especiação simpátrica é o processo no qual espécies emergem em uma comunidade na presença de fluxo gênico. Embora tendo sido descrito por Darwin já em 1859, esse processo de diversificação continua bastante controverso, para o qual evidências robustas não são fáceis de encontrar. Apesar disso, em 1991, Derrida e Higgs mostraram que indivíduos haploides evoluindo em simpatria sob forças neutras de evolução poderiam formar grupos reprodutivamente isolados. O modelo dos autores considera a evolução de uma população finita em panmixia, com reprodução sexuada e taxa de mutação fixa, na qual os indivíduos são descritos por meio de sequências binárias, que representam o seu genoma. No limite de genomas indefinidamente grandes, pode ser observada uma transição entre regimes de alta e baixa diversidade caso restrições à reprodução, baseadas na similaridade genética entre indivíduos, seja incluída no modelo. Contudo, a mesma transição possui comportamento diferente quando o genoma é finito. Essa tese apresenta a análise teórica dos diferentes regimes que o modelo de Derrida-Higgs finito apresenta, i.e., os regimes de alta e baixa diversidade, incluindo uma aproximação heurística para a transição entre os dois. Não obstante, apresentamos algumas aplicações tanto do modelo quanto da teoria apresentada. O modelo de Princepe-Aguiar para coevolução entre os materiais genéticos mitocondrial e nuclear é analisado no caso de comunidades em simpatria e os nossos resultados corroboram as conclusões dos autores originais do modelo, estabelecendo que a propriedade de barcode do DNA mitocondrial não emerge na ausência de estruturas espaciais. Finalizamos o presente trabalho com a introdução de um modelo de evolução viral durante uma epidemia, no qual estudamos a variabilidade genética em um espalhamento sobre redes de contato livres de escala. Assim, introduzimos a primeira teoria completa para a dinâmica de Derrida-Higgs a qual, embora contendo aproximações heurísticas, pode ser estendida para outros estudos (e.g. o modelo de coevolução mito-nuclear). Além disso, defendemos que o nosso modelo epidemiológico oferece uma ferramenta bastante geral para estudar os padrões evolutivos de um patógeno se a rede de contatos for considerada.

Contents

Part I

Introductory Remarks

Chapter 1

Introduction

1.1 Initial words and on the aims of this thesis

This work started to be developed in 2019, almost a year before the COVID-19 pandemic changed many rules of our daily lives. The project has began as a well-defined mathematical question in a model of species formation but it has suddenly branched and we found ourselves also working on epidemiology, giving our contribution to the enourmous mass of scientists that were trying to somehow help with the ongoing outbreak of the new coronavirus SARS-CoV-2. But this should not be understood as a complete detour: mathematical, computational, and probabilistic modeling and biological evolution have always been the backbone of this Ph.D. research. This is why it can all be combined in a single manuscript without lack of continuity.

By the end of a Ph.D., writing a thesis seems to be only a formality, a well-posed ending delimiter of a degree. But I have never wished it to be like this. The text in a thesis can be very technical, intricate, and very niche-specific and this is exactly how I *do not* want it to be. I cannot let the formalism go away from these pages, but I intend it to have meaningful content for the general reader. Although some background in mathematics is required, I hope it to be a textbook for anyone who would like to learn about agent-based modeling, probability theory, and biological evolution (altogether).

I want the knowledge I acquired in the course of the last five years to be well described in this text, and that is what all the effort I have put into these pages is about. Let our difference at the end of this text be our experience on the matter, not any hidden knowledge. Along these pages, some calculations can be tedious, but the algebraic manipulations can also be very instructive.

This manuscript talks about three different *agent-based models* (ABM) to which mathematical results are achieved. They are all placed in a biological evolution background in which binary chains are used as a proxy for the genetic material. We start with the Derrida-Higgs model of sympatric species formation, pass through the Princepe-Aguiar model of mito-nuclear coevolution, and end up with a model in epidemiology. My goal is to build a not so much rigid work line that can teach the reader instead of simply exposing the results of a research. In the following sections, a summary of what is ahead is presented.

1.2 A bit of Evolution

Biological evolution is one of the most successful theories in science. It constitutes a set of many evidences and fruitful explanations connecting them altogether concerning the origins and patterns of diversity of living organisms. However, what is seen nowadays as a solid and well accepted scientific theory, has once been thought of as very contentious. In the 19th century, Charles Robert Darwin and Alfred Russel Wallace independently proposed that the environment selects different *characters*, thus small variations along generations of the same species would be responsible for the emergence of different species in different environments. The long-lasting problem was a theory for variability along generations. Where is it coming from? Why do the offspring can be different from its parents? How much different? The lack of answers to these questions made the so-called *evolution theory* very controversial [1, 2].

Some years later, Gregor J. Mendel published his work on trait heritage, grounding the basis of genetics, which were believed to be incompatible with the Darwinian theory[1, 3]. Only in the 20th century, both theories were reconciled, mainly due to the works of Ronald A. Fisher (1930) [4], J. B. S. Haldane (1932) [5], and Sewall Wright (1931) [6], who showed that they are not only compatible but that the Mendelian Genetics is indeed necessary for the evolutionary theory introduced by Darwin and Wallace. The synthesis of both theories is known as *Neo-Darwianism*, *Synthetic Theory of Evolution*, or the *Modern Synthesis* [7, 1].

The synthetic theory set the framework for studying diversity as a process of variation from one generation to another and fixation of the variants according to environmental conditions. Nowadays, biomolecular processes involving the *genetic material* and its modifications are rather understood [8] and the theory of evolution is known to act in all different life scales, from single cell bacteria to complex vertebrate systems, passing through the strange archaea and the very well adapted viral world [9, 10, 11, 12].

One of the most striking conclusions of evolution is that the origin of all observed (and possibly the non-observed) diversity is a single ancestor, the *Last Universal Common Ancestor* (LUCA) [13, 14]. All life forms emerged from this unique "creature" billions of years ago [14], and a very significant evidence of this conclusion is the *universality* of the *genetic code* [15, 16]. Living beings are governed by chemical reactions. Many of these reactions are catalyzed by *proteins* called enzymes [8]. Proteins are biomolecules formed by sequences of smaller units, the *amino acids*. The amino acid sequence determines the protein spatial structure which is crucial for its function, which can also be structural, responsible for signaling and cell binding besides the enzymatic function.

Behind this protein world [17], there are the *nucleic acids*: the RNA (*ribonucleic acid*) and the DNA (*deoxyribonucleic acid*). These two molecules can be described as long chains of four letters ("A, U, C and G" for the RNA and "A, T, C and G" for the DNA) representing their smallest units, the *nucleotides*, and each triple of nucleotide is a code for a different amino acid. The genetic material, i.e., the set of nucleic acids in a given life form, acts like a "cookbook". A sequence of amino acids is formed from the "recipe" contained in the DNA (RNA) and therefore the proteins in a given living being[1]. The remarkable finding is that which triple corresponds to which amino acid is the same for every organism studied so far, and therefore this *genetic code* is *universal*. A sequence of nucleotides that codifies a given protein is called a *gene* [8].

[1]This is a very simplistic metaphor; for instance, due to a molecular mechanism called *alternative splicing* [18], the same "recipe" could result in different "dishes".

DNA and RNA are related to each other, in the language of the code letters, U and T are interchangeable, and then RNA and DNA are "the same" at this level. Some life forms have their information all stored in RNA molecules, and others in DNA molecules, but the genetic code is still the same. Thus, "genetic material" and "DNA" are going to be general expressions to designate *both molecules* in this text, and we are going to make it clear which one it is if needed.

The universal property of the genetic code is not needed if different species emerged from different life forms. All the diversity observed so far shares this same set of rules to build their proteins, which are indispensable for their existence [17]. But where does the variability come from? Small errors in this process (e.g. from the gene to the protein) can lead to different structures, which can be functional or not. If the mutant type is *more adapted* to the environment than others, then a mutant can fixate [1]. Being able to explain adaptation is a great hallmark of the evolutionary theory. Mutations occur randomly across generations, they are not "good or bad", but those that can lead to a greater survival rate of the individual tend to exist for a longer time, therefore more adapted individuals are *naturally selected*.

Different *species* can emerge in this context: the increase of variability due to mutations can end up in so different individuals that one cannot recognize them as being representative of a single "group" [2]. I must admit that this text has taken a great jump now, grouping individuals is not an easy task and the concept of species is also contentious [19], but this is maybe the high point of the evolution theory: individuals that we classify as different "groups" (in some sense[2]) are the result of evolution acting on a previous single group [1]. To give a simple example, thousands of years ago, a lineage of wolves started to become different from the others, living closer to humans, being domesticated and, after years of accumulating changes, they are now recognized as dogs [20].

How exactly this intense diversification happens and what defines different species is going to be more discussed later in this text, but we have reached our first milestone: much of this work is about a biological process that gives rise to *different species*, which is known as *speciation*.

1.3 Agent-based modeling

Much of science is about raising hypotheses that can explain observed patterns. Good hypotheses not only explain but are also able to predict non-observed patterns, and are thereafter tested. The way someone joins different hypotheses and/or develops a sequence of reasonable arguments to show a specific pattern is what we call a *model*. Different models for the same observation can be proposed; they can be very complex, contain many steps, a long list of hypotheses, and a very intricate set of rules (in many cases these are mathematical rules – equations). But models can also be very simple, aimed to describe some general aspect of an observed phenomenon, or they even can be grounded on a unique rule that "magically" works with astonishing precision.

In *population dynamics* [21], models are intended to describe the variations in the number of individuals and their ages over time. For instance, suppose that in a certain farm, there are rabbits and foxes. The foxes predate the rabbits. If there are no rabbits, the fox population decreases, because they die from starvation, and then the rabbits can reproduce without being predated. Once the number of rabbits increases, there are going

[2]To classify individuals into different groups is the task of a field called *Taxonomy*.

to be many prey for the foxes and then their population can also increase, hence reducing the number of rabbits, restarting, this way, the cycle. This ecological system, known as a predator-prey model [22], can be written as a set of ordinary differential equations, which were first introduced by Alfred J. Lotka and Vito Volterra in 1920 and 1926 [23, 24]. These equations do not name the individuals in the community, they only describe what is observed at the *population level*.

Population level descriptions like this follow a long tradition from the physics of thermodynamics. The amount of molecules that are present in a gas is so huge that it is not only unfeasible to solve all the equations of motion, but it is also useless. State equations, describing the relations among macroscopic quantities (like temperature, pressure and volume) are much more significant. The field of statistical mechanics, developed in the mid 19th century, has set the mathematical background of the thermodynamic results, and later, with stochastic analysis, has created the basis for the *reaction networks theory* [25, 26]. In a given system, a set of compounds reacts forming some *products*; the concentration of the *reactants* changes as the products are being formed. Many systems can be described in this way; in the predator-prey dynamics, for instance, a predator "reacts" with a prey and the product is more predators. Ecological systems, epidemiology, evolutionary systems, chemical reactions, and gene expression schemes, are all well suited to be described as reaction networks, and this makes population level descriptions a very powerful and general way of modeling [27].

But a direct consequence of this picture is that there is no individual tracking of organisms. As said before, it can be useless, but there is no reason for it to not be required in some cases. In evolutionary systems, for instance, mutations appear at the level of the individual, and can or not be fixated in a way to affect the population in an *average* way. This is the second problem of this type of modeling: they assume *well-mixed* populations, thus the behavior of a given reactant is the same as the average of the population of the same reactant. For this reason, these models are also called *mean-field* models and are not representative of every system in its simplest versions [28]. Also, mean-field models can become highly non-linear, hindering their analysis the more realistic they attempt to be. As a way out, one can consider models at the *individual level* in order to introduce heterogeneities.

Individual-based models (IBM) describe the behaviors of each smallest (and autonomous) constituents of a system [29]. For instance, it would describe each one of the rabbits and each one of the foxes; one would know if *Bugs Bunny* has survived the foxes attacks, grew old, and had kids, or if it has lost for *Vulpix* and his gang. Since they can be very general, these models are also called *agent-based models* (ABM), because individuals can be genes, animals, market agents, and so on. The interactions between agents are defined as rules, which can be very simple, like "when a blue agent encounters a red one, they both become blue", but (in this case) the important aspect is that the color of every agent is well defined, not only the proportions or the total amount of blues and reds.

Agent-based modeling is our second milestone. The diversification processes we are going to introduce occur at the level of the individuals, not of the population. This way, heterogeneities are easily introduced, although the mathematical tractability gets harder. As mean-field models have very standard techniques, grounded on dynamical systems mathematics, when dealing with ABMs, each model can be a completely new system, making it a challenge to find analytical results. ABMs are in general introduced as computational models, to which specific algorithms are written in order to look for strong results. On the other hand, the specificity of the codes and the (often) long list

of individual rules make them hard to be scientifically described in papers, utilized in different contexts, and even reproduced in different works.

The difficulties concerning ABMs led a group of scientists to propose a standard protocol for writing about such models, the *ODD Protocol*, which states for Overview, Design concepts and Details [30]. Despite its "best-practice" purpose, we follow the original publications on the models we deal with and do not believe the set of rules within them is big enough to hinder its usage and reproducibility.

1.4 What follows

This thesis is divided into three parts. In this first one, I have introduced the upcoming content and within the following chapters, three different theoretical tools are going to set the mathematical framework on which we are going to work. Chapter 2 introduces the probability toolkit we are going to need to develop our goals. We will start with the axiomatic probability theory, showing the properties of a probability measure and reach the concept of conditional probability, which is definitely the main tool one can find in this text. This chapter finishes with an introduction to Markovian processes and some of their properties.

Chapter 3 discusses biological evolution in more detail. We are going to talk about evolutionary forces, of some models of evolution, and introduce the concept of species and speciation processes. Our main goal is to place the reader in this world of different concepts; this is not a biology text, so many theoretical questions might be not approached, but we hope it to be enough for the full comprehension of the present work.

Chapter 4 introduces important concepts in network theory, which is also a language that is going to be spread all around this text. It could be easily placed as a section somewhere else in this introduction, but as a matter of organization, we make it a separate chapter, finishing this way our theoretical background.

The second and third part of this thesis regards different agent-based evolutionary models. In Part 2, a model of sympatric species formation is presented: the Derrida-Higgs model. Its introduction and analytical theory are developed in Chapter 5. This model presents an interesting transition between low and high diversity regimes, and a heuristic theory for this transition is presented in Chapter 6. A model for the high diversity regime appears in Chapter 7.

The Derrida-Higgs model is easily adapted to different evolutionary scenarios. In 2020, a follow-up version to study the coevolution between the *Mitochondrial* and Nuclear DNA was proposed by Princepe and Aguiar. We present and study this model in Chapter 8, thus finishing Part 2.

Part 3 is devoted to a third evolutionary agent-based model, which uses the same guidelines as in the previous models, but in the context of epidemiology. The model is introduced in Chapter 10 and two different results are presented and discussed in Chapters 11 and 12. These two chapters integrally display the Results, Discussion, and Conclusion parts of two already published works. This part is concluded in Chapter 13.

Brief final words are left for the end of the text. Some appendixes can also be found at the end of the text. Appendix A supplements the Derrida-Higgs theory, while appendixes B, C, and D, supplement the epidemics model.

Chapter 2

A probabilistic background

This chapter closely follows the notations and definitions from the book on probability and random variables of M. N. Magalhães [31].

The idea behind probability theory dates from the Renaissance, when, in the 16th century, Girolamo Cardano first studied the mathematics of gambling [32]. His work "Book on Games of Chance" was only published a hundred years later, posthumously [33]. As the previous sentence may sugest, probability regards chances, it measures how likely events can be. Such chances may be related to some trust in a given outcome; different gamblers could have different trusts in how likely an ace is to show up in a certain poker round. Nowadays, it might seem obvious that the probability of drawing any card from a deck should not depend on the players, but this has not always been a straightforward argument.

A lot has been developed since the works of Cardano [34], and despite a quite subjective way of defining what *probability* is, an axiomatic theory appeared in 1933, due to Andrey Kolmogorov [35], allowing these quantities to be calculated unambiguously [36, 37]. And this is where we start the following section.

2.1 The Kolmogorov axioms

Suppose a game is played by throwing a dice and checking its upper face. The set of all possible results of this game is $\Omega = \{1, 2, 3, 4, 5, 6\}$ for a common cubic dice. Ω is called the *sample space*. But sometimes we are not exactly interested in single elements of the sample space; one may ask if the result is even or odd, or higher than 1, or a multiple of 3. "Combinations" of possible "types" of outcomes, i.e., any subset of the sample space, may be of interest to the one who is playing the game.

This observation leads us to introduce a second concept: the *σ-algebra* (also called *σ-field*). Let $\mathcal{F}$ be a class of subsets of Ω. $\mathcal{F}$ is a σ-algebra of Ω if it satisfies

1. $\Omega \in \mathcal{F}$;
2. Let $A \subset \Omega$. It $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$, where A^c denotes the complement of A;
3. If $A^i \in \mathcal{F}$, $i \geq 1$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

In simple words, a σ-algebra is a set of subsets of the sample space that gives meaning to "groups" of results. First, the whole sample set should be one of these groups. Second, if a given group is considered, this group can also be *not* considered. Finally, if many groups are considered, unions of these groups should also be considered. For instance, think about the game of throwing a dice. Suppose we are interested in knowing whether it is even ($A^{\text{even}} = \{2, 4, 6\}$) or odd ($A^{\text{odd}} = \{1, 3, 5\}$). Notice that $(A^{\text{even}})^c = A^{\text{odd}}$. Also, any face can be shown up after throwing the dice, thus $A^{\text{all}} = \{1, 2, 3, 4, 5, 6\}$ should also be a set of interest. Moreover, $A^{\text{even}} \cup A^{\text{odd}} = A^{\text{all}}$ (which is a set of interest). Take the sample space $\Omega = A^{\text{all}}$. Hence, the set $\mathcal{F} = \{A^{\text{even}}, A^{\text{odd}}, A^{\text{all}}, \varnothing\}$ is a σ-algebra of Ω (where the empty set $\varnothing = \Omega^c$).

Now, we are able to ask *how likely* one particular type of outcome to appear is. In the previous example, one can ask about the chances of the dice to showing up as an even or an odd number. Observe that this question directly regards the elements of the σ-algebra, not exactly of the sample space Ω. Hence, we define the *probability* $\mathcal{P}(A)$ of an *event* $A \in \mathcal{F}$ as a function $\mathcal{P} : \mathcal{F} \to [0, 1]$ satisfying

1. $\mathcal{P}(\Omega) = 1$;
2. $\mathcal{P}(A) \geq 0$ for every $A \in \mathcal{F}$;
3. For every sequence $A_1, A_2, \ldots \in \mathcal{F}$, with $A_i \cap A_j = 0$ for $i \neq j$,

$$\mathcal{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathcal{P}(A_i). \tag{2.1.1}$$

These axioms are known as the Kolmogorov's axioms.

A *probability space* is the triple $(\Omega, \mathcal{F}, \mathcal{P})$. Problems involving the calculation of probability can be subjective because all the elements of the constructed probability space should be well defined, but once they are given, the calculations are straightforward, without ambiguity.

2.1.1 Properties of a probability

From the Kolmogorov's axioms, the following properties can be proved [31]. Given the probability space $(\Omega, \mathcal{F}, \mathcal{P})$, we have

1.
$$\mathcal{P}(A^c) = 1 - \mathcal{P}(A). \tag{2.1.2}$$

2. Let A and B be two events. Then,

$$\mathcal{P}(B) = \mathcal{P}(B \cap A) + \mathcal{P}(B \cap A^c); \tag{2.1.3}$$

3. If $A \subset B$, then
$$\mathcal{P}(A) \leq \mathcal{P}(B); \tag{2.1.4}$$

4. For the event $A \cup B$,

$$\mathcal{P}(A \cup B) = \mathcal{P}(A) + \mathcal{P}(B) - \mathcal{P}(A \cap B); \tag{2.1.5}$$

5. For any events $A_1, A_2, \ldots$

$$\mathcal{P}\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \mathcal{P}\left(A_i\right). \tag{2.1.6}$$

Proofs [31]

1. This property follows from the fact that $A \cap A^c = \varnothing$. Then, from the third axiom,

$$\mathcal{P}(\Omega) = \mathcal{P}(A \cup A^c) = \mathcal{P}(A) + \mathcal{P}(A^c) = 1 \Rightarrow \mathcal{P}(A^c) = 1 - \mathcal{P}(A).$$

2. By noticing that $B = (B \cap A) \cup (B \cap A^c)$ and that $(B \cap A) \cap (B \cap A^c) = \varnothing$, then

$$\mathcal{P}(B) = \mathcal{P}((B \cap A) \cup (B \cap A^c)) = \mathcal{P}(B \cap A) + \mathcal{P}(B \cap A^c).$$

3. With the same reasoning, we can write

$$B = (B \cap A) \cup (B \cap A^c) = A \cup (B \cap A^c)$$

in which the second equality comes from $A \subset B$. Now, from the third axiom,

$$\mathcal{P}(B) = \mathcal{P}(A) + \mathcal{P}(B \cap A^c) \geq \mathcal{P}(A).$$

4. Write $A \cup B$ as

$$A \cup B = (A \cap B^c) \cup (A \cap B) \cup (A^c \cap B),$$

and notice that these sets are mutually disjointed. Thus,

$$\mathcal{P}(A \cup B) = \mathcal{P}(A \cap B^c) + \mathcal{P}(A \cap B) + \mathcal{P}(A^c \cap B).$$

From the second property

$$\begin{aligned} \mathcal{P}(A) &= \mathcal{P}(A \cap B) + \mathcal{P}(A \cap B^c), \\ \mathcal{P}(B) &= \mathcal{P}(B \cap A) + \mathcal{P}(B \cap A^c), \end{aligned}$$

we find that

$$\mathcal{P}(A \cup B) = \mathcal{P}(A) + \mathcal{P}(B) - \mathcal{P}(A \cap B).$$

5. By writing $\bigcup_{i=1}^{\infty} A_i$ as the union of a sequence of disjoint sets

$$\bigcup_{i=1}^{\infty} A_i = A_1 \cup (A_1^c \cap A_2) \cup (A_1^c \cap A_2^c \cap A_3) \cup \ldots$$

and then from axiom 3,

$$\mathcal{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \mathcal{P}(A_1) + \mathcal{P}(A_1^c \cap A_2) + \mathcal{P}(A_1^c \cap A_2^c \cap A_3) + \dots.$$

For every j,

$$A_1^c \cap \ldots \cap A_{j-1}^c \cap A_j \subset A_j,$$

and then, from property 3,

$$\mathcal{P}\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \mathcal{P}(A_1) + \mathcal{P}(A_2) + \dots.$$

2.2 Conditional probability and independence

Many questions on probabilities emerge when previous information is considered. For instance, after throwing a cubic dice, what is the probability of getting a 3 *given that* it has shown up an odd number? *Conditioning* an event to another event is the concept we introduce now.

Let us suppose that we know that the result of a given experiment is a point $\omega \in B$, and B is an event of a σ-algebra $\mathcal{F}$ of a probability space $(\Omega, \mathcal{F}, \mathcal{P})$. Then, if one is interested in the occurrence of an event A, then ω must also belong to A. Hence, $\omega \in A \cap B$. Let $\overline{\mathcal{P}}(A)$ denote the probability of A when it is restricted to the occurrence of B, thus

$$\overline{\mathcal{P}}(A) \sim \mathcal{P}(A \cap B).$$

To satisfy the Kolmogorov axioms, $\overline{\mathcal{P}}(A)$ is a probability if the proportionality constant is chosen to be $1/\mathcal{P}(B)$ [1]. This new probability is known as a *conditional probability*, and it is written as $\mathcal{P}(A|B)$,

$$\mathcal{P}(A|B) = \frac{\mathcal{P}(A \cap B)}{\mathcal{P}(B)}, \tag{2.2.1}$$

whenever $\mathcal{P}(B) > 0$, and it is read as "the probability of A *conditioned to* (or *given*) B". Notice that $\mathcal{P}(A|\Omega) = \mathcal{P}(A)$.

2.2.1 Multiplication rule

Events that mutually happen are generally described as an "and": A *and* B happen. This is the same as saying that the event $A \cap B$ happens. This event has probability

$$\mathcal{P}(A \cap B) = \mathcal{P}(A|B)\mathcal{P}(B) = \mathcal{P}(B|A)\mathcal{P}(A). \tag{2.2.2}$$

[1] This new probability is defined over a σ-algebra $\mathcal{F}_B$ defined as the set of all events $B \cap C$ for any $C \in \mathcal{F}$. $\mathcal{F}_B$ is called a *restriction* of $\mathcal{F}$ to the *sample space* B.

Suppose now a sequence of events $A_1, A_2, \ldots, A_n$, with no empty intersection. What is the probability of all events to happen, i.e., what is the probability $\mathcal{P}\left(\bigcap_{i=1}^{n} A_i\right) > 0$?

From (2.2.2),

$$\mathcal{P}(A_1 \cap A_2) = \mathcal{P}(A_1)\mathcal{P}(A_2|A_1). \tag{2.2.3}$$

Now, rename $A_1 \to A_1 \cap A_2$ and $A_2 \to A_3$. Then,

$$\mathcal{P}(A_1 \cap A_2 \cap A_3) = \mathcal{P}(A_1 \cap A_2)\mathcal{P}(A_3|A_1 \cap A_2). \tag{2.2.4}$$

But from (2.2.3),

$$\mathcal{P}(A_1 \cap A_2 \cap A_3) = \mathcal{P}(A_1)\mathcal{P}(A_2|A_1)\mathcal{P}(A_3|A_1 \cap A_2). \tag{2.2.5}$$

Then, by induction,

$$\mathcal{P}\left(\bigcap_{i=1}^{n} A_i\right) = \mathcal{P}(A_1)\mathcal{P}(A_2|A_1)\mathcal{P}(A_3|A_1 \cap A_2)\ldots\mathcal{P}(A_n|A_1 \cap \ldots \cap A_{n-1}). \tag{2.2.6}$$

2.2.2 Total probability law

This is a very important result, which is going to be used a lot in this text. Suppose someone wishes to know the probability of a given event A, e.g., the probability of a dice showing up a 6. Given the probabilities of a 6, when given that the showed face is a multiple of 2 and when it is *not* a multiple of 2, is it possible to know the probability of showing the 6? Notice that "being a multiple of 2" (thus 2, 4 and 6) and "not being a multiple of 2" (thus 1, 3, and 5) are complementary events, say it $M_2 = \{2, 4, 6\}$ and $M_2^c = \{1, 3, 5\}$. $M_2 \cup M_2^c = \Omega \equiv \{1, 2, 3, 4, 5, 6\}$.

Let M_6 be the event of the dice showing a 6. $M_6 = (M_6 \cap M_2) \cup (M_6 \cap M_2^c)$. Hence,

$$\begin{aligned}\mathcal{P}(M_6) =& \mathcal{P}(M_6 \cap M_2) + \mathcal{P}(M_6 \cap M_2^c) \\ =& \mathcal{P}(M_6|M_2)\mathcal{P}(M_2) + \mathcal{P}(M_6|M_2^c)\mathcal{P}(M_2^c).\end{aligned} \tag{2.2.7}$$

Suppose $M_2 = M_2^c = 1/2$ and that the 6 appears half of the time when the result is even, thus $\mathcal{P}(M_6|M_2) = 1/2$, (obviously $\mathcal{P}(M_6|M_2^c) = 0$). Then, $\mathcal{P}(M_6) = 1/4$ and anyone can believe that there is something weird with the dice[2]. Albeit the information about the number 6 was given in terms of being an even or odd face, it was still possible to know its probability, and the reason was because being even or odd are complementary sets. This result (expressed in equation (2.2.7)) can be extended to what is known as the *total probability law*.

A partition $P = \{C_1, \ldots, C_m\}$ of F is defined as a set of non-empty subsets ($C_i \neq \varnothing$) of F, such that they are all pairwise disjoint, ($C_i \cap C_j = \varnothing$) for $i \neq j$ and their union equals F, ($\bigcup_{i=1}^{m} C_i = F$). Given a probability space $(\Omega, \mathcal{F}, \mathcal{P})$ and a partition $P = \{C_1, \ldots, C_m\}$

[2]We have assumed that the reader is familiar with the idea of *equiprobability* of the state space, i.e., every face of the dice has the same chance to appear, 1/6.

of Ω,

$$\mathcal{P}(A) = \sum_{i=1}^{m} \mathcal{P}(A|C_i)\mathcal{P}(C_i). \tag{2.2.8}$$

Proof:

$$\sum_{i=1}^{m} \mathcal{P}(A|C_i)\mathcal{P}(C_i) = \sum_{i=1}^{m} \mathcal{P}(A \cap C_i) = \mathcal{P}\left(\bigcup_{i=1}^{m} A \cap C_i\right) = \mathcal{P}\left(A \cap \bigcup_{i=1}^{m} C_i\right) = \mathcal{P}(A) \tag{2.2.9}$$

in which in the first equality we used the definition of a conditional probability; in the second, the fact that C_i are pairwise disjoint; and in the last, that they cover Ω.

2.2.3 Bayes' theorem

The Bayes' appeared in 1764 [38] and it is used to calculate probabilities of an event based on previous knowledge [39]. It is given by the formula

$$\mathcal{P}(A|B) = \frac{\mathcal{P}(B|A)\mathcal{P}(A)}{\mathcal{P}(B)}. \tag{2.2.10}$$

which updates *prior* probabilities $\mathcal{P}(A)$ given that experiments resulted in B.

The Bayes' theorem is the basis of *Bayesian statistics* (which opposes the so-called frequentist approach [39]) and can be easily shown from equation (2.2.2). Also, if $P = \{C_1, \ldots, C_m\}$ is a partition of Ω, the Bayes' theorem can be written as

$$\mathcal{P}(C_i|B) = \frac{\mathcal{P}(B|C_i)\mathcal{P}(C_i)}{\sum_{j=1}^{m} \mathcal{P}(B|C_j)\mathcal{P}(C_j)}. \tag{2.2.11}$$

2.2.4 Independent events

Asking the probability of an event A given that another event B has happened raises the question what if B does not interfere at all on the occurrence of A? This brings the concept of *independence between events*. Two events A and B in $(\Omega, \mathcal{F}, \mathcal{P})$ are *independent* if and only if

$$\mathcal{P}(A \cap B) = \mathcal{P}(A)\mathcal{P}(B). \tag{2.2.12}$$

Equivalently,

$$\mathcal{P}(A|B) = \mathcal{P}(A), \tag{2.2.13}$$

i.e., the occurrence of B does not change the probability of A.

In some cases, two events are independent only if they are conditioned to the occurrence of a third event. This is called *conditional independence*, and it is going to be used in key points of this thesis. It reads as A and B are *conditionally independent* given C if and only if

$$\mathcal{P}(A \cap B|C) = \mathcal{P}(A|C)\mathcal{P}(B|C). \tag{2.2.14}$$

For completeness, suppose now a sequence of events $A_1, \ldots, A_n$. In order to define independence among these many events, pairwise independence is not enough, i.e., $\mathcal{P}(A_i \cap A_j) = \mathcal{P}(A_i)\mathcal{P}(A_j)$ for every $i \neq j$ is not sufficient (albeit necessary). Thus, we define the events A_i to be independent if and only if for every colection of indexes $1 \leq i_1 < i_2 < \ldots < i_k \leq n$, with $2 \leq k \leq n$,

$$\mathcal{P}(A_{i_1} \cap \ldots \cap A_{i_k}) = \mathcal{P}(A_{i_1}) \ldots \mathcal{P}(A_{i_k}) \tag{2.2.15}$$

is satisfied.

2.3 Random variables

The theory introduced so far is very rich and still leads to many other interesting results, but now we are switching to a new definition, which is very rich itself and would deserve its own chapter in more specific texts. It is very useful to describe events as numbers. As an example, if an experiment is to throw a coin, one could assign the number 0 to heads and 1 to tails. This simple idea gets a deeper meaning and formal definition as *random variables*.

Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space. The random variable $X : \Omega \to \mathbb{R}$ is a real-valued function defined on the sample space Ω such that

$$X^{-1}(I) = \{\omega \in \Omega : X(\omega) \in I\} \in \mathcal{F}, \tag{2.3.1}$$

for every $I \subset \mathbb{R}$. Thus, X is a random variable if for any set $I \subset \mathbb{R}$, its *inverse image* belongs to the σ-algebra $\mathcal{F}$. Since it is guaranteed that $X^{-1}(I)$ is an event (i.e., belongs to the σ-algebra), it is possible to assign probabilities to random variables. The usual notation is capital letters for the function and small letters to its value, $X(\omega) = x$.

Despite its technical definition, using random variables makes the calculus of probabilities something very natural. The next concept regards the probabilities when dealing with random variables.

2.3.1 Probability distribution

Consider a random variable X. If X assumes only a countable number of values, then it is defined as a *discrete random variable*. The *probability distribution function* (PDF) of a discrete random variable is a function that assigns probabilities to the values of X. If X can have the values $x_1, x_2, \ldots$, then the probability distribution function $p : \mathbb{R} \to [0, 1]$ is

$$p(x_i) = \mathcal{P}(X = x_i) \tag{2.3.2}$$

where $X = x_i$ is the short notation for the set $\{\omega \in \Omega : X(\omega) = x_i\}$. Albeit the notation may look "easy to get", always keep in mind that a probability is assigned to an event (i.e., a set), while the PDF is assigned to a real number.

Two important properties of p are [31]

1.

$$0 \leq p(x_i) \leq 1, \text{ for every } i; \tag{2.3.3}$$

2.
$$\sum_i p(x_i) = 1. \tag{2.3.4}$$

The first property regards the non-negative values of a probability and the second property is the *normalization* property, i.e., it reflects that $\mathcal{P}(\Omega) = 1$.

The PDF emerges quite naturally in this theory, but it is possible to define an even more general quantity: the *cumulative distribution function* (CDF). The cumulative distribution function of a random variable X in $(\Omega, \mathcal{F}, \mathcal{P})$ is defined by

$$F_X(x) = \mathcal{P}(X \in (-\infty, x]), \tag{2.3.5}$$

with $X \in \mathbb{R}$. The event $X \in (-\infty, x]$ is commonly written as $X \leq x$. It worths mentioning that different authors give different names to the above defined functions,[3] but we follow here the usual nomenclature used by physicists.

The CDF [31] satisfies the following properties (which we are not going to prove)

1.
$$\lim_{x\to-\infty} F(x) = 0 \text{ and } \lim_{x\to+\infty} F(x) = 1; \tag{2.3.6}$$

2. F is a right-continuous function;

3. F is non-decreasing, i.e., $F(x) \leq F(y)$ whenever $x \leq y$ for any real numbers x and y.

Moreover, for a discrete random variable, the CDF is given by

$$F(x) = \sum_{x_i \leq x} p(x_i), \tag{2.3.7}$$

and

$$p(x_i) = F(x_i) - F(x_i^-), \tag{2.3.8}$$

where $F(x_i^-) \equiv \lim_{x\to x_i^-} F(x)$, ($\lim_{x\to x_i^-}$ denotes the left-sided limit as x approaches x_i).

Continuous random variables

Although much of this text is going to consider discrete random variables, another important case is when the CDF, instead of being written as a sum (Eq.(2.3.7)), is given by

$$F_X(x) = \int_{-\infty}^{x} f(\omega)d\omega, \text{ for any } x \in \mathbb{R}, \tag{2.3.9}$$

for a random variable X in $(\Omega, \mathcal{F}, \mathcal{P})$, where $f : \mathbb{R} \to \mathbb{R}$ is a non-negative function. f is called the *probability density function*. This function satisfies

1.
$$f(x) \geq 0, \text{ for any } x \in \mathbb{R}; \tag{2.3.10}$$

[3] In many cases, the CDF is simply called the *probability function*, which can cause a lot of confusion regarding the names of these quantities.

2.
$$\int_{-\infty}^{+\infty} f(\omega)d\omega = 1; \tag{2.3.11}$$

which is the normalization property.

Analogously to the equation (2.3.8), we can also show that, for continuous random variables,

$$\mathcal{P}(a < X \leq b) = \int_a^b f(x)dx = F(b) - F(a), \tag{2.3.12}$$

where we have used our simplified notation for the event $X \in (a, b]$. It is very interesting to notice that for continuous random variables, the probability of a point is zero, given the definition of the CDF as the integral of a density [31]. In physics texts, it is common to write that the probability of a random variable to be within the interval $(x, x + dx)$ is equal to $f(x)dx$ [40].

2.3.2 Expected value and moments of a random variable

When performing any experiment, e.g., to calculate the light velocity, the air dielectric constant, or the Young modulus of some solid, there is no way of knowing what is the *exact* result of the experiment. One may repeat the experiment many times and find very numerically close results, but after some decimal places, the numbers diverge in a non-predictable way. This is because there are many errors involved in the experiment and there can be no way to control some of them. Thus, the result of an experiment might be described as a random variable, but it does not mean that its non-exact predictability forbids learning from experiments.

For instance, distribution probability functions can be peaked around some number, and this fact provides important information concerning what someone may *expect* from an experiment. And that is where we find the concept of *mathematical expectation* or *mean* of a random variable.

Let X be a random variable in $(\Omega, \mathcal{F}, \Omega)$. If X is discrete and takes the values x_i for $i \in I$, then the expectation of X (or the mean value of the distribution of X, or its expected value) is defined as

$$\mathbb{E}(X) = \mu_X = \sum_{i \in I} x_i p_X(x_i), \tag{2.3.13}$$

where p_X is the PDF of X (if the sum is determined). If X is continuous, then

$$\mathbb{E}(X) = \mu_X = \int_{-\infty}^{+\infty} x f(x)dx, \tag{2.3.14}$$

where f_X is the density probability function of X (if the integral exists).

The expected value of X is connected to the *statistical average* of the results of an experiment. On the other hand, it can also happen that this value is not so meaningful. Suppose an experiment is to throw an honest cubic dice. The showing face is the random variable we are interested in, which can result in $x_i = i$ for $i = 1, \ldots, 6$. Since the dice is cubic and honest, $p_X(x_i) = 1/6$. From Eq.(2.3.13), $\mathbb{E}(X) = 3.5$, and this value has almost no meaning to the player.

It is possible to define functions of random variables. These functions are also random variables with different associated probability distributions. Then, concerning this new PDF, one can also calculate the expected value of a function. We define the *k−th order moment* of a random variable X, as

$$\mathbb{E}(X^k) = \sum_{i \in I} (x_i)^k p_X(x_i), \tag{2.3.15}$$

for X discrete, or

$$\mathbb{E}(X^k) = \int_{-\infty}^{+\infty} x^k f(x) dx, \tag{2.3.16}$$

for X continuous (if these operations are determined).

These quantities can provide other information about the distribution, e.g. if it is symmetric around some point or if it decays fast enough [41]. It is useful to redefine the moments by discounting the mean value from the random variable: $X - \mu_X$. Hence, the *k-th order central moment* is defined as

$$\mathbb{E}((X - \mu_X)^k) = \sum_{i \in I} (x_i - \mu)^k p_X(x_i), \tag{2.3.17}$$

for the discrete case and analogous for the continuous case.

Variance of X

For instance, the second central moment of a random variable X is called its *variance*

$$\mathrm{Var}(X) = \sigma^2 = \mathbb{E}((X - \mu_X)^2). \tag{2.3.18}$$

It is easy to show that $\sigma^2 = \mathbb{E}(X^2) - \mathbb{E}(X)^2$. The variance of a distribution gives information about how much someone can expect that a value of an experiment deviates from the expected value. For the honest cubic dice mentioned before, $\sigma^2 \approx 2.9$. The square root of the variance is called the *standard deviation*.

The next important step of this introduction would be to define *random vectors* [31], which are indeed used in the following parts of this thesis, but in order to keep it gentle, we are going to skip it. As a hint, a random vector can be taught as a multidimensional real function, in which each coordinate is given by a different random variable. Probability distribution functions in this case are called *joint distributions* and they represent the probability of the intersection of the events that happen in each vector coordinate. All the theory is developed in an analogous way to what we have done so far and the number of formal references to the matter is very large [31, 36, 42].

2.4 Markov chains

The subject of this section would deserve its own chapter, but we are not planning to exhaust its details, instead, we are only going to introduce what is sufficient to understand

and develop the theory of Chapter 7. Notwithstanding, a *Markovian process* is a very important concept when modeling biological systems and it can be very fruitful to delve deeper into it. Many references can be suggested [37, 43, 44].

Previously, we talked about single experiments, e.g., to throw a dice once. But what if someone is interested in throwing a dice many times and the specific sequence of results is important? For instance, a gambler needs a specific sequence of results in order to not lose all of his money. Suppose he bets all his money on the occurrence of a 6 and that he does this procedure repeatedly. If the dice shows first a 6, then a 6 again, and then a 4, then he loses all his money in the third round. However, if it shows a 6 and then a 4, then he loses his money in the second round. The order of the results matters! And the results in every round can be modeled as random variables.

We define a *stochastic process* as a family of random variables $\{X_t\}$ indexed on $t \in I \subset [0, +\infty)$. If t assumes only integer values, then it is a discrete-time stochastic process; if t can assume any non-negative real values, then it is a continuous-time stochastic process. This is a very powerful tool, which is able to describe many different dynamical systems. The sequence of dice was just a very simple example. The price of an asset in the stock market is very well described as a stochastic process [45]; the motion of pollen grains in water [46], the electrical current in a noisy circuit [47], the number of predators and prey in a given environment [48], the human population on Earth [49], they are all described as sequences of random variables indexed on time. Even deterministic systems can be accommodate in this definition with suitably chosen probability distributions (Dirac delta functions).

For a discrete-time stochastic process on integer indexes $\{0, 1, 2, \ldots\}$, up to time t, the process is completely described by

$$\{X_0 = x_0, X_1 = x_1, \ldots, X_t = x_t\}, \tag{2.4.1}$$

for any sequence of real values $x_0, x_1, \ldots, x_t \in S \subset \mathbb{R}$ that can be assumed by the random variables. Suppose now that given the description of the process up to time t, someone wants to calculate the probability of the next random variable

$$\mathcal{P}(X_{t+1} = x_{t+1} | X_0 = x_0, X_1 = x_1, \ldots, X_t = x_t).$$

The *Markovian property* is defined by

$$\mathcal{P}(X_{t+1} = x_{t+1} | X_0 = x_0, X_1 = x_1, \ldots, X_t = x_t) = \mathcal{P}(X_{t+1} = x_{t+1} | X_t = x_t), \tag{2.4.2}$$

i.e., the probability of the next value of the process, given all its history, depends only on its last value. Such system is called a *Markov chain*[4], and the space S is called the *state space* while the values $x \in S$ are the *states* of the system. S can be finite or infinite.

[4] *Higher order* Markov chains (of order k) can also be defined by considering that $\mathcal{P}(X_{t+1} = x_{t+1} | X_0 = x_0, X_1 = x_1, \ldots, X_t = x_t) = \mathcal{P}(X_{t+1} = x_{t+1} | X_t = x_t, \ldots, X_t = x_{t-k+1})$ [44].

Thus, for a Markov chain

$$\begin{aligned}&\mathcal{P}(X_0 = x_0, X_1 = x_1, \ldots, X_t = x_t, X_{t+1} = x_{t+1})\\ &= \mathcal{P}(X_{t+1} = x_{t+1}|X_0 = x_0, X_1 = x_1, \ldots, X_t = x_t)\mathcal{P}(X_0 = x_0, X_1 = x_1, \ldots, X_t = x_t)\\ &= \mathcal{P}(X_{t+1} = x_{t+1}|X_t = x_t)\mathcal{P}(X_0 = x_0, X_1 = x_1, \ldots, X_t = x_t) \qquad (2.4.3)\end{aligned}$$

and then (and simplifying the notation),

$$\mathcal{P}(x_0, x_1, \ldots, x_t) = \mathcal{P}(x_t|x_{t-1})\mathcal{P}(x_{t-1}|x_{t-2})\ldots\mathcal{P}(x_1|x_0)\mathcal{P}(x_0). \qquad (2.4.4)$$

A Markov chain is then completely defined by the *transition probabilities*

$$P_{ij} = \mathcal{P}(x_i|x_j), \qquad (2.4.5)$$

which corresponds to the probability of transition from the state j to the state i.

From the total probability law (Eq.(2.2.8)),

$$\mathcal{P}(X_{t+1} = x_i) = \sum_j P(x_i|x_j)\mathcal{P}(X_t = x_j). \qquad (2.4.6)$$

If $P(x_i|x_j)$ is a function not only of x_j, but of $\mathcal{P}(X_t = x_j)$, then this system is called a *non-linear Markov chain* [50]. We may simplify now the notation as $\mathcal{P}(X_t = x_j) \to \mathcal{P}_t(x_j)$.

2.4.1 Perron-Frobenius theorem

Consider the state space to be the finite set $\{n_1, n_2, \ldots, n_m\}$. Thus, equation (2.4.6) defines a matrix equation. Consider $\mathcal{P}_t = (\mathcal{P}_t(n_1), \mathcal{P}_t(n_2), \ldots, \mathcal{P}_t(n_m))$ and the *transition matrix* $\mathbb{P}$, whose elements are $(\mathbb{P})_{ij} = P_{ij}$. Then,

$$\mathcal{P}_t = \mathbb{P}\mathcal{P}_{t-1}, \qquad (2.4.7)$$

and for a non time-dependent transition matrix,

$$\mathcal{P}_t = (\mathbb{P})^t\mathcal{P}_0. \qquad (2.4.8)$$

The transition matrix has two important properties [44]

$$\sum_{i=1}^{m}(\mathbb{P})_{ij} = 1 \text{ and } (\mathbb{P})_{ij} \geq 0,$$

which defines a *(left) stochastic matrix*.

Given a stochastic matrix $\mathbb{P}$, it is called *irreducible* if, given i and j, there is a positive integer $L(i, j)$ such that $\mathbb{P}^L > 0$. In other words, it means that after a finite number of steps, there is a positive probability to occupy a state i given that the system started on

j. Moreover, if there is an integer L such that for every i and j, $\mathbb{P}^L > 0$, then the matrix is called *regular* [44].

Important properties of a stochastic matrix $\mathbb{P}$ are [44]:

1. $\mathbb{P}$ has at least one eigenvalue $\lambda = 1$.
2. Any eigenvalue of $\mathbb{P}$ satisfies $|\lambda| \leq 1$.
3. The eigenvector of the eigenvalue $\lambda = 1$ (which can be degenerate) corresponds to a vector with no negative components.
4. *The Perron-Frobenius theorem.* If $\mathbb{P}$ is irreducible, then the eigenvalue of $\lambda = 1$ is non-degenerate. Also, the eigenvector has all components strictly positive. Non regular matrices can also have other eigenvalues such that $|\lambda| = 1$.
5. For a regular matrice, all eigenvalues except for the $\lambda = 1$ satisfy $|\lambda| < 1$.
6. Let $\mathcal{P}$ be the eigenvector of the eigenvalue $\lambda = 1$ of a regular matrix. Then, for $t \to \infty$, $\mathbb{P}^t$ converges, and all its columns equal $\mathcal{P}$.

Therefore, for a regular stochastic matrix, equation (2.4.8) has a unique stationary solution, (for $t \to \infty$) regardless of the initial condition $\mathcal{P}_0$, and it is equal to the eigenvector of eigenvalue $\lambda = 1$ of the transition matrix,

$$\mathcal{P} = \mathbb{P}\mathcal{P}. \tag{2.4.9}$$

We shall finish now our introduction to Markov chains. Much more can be studied, but what we have presented should be enough for the developments in Chapter 7. Now, let us move on to another important subject of this thesis: the evolution theory and species formation.

Chapter 3

A bit more of Evolution

The word "evolution" has gotten its place in our daily vocabulary; people talk about the evolution of cars, of medicine, of technology, and none of these pose the actual meaning of what we call biological evolution. In biology, *evolution* means "descent with modification" [1]. It is a very straightforward definition but it opens a vast world of implications. To start with, evolution does not mean individuals are *getting better*, or *improving*, it only means they change *over generations* (descents). What makes individuals change is now known to be molecular processes, variations occurring at the gene expression level, mutations appearing during the cell replication steps, the horizontal transference of pieces of genetic material, or the recombination of parental DNA [8]. Nonetheless, what keeps these modifications in a population, or drives them away, thus shaping diversity, together with what produces them, are known as *evolutionary forces*.

When the evolution theory was conceived, the biomolecular origins of diversity were not known. After the rediscovery of Mendelian genetics at the beginning of the 20th century [51], the theory introduced by Wallace and Darwin [52] was perfected by Fisher, Haldane, Wright, and others, and is now called the Synthetic Theory of Evolution [1, 7]. On the other hand, even in Darwin's time, it already made sense that, since variants were introduced, there would be mechanisms for them to remain, to accumulate, and to give rise to so distinct organisms that a whole diverse set of life forms would be observed. That is about such sense and the mechanisms behind, it that we aim to talk about in this very short introduction, reaching, in the end, the concept of species and speciation.

3.1 Evolutionary forces

When an organism different from its ancestors (let us call it a mutant) appears in a population, it can give rise to new individuals that are similar to it. Think about a red bacteria that, due to a mutation, appears in amid a blue bacteria population. If it gets to replicate, then more red bacteria would emerge, and, in an environment with limited availability of nutrients, there could be fewer blue bacteria than in a previous generation. Imagine then that, suddenly, a purple bacteria appears within the red population and starts to replicate; the red population starts to decrease while the blue remains constant. These changes in frequency appear, are sustained, or are carried away by the *evolutionary forces* that act on the system and, particularly, on each trait (blue, red, and purple, in the example).

There are four different evolutionary forces, which can act in different organism scales and account for different effects in the composition of a community:

- **Mutations** are the primary source of diversity;
- **Gene flow** sustains the genetic diversity;
- **Random genetic drift** decreases diversity;
- **Selection** accounts for *adaptation* and it increases or decreases diversity, depending on the system.

Let us say a few words about each of these concepts.

3.1.1 Mutations

A mutation is a difference that appears in the genetic material when it is compared with the previous generation. They thus can affect the gene expression of important proteins, usually resulting in the cell's death. They can happen naturally (molecular decay), due to *errors* in the molecular processes behind replication or can also be induced by radiation or mutagenic chemicals. In sexually reproducing individuals, when mutations affect *reproductive* cells, they can be passed to the next generation. On the other hand, *somatic* cells do not pass mutations forward, although a whole lineage of somatic cells is going to be affected [8, 53, 54, 55].

Suppose that, in a given community, no mechanism can keep different organisms alive, and every mutant dies after one generation. Thus, the only way to sustain diversity is to keep a constant mutation rate: every couple of generations a new variant appears. This is the role of mutations: it sustains genetic variability, thus being the ultimate source of diversity [54, 55].

Different types of mutations can be identified. When happening within a sequence, from one generation to the other, nucleotides can be *inserted* or *deleted*, which can affect the expression of many genes; they can also be *substituted* for a different base, or subsequent bases can be *inverted*. Moreover, another type of mutation may occur in larger groups of genes called *chromosomes*. The *Down Syndrome* [56], for instance, is a human condition resulting from the existence of three copies of the chromosome number 21 (in non-mutant cells, there are only two copies).

An important feature of mutations is their random aspect. Mutants are not predictable, i.e., there is no way to know when or what is going to happen, and that is where probability theory and population genetics come to the stage.

3.1.2 Gene flow

The exchange of genes among populations is called *gene flow* [57]. Individuals can interchange genes horizontally [58], e.g. bacteria can perform what is known as *conjugation* [59], acquiring small fragments of genetic material dispersed in the cytoplasm of different cells; sexual reproduction mixes the genetic material of different individuals within the offspring, which can homogenize different populations since it allows them to evolve – to change – in correlated ways.

Some individuals have multiple copies of the same set of genes. Diploid species have two: one copy coming from the mother and the other from the father. The copies may be different, having different *alleles* of a given gene. However, some alleles may show *dominance* behaviors over others, thus governing the gene expression regardless of the other copy, the *recessive* allele. Hence mating is not always going to result in diverse phenotypic populations, but it can silently spread genes out, sustaining genetic diversity.

Mobility is an important factor for gene flow, as well as dispersal and migration [60]. Constant gene flow within a community correlates the evolution of the individuals, providing a major hurdle to diversification in larger scales (species formation).

3.1.3 Genetic drift

While mutations are a source of variability and gene flow is a way of spreading it out, the *random genetic drift* is a sink of diversity [61]. As the name points out, this is a random process, through which genes can be lost. Suppose only one individual carries a mutant allele of a gene. If this individual dies before reproducing, then this allele is going to be lost. Indeed, if only a few individuals carry a given allele, it is easier for it to be lost due to their death. On the other hand, if these individuals reproduce more than the average, they can randomly drive other alleles to extinction. This random frequency change of genetic material is called random genetic drift.

It strongly affects small populations, since they are more susceptible to random events than bigger ones. As an example, suppose you toss a coin three times. The probability of getting only heads is 1/8. If you toss it 5 times, it decreases to 1/64, and it keeps decreasing for larger numbers of tosses. This is the same reasoning behind the fixation of a gene through random drift: the chance of *all* individuals that carry a given allele to die without any offspring decreases as the population size increases[1].

Mutations are then balanced by genetic drift. As soon as a mutant appears, it needs to "win" over random death events to pass this variation on. In an infinite population, any allele with positive frequency is conserved as long as it does not display disadvantages for survival when compared to other alleles.

3.1.4 Selection

As the reader may have noticed, the sink and the source of variability are governed by random events: mutations appear randomly and genes can be lost randomly. *Selection* is what gives some direction to randomness. If a given trait increases the survival rate of an individual, then its reproduction rate may increase as well as its frequency. It is like tossing a biased coin, in which it is more likely to get "all heads" than "all tails". A trait that survives under this mechanism is said to be *positively selected*, while if it decreases the rates of reproduction, it is said to be *negatively selected*. A trait that evolves without any selection is called *neutral* [1].

[1]In probability theory, this correlates with *first passage* problems, which is usually exemplified with the *gambler's ruin*: once a gene is lost, it is lost (although nothing forbids it from appearing again because of a mutation), like a gambler when it runs out of money (although nothing forbids it from receiving more money from another person). This effect is also related to *absorbing states* of Markov chains [62, 63].

Selection can be *natural*, as pointed out by Darwin in *On the Origin of Species* [64]. He noticed that finches have different beak types in different Galapagos islands, but they were all well *adapted* to the existent food type. *Natural selection* explains *adaptability*. Bacteria that are resistant to a given antibiotic are not *better* than the non-resistant, but they are well adapted, i.e., they can survive and successfully replicate, in an environment containing that chemical [65]. Every living system is subjected to natural selection. The frequency of adults able to digest lactose – a sugar present in milk and its derivatives – is higher in regions where the society developed around dairy farming. Mammals stop digesting lactose after a certain age, at which they stop to breastfeed, but humans developed the ability to digest it at older ages, giving them the possibility to survive in communities where diet was based on milk derivatives. This ability is tracked back to a mutation that happened only a few thousand years ago, which was positively selected by the local food availability [66].

But selection can also be *artificial*. Scientists are constantly choosing in laboratory different strains of plants with higher resistance to certain environmental conditions [67]. A certain corn variety could be better to resist in dry weather than in rainy one. Thus it may not be feasible to harvest it in a tropical region, but after carefully separating it in lab cultures, they could be exported to dry countries. More docile lineages of animals are favored for farming over aggressive ones; plants are selected concerning their "food quality" over their natural survival.

Selection is also classified according to its effects. If it keeps diversity as it is, eliminating very different traits and favoring the common ones, it is known as *stabilizing selection*. When it constantly moves the common trait to a different one, it is a *directional selection*. Selection can also divide the population into different traits, favoring extreme characters over the common ones, which is known as *disruptive selection* [1].

Other types of selection can be identified, such as *sexual selection* [68] and *kin selection* [69]. Some traits do not confer survival benefits but even so, increase the reproduction rates of individuals. Some of them are correlated to good survival genes, despite being very inefficient, like the beautiful and big tails of peacocks [70, 71]. Sexual selection acts in favor of these traits, making individuals to preferentially mate with those that pose a specific character over the others. Individuals also get benefits from genealogically closer individuals than from farther ones. This is called kin selection. An extreme case is infanticide practices documented in many species [72, 69, 73]. In non-human primates, for instance, males kill the offspring of a female and are then able to reproduce with her. This keeps the community with close related genetics, diminishing the gene flow between far lineages.

3.2 Quantitative models of evolution

It is possible to model a system subjected to evolution by including the evolutionary forces one by one. Let us briefly study, as an introduction to the mathematical journey we are going to engage in the next part of this thesis, three different models, which show the quantitative descriptions of the concepts introduced so far.

3.2.1 The Hardy-Weinberg equilibrium

We must now introduce other concepts of genetics. The set of all genetic material contained in an individual is called its *genotype*. The genes are expressed as proteins whose "sum" ends up in observable traits, which describe the *phenotype* of the individual. Different genotypes can be related to the same phenotype. For communities evolving under sexual reproduction, there are specific cells that are combined to generate the offspring, the *gametes*. Gametes are formed through a cell division named *meiosis*, which give rise to four daughter cells. In diploid individuals, each daughter cell contains only one copy of the genetic material: two with the paternal copy and the other two with the maternal copy [1].

In such a system, a trait is hence given by the combination of the genes coming from its parents. The different forms a gene can assume are called its *alleles*. Suppose a given gene has two alleles, i.e., it is *biallelic*. An important observation in genetics is that genes show *dominance* patterns, in which one of the alleles (the *dominant*) can suppress the effect of the other, the *recessive*. Let us call them A for the dominant and a for the recessive. Whenever A is present, the effect of the recessive is not shown. Therefore, an individual can have 3 different genotypes regarding this gene:

$$AA; Aa; aa$$

and two different phenotypes: AA and Aa expresses the same observable trait (e.g. eye colors) while aa express a different one. If the two copies of a gene in an individual have the same allele, it is called an *homozygote*, and when they have different alleles, it is an *heterozygote* [1].

A very natural question that arises concerns the frequencies of genotypes and phenotypes in a population. Given the number of individuals of a given genotype, what can someone expect concerning the next generation? The answer to this question was given first by W. Weinberg in 1908 and, later and independently, in the same year, by G. H. Hardy [74, 75].

Suppose an *indefinitely large population* with three different genotypes and proportions:

AA	Aa	aa
P	Q	R

Therefore, the proportions of the gametes (given the rules of meiosis) are given by:

A	a
$p = \mathrm{P} + (1/2)\mathrm{Q}$	$q = \mathrm{R} + (1/2)\mathrm{Q}$

Thus, in the next generation (represented with a prime '), if all alleles are equally likely to reproduce, the genotype proportions are going to be:

AA	Aa	aa
$\mathrm{P}' = p^2$	$\mathrm{Q}' = 2pq$	$\mathrm{R}' = q^2$

with $p + q = 1$. Observe that, in the following generation,

$$\begin{aligned}
&\mathrm{P}'' = (\mathrm{P}' + (1/2)\mathrm{Q}')^2 = p^2 \Rightarrow \mathrm{P}'' = \mathrm{P}',\\
&\mathrm{Q}'' = 2(\mathrm{P}' + (1/2)\mathrm{Q}')(\mathrm{R}' + (1/2)\mathrm{Q}') = 2pq \Rightarrow \mathrm{Q}'' = \mathrm{Q}',\\
&\mathrm{R}'' = (\mathrm{R}' + (1/2)\mathrm{Q}')^2 = q^2 \Rightarrow \mathrm{R}'' = \mathrm{R}',
\end{aligned}$$

and thus the genotypic proportions are conserved from the first to the second generation. This is called the *Hardy-Weinberg equilibrium*, which is achieved after one reproduction step and is therefore a stable equilibrium. The Hardy-Weinberg's law starts the field of *population genetics* and it is the starting point for quantifying the effect of other evolutionary forces on a given system [75].

3.2.2 The effect of selection

The previous model considers a situation in which all individuals have the same chances of reproduction, called a *panmictic* population. Because its size is considered to be indefinitely large, there is no random drift; also, mutants are not introduced in the system. Therefore, the only evolutionary force considered is the gene flow. We shall now consider a case in which the dominant allele is positively selected. To quantify the selection, we introduce the concept of *fitness*, which is a measure of adaptation. The greater the fitness of an individual, the greater its reproductive success, i.e., its chance of reproduction [1].

Let the *selection coefficient* s be a decrease in the fertility rate of the recessive homozygote, and let the phenotype proportions be:

Genotype:	AA	Aa	aa
Proportions:	P	Q	R
Fitness:	1	1	1-s

Hence, in the next generation,

$$\begin{aligned}
\mathrm{P}' &= \overline{N}(\mathrm{P} + (1/2)\mathrm{Q})^2 = \overline{N}p^2, \\
\mathrm{Q}' &= \overline{N}2(\mathrm{P} + (1/2)\mathrm{Q})(\mathrm{R} + (1/2)\mathrm{Q}) = \overline{N}2pq, \\
\mathrm{R}' &= \overline{N}(1-s)(\mathrm{R} + (1/2)\mathrm{Q})^2 = \overline{N}(1-s)q^2,
\end{aligned}$$

where $\overline{N}$ is a normalization factor calculated with $\mathrm{P}'+\mathrm{Q}'+\mathrm{R}' = 1$, hence $\overline{N} = 1/(1-sq^2)$. These frequencies are no longer in equilibrium, as can be seen by the new frequency p' of the dominant allele A,

$$p' = \mathrm{P}' + (1/2)\mathrm{Q}' = \frac{p}{1-sq^2},$$

which increases by an amount

$$\Delta p = p' - p = p\frac{sq^2}{1-sq^2}. \tag{3.2.1}$$

Therefore, selection is able to fixate a gene in the population, by increasing its frequency in the course of generations. This model is due to Haldane [75] and we can easily include an effect of mutations between both alleles. Suppose A is converted into a with mutation rate μ. Then, if the system is found to be in equilibrium,

$$\Delta p_{\text{selection}} = -\Delta p_{\text{mutation}}$$

where $\Delta p_{\text{mutation}} = -\mu p$. This way, for small selection s,

$$q = \sqrt{\mu/s} \tag{3.2.2}$$

and thus the selection-mutation ratio is easily calculated as the square of the recessive allele frequency [75].

3.2.3 The effect of genetic drift

The Hardy-Weinberg equilibrium is satisfied in the absence of genetic drift. When populations are finite, there is a non zero chance of losing genes due to random fluctuations. If the nutrients are limited, an environment can support only a certain number of individuals, which gives rise to variations in the number of genes from one generation to the next by pure chance. Eventually, one gene disappears and the other is fixated (in the absence of mutations).

(Here we follow the discussion in M. Ridley, 2006 [1]).

Suppose a population of N diploid individuals with a biallelic gene; there are thus $2N$ genes. In the Hardy-Weinberg equilibrium, the homozygous frequency is kept the same. How does it change for a finite population? Suppose that a gamete with an allele a finds another gamete with allele a coming from the same progenitor, in a process called *self-fertilization*. This happens with probability $1/2N$.[2] But a gamete can also combine with another gamete, containing the same allele, coming from different progenitors. The chance of such case is $1 - 1/2N$.

Now, suppose that, in the parental generation, the frequency of the allele A is p, and a is $q = 1 - p$. Thus, the chance of forming a homozygote with two independent gametes is $(p^2 + q^2)(1 - 1/2N)$. But $p^2 + q^2 = f$ is the homozygose frequency in the parental generation[3]. Combining this result with the self-fertilization probability, we find that

$$f' = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) f \tag{3.2.3}$$

is how the homozygous frequency changes from one generation to the next. Notice that $f = 1$ is a stable equilibrium of the system, and thus one gene is going to disappear due to the random genetic drift.

Including mutations

If we add a mutation chance μ for each gene, with which an allele is changed to the other, then the chance of non-mutation is $1 - \mu$, and the homozygosity changes according

[2]The picture here is a population whose individuals release their gametes in the environment, and the gametes find each other. The number of gametes per individual is very large and it is considered to preserve the parental frequencies

[3]Notice that this quantity is given at the gametes level. A population composed of only AA and aa has only homozygotes, but f equals 1/2. Nonetheless, it equals the fraction of homozygote individuals in the Hardy-Weinberg equilibrium.

to

$$f' = \left(\frac{1}{2N} + \left(1 - \frac{1}{2N}\right) f\right)(1-\mu)^2, \tag{3.2.4}$$

whose equilibrium $f' = f = f_{eq}$ is now given by

$$f_{eq} \approx \frac{1}{1+4\mu N} < 1, \tag{3.2.5}$$

where the approximation holds for $\mu \ll 1$, and we then see that mutation can balance the effects of diversity loss [1]. The result in equation (3.2.5) is going to appear again in this text, but for a system with many loci, thus being a quite remarkable result.

The Wright-Fisher model

For completeness, we shall mention the Wright-Fisher model, in which a haploid, asexual finite population evolves in non-overlapping generations. It was introduced by Sewall Wright (1931) [6] and Ronald Fisher (1930) [4]. In the model, individuals carry the allele *a or* the allele *A*. If an individual reproduces, it is going to generate offspring carrying the same gene. Given a fixed population size N, and the number X_t of individuals that carry the gene a at generation t, we can ask what the number of individuals carrying a in the next time step $t+1$ is. Hence, the system defines a Markov chain, with transition probabilities [76]

$$P_{ij} = \binom{N}{i}\left(\frac{j}{N}\right)^i \left(1 - \frac{j}{N}\right)^{N-i}, \tag{3.2.6}$$

to move from state $X_t = j$ to $X_{t+1} = i$. In other words, this model is like an urn with two colors of balls, from which someone samples sequences of balls with replacement.

We can see that the states $j = 0$ and $j = N$ are absorbing states, i.e., once the system reaches $X_t = 0$ or $X_t = N$, it is going to remain there with probability 1. These states correspond, respectively, to the fixation of the allele A and the allele a.

We are going to use the same type of modeling when talking about abundance distributions in Chapter 7.

3.3 What are Species?

The pinnacle of diversity is to be able to distinguish between types of individuals. To claim that an organism is not the same as another is where daily observations come to place and the idea of biodiversity makes sense to the common knowledge: a cat is not the same as a dog, an eagle is not a falcon, a bee is not ant, or a palm tree is not a pine tree. These different types of living beings are known as *species*, the thinnest division of life forms. But what does separate them? What does differentiate a *Panthera leo* (lion) from a *Panthera tigris* (tiger)?

Another important question is whether the concept of a species, i.e., the idea of dividing types of living beings into groups, is something natural or artificial [77]. The discussion is not fruitless but still unsolved; it regards the types of communities and the reproduction types they have. We hence stick to the fact that life can be shown in different forms, from mosquitoes to bears, from bacteria to plants, and then there are different

"things", units that we can recognize. These units are the *species* and then we take them as real entities, not mere artifacts of science. However, what defines the boundaries of closely related species is still a contentious problem [78]. In practice, taxonomists have no problems in differentiating lions from tigers (not even a child would have this problem, I must say), but what if a tiger gives birth to tigers that have a very different pattern of lines? Do they belong to the same species?

In order to easily state how far this subject can dive, we observe that there are more than 30 different concepts of species[4] [19, 79]. Two important concepts are:

- **Biological concept**: *"interbreeding natural populations reproductively isolated from other such groups;"* [19]
- **Ecological concept**: *"a lineage (or a closely related set of lineages) which occupies an adaptive zone minimally different from that of any other lineage in its range and which evolves separately from all lineages outside its range."* [19, 80]

The biological species concept was introduced by Dobzhanski (1935) and restated in many ways, being very advocated by Ernst Mayr (1942), who detains a lot of its credits [81]. This concept defines a species as groups of individuals that can interbreed, i.e., mate and produce fertile offspring. Reproduction leads to closely related individuals, who share large sets of genes and hence have many characters in common. Therefore it well suits the works of taxonomists but does not exhaust it [5]. The reproductive isolation, for instance, is not easy to determine – indeed not possible in the fossil record. We shall emphasize that isolation here regards the existence of reproductive barriers, which are characters that evolved and led to the hindering of reproduction. We thus exclude geographical barriers from this definition.

If a given species inhabits two different geographical places, there is no breeding between the two populations, thus they do not naturally reproduce, but it does not mean they are not the same species, there is just no easy way to test their breeding behavior. Reproductive isolation can be distinguished between *pre-zygotic* and *post-zygotic*. In pre-zygotic isolation, individuals do not copulate due to some mechanism: body incompatibilities, non-matching of fertile periods, sexual selection, etc, but it does imply that, if they mate, they cannot produce a fertile offspring. Post-zygotic isolation means that the different species can mate and the zygote is formed, but this is non-viable or sterile [1].

The borders of a biological species can therefore be ill-defined since reproductive isolation is not straightforwardly testable. The other concept listed here, in contrast, is the ecological species concept, which decides between species according to their *niche* occupation. The niche is the habitat and the resources consumed by the individuals, which therefore define a species. Adapting to an environment can lead to reproductive isolation to some extent. For instance, reproduction between different ecological species can decrease the fitness of the offspring, who could be poorly adapted to any of the parental niches. Ecological adaptations can thus produce a biological species with no need for intricate arguments, but this is not valid for every case [1].

The species problem is indeed a subject for theoretical biology, not an exactly great issue for species identification. Citing Orr (2022), "*taxonomists have been describing*

[4] As stressed by Wilkins [2], the concept is actually only one: *species*, which has many different *conceptions* or definitions.

[5] Taxonomists classify species according to morphology and other observable and measurable traits [1].

species for centuries (Costello 2022), so a universally accepted definition for species is clearly not a prerequisite for taxonomic research."[82].

3.4 Species formation processes

The mutability of species was not a problem for pre-Darwinian beliefs; the bible itself does not request the constancy of living beings and even Christian thinkers, such as Saint Agostine, accepted that old species could give rise to new ones [2]. Linneaus intervened in the name of the *species-fixism* in 1735, when he came up with his classification system [83]. Evolution puts this idea in check, by successfully explaining diversity and (importantly) adaptation. The forces involved in the diversification process have already been presented, but its utmost effect is the emergence of new species, which we address in this section now.

By species, we mean the biological concept. New species are then formed when individuals evolve to reproductively isolated groups. Natural selection is an important force in this process, which is called *speciation*. When individuals are separated and there is no gene flow between different groups, evolution in each subgroup becomes uncorrelated, and different selection pressures can drive different adaptations. What is surprising (but well documented) is that ecological adaptations are followed by reproductive isolation. In an experiment by Diane Dodd (1984, 1989), different populations of flies (*Drosophila pseudoobscura*) grew in different media (one starch-based and one maltose-based) [84, 85]. After getting adapted they were shown to be under (pre-zygotic) reproductive isolation.

Geographical modes of speciation

In Dodd's experiment [84, 85], it is noteworthy that the gene flow was interrupted between the populations. This situation is called *allopatry* and hence the emergence of reproductively isolated species without gene flow is called *allopatric speciation*. This is a non-contentious process, with lots of evidence, ranging from natural observation to lab experiments [77]. The strife arises when gene flow is not completely absent from populations. When individuals still mate, regardless of their place of birth [86], the community is *sympatric* and when species emerge in such a case, it is known as *sympatric speciation*. These different modes of speciation directly correlate with the communities' spatial structure, thus they are called *geographical modes of speciation* [77]. Geographical barriers, like rivers and mountains put populations in allopatric conditions, while the coexistence in a lake, for instance, is likely to put the aquatic life in sympatry. Intermediate cases, with non-complete gene flow, are called *parapatric*.

The sympatry problem

The effects of divergent selection in a sympatric community were already considered by Darwin as an important process of diversification and species origin, but as Mayr (1963) showed inconsistencies between theory and observations, a lot of work was devoted to show that this mode of speciation, although not the norm, could be reproduced in the laboratory as also found in the field [77].

It is rather counterintuitive that different species can be formed within a community in the presence of gene flow. Recombination is taught to balance mutations and thus

selection should play a major role if species are in fact emerging in such a situation [87]. On the other hand, there is some evidence of this process [88], but the definition of what a *sympatric speciation* really means should be under debate [89]. Gavrilets [86] points out the need for a more generical definition (the one we used here), for modeling purposes, as a way of diminishing the number of parameters. But an "easier to take" is the emergence of species within the cruising range (which is similar to the one given by Mayr, 1942 [77, 86])[6].

C. H. Martin calls [89] the definition above as *easy* , and argues that the "*truly elusive beast, that original sought-after Lernean Hydra, is empirical evidence in nature for sympatric speciation under the population genetic definition (Richards et al. 2019).*" [89] For the easy case, Coyne and Orr established four criteria of analysis [77]:

1. *Species should be found in sympatry*;
2. *Species should be under reproductive isolation*;
3. *Sympatric taxa must be sister groups*[7] *and not resulting from hybridization processes*;
4. *Evolution and geography should make the existence of an allopatric phase very unlikely.*

Martin then includes a fifth challenge, that any secondary gene flow (coming from other species) should be proven to not contribute to reproductive isolation [89].

But apart from the defiances enlighted by genomic data, there are examples of (what can be a case of) sympatric speciation. Perhaps the most elusive are the cichlid fishes, a family of fishes (*Cichlidae*) that can be found in the whole African continent, Central and South America, as well as other smaller locations. A study with two sister species in Lake Victoria[8] shows that they are in reproductive isolation due to sexual selection [90]. The males of the species *Pundamilia nyererei* and *Pundamilia pundamilia*, which are sympatric and sister species inhabiting the lake, have different color patterns. Under white light, the females mate with their own species. However, under monochromatic light, when the colors are indistinguishable, there is interspecific breeding.

Another example of cichlids comes from the lake Apoyo, in Nicaragua, where the two species *Amphilophus citrinellus* (midas cichlid) and *Amphilophus zaliosus* (arrow cichlid) live in sympatry [91]. The authors show lots of evidence for the sympatric speciation process from which *A. zaliosus* evolved from *A. citrinellus*, including that the Midas cichlid was seeded only once by an ancestral in the lake. Another study [92] also shows strong evidence, even regarding secondary gene flow, for sympatric speciation of other species of midas fish on different other crater lakes in Nicaragua, close to Lake Apoyo.

Sympatric speciation, albeit still contentious, does seem to play an important role in natural communities. Different definitions of what it is and which are its constraints

[6]This definition is mentioned here because it directly defines the processes we discuss in Part II as sympatric processes. Notwithstanding, the process also falls into Gavrilets' definition [86], but because of its absence of spatial structure, there is no sense in the expression *place of birth*.

[7]It means the species are the closest relatives they have, descending from the same *node* in a genetic tree.

[8]One of the *African Great Lakes*, in East Africa, with its area extending through Tanzania, Uganda and Kenya.

entangle with the challenges in defining a universal concept of species. In speciation research, Gavrilets [86] points out the necessity of general analytical results, as models have been developed to account for specific cases in numerical ways, being hardly reproducible and reused in different contexts – he seems to outline the same general issues concerning agent-based modeling. But it is important to highlight that research in evolution did not stop and neither on speciation. The genomic revolution is still one step more towards the end of the debates and the controversies concerning diversity origins, but it is still not the last one (in the case there is a last one).

We have now finished this very brief introduction to the theory of evolution. A lot is still missing, but in the case I have put some curiosity pill into the reader's mind, there are many specific and complete books on the field [1, 93, 77]. We aimed to reach speciation and its counterintuitive sympatric case, to which we hope to make some contribution. Nonetheless, it is going to be even more curious that the model we present in Part II, introduced by Derrida and Higgs in 1991 [94], resembles the sympatric species formation but without differential fitness: thus in the absence of disruptive selection. But let us take baby steps and introduce now (in an even more compact way) the next and last tool of this manuscript: *the network theory.*

Chapter 4

Network theory in a nutshell

This chapter is as brief as it can be, acting much more as a glossary than a proper introduction to the field of complex networks. This field is a very rich world, with applications ranging from social sciences to electrical power grid systems, and here are only presented the necessary concepts that are used later in the text. The reader is then referred to specific literature for further readings [95, 96, 97].

4.1 What are networks?

Complex systems are those whose behavior cannot be directly explained by the sum of the behavior of its smaller components [98]. *Emergent properties* appear in such systems, in spite of the involved rules do not directly describe the observed behavior. A beautiful example is the flock of birds, or fish schools [99], in which interactions are modeled as being short-distance but their movement shows long-distance correlation lengths [100]. The actual description of such systems passes through depicting which animal interacts with whom. For instance, by numbering them, we could say that the bird number 52 interacts with the 51 and the 53. The *interaction network* is then the set of birds and the set that describes who interacts with whom.

Metabolites in a cell can also be described in the same way , with their connections defined as chemical reactions taking from one compound to another. Actors who worked together in a movie can also define a network as scientists who have coauthored papers [101, 102, 103]. Hence, a network is a list of the smallest components (called *nodes*) of a system and the connections between them (called *links*).

The mathematical theory that deals with networks is the *graph theory* [104], in which nodes are called *vertices* and links are called *edges*. A graph $\mathcal{G}$ is an abstract mathematical formulation of a network and it is defined as the pair of sets $(\mathcal{V}, \mathcal{E})$, where the set $\mathcal{V} = \{v_1, \ldots, v_N\}$ denotes the vertices and the set $\mathcal{E} = \{e_1, \ldots, e_M\}$ denotes the edges. The words "Graph" and "network" in this text are then used interchangeably.

4.1.1 Definitions and characterization

There are important definitions in graph theory, as well as features of networks that distinguish them. Let us introduce now some of the important concepts, which are going

to be used further in the text [96, 97].

- **Network size**. The size N of a network is the number of its nodes, i.e., the cardinality of the set $\mathcal{V}$: $N = \#(\mathcal{V})$.

- **Directed and undirected graphs**. The connections between nodes can be *directed* or *undirected*. For instance, a paper A may cite a paper B, however, B does not cite A. On the other hand, if Alice publishes a paper with Bob, then Bob has published a paper with Alice. Hence direction can be a feature of links or not. When all links are directed, then the network is directed. If all nodes are undirected (as for the networks that are going to appear in the following chapters), then the network is said to be undirected.

- **The adjacency matrix**. A network is defined by its nodes and links. A link does not need to be named: it can be given according to its initial and final nodes, e.g., the link (i, j) connects node i to node j. Therefore, a network can be fully described by a square $N \times N$ matrix $\mathbb{A}$, with elements

  $$A_{ij} = 1, \text{ if there is a link between nodes } i \text{ and } j;$$
  $$A_{ij} = 0, \text{ otherwise.}$$

  The matrix $\mathbb{A}$ is called the *adjacency matrix*. If $\mathbb{A}$ is symmetric, $A_{ij} = A_{ji}$, then the network is undirected. Notice that if $A_{ii} = 1$ for some i, then the node i is connected to itself, which is defined as a *loop*.

- **Complete network**. If every node is connected to all other nodes (without loops), then the network is called *complete*. For a complete network, $A_{ij} = 1$ for every $i \neq j$ and $A_{ii} = 0$ for every i.

- **Path between nodes**. If one can follow the links in a network (respecting their directions), from a node i to another node j, then there is a *path* between nodes i and j. A path is then a sequence of nodes that are connected by links.

- **Connected nodes**. Two nodes i and j are then *connected* if there is a path between them. If all pairs of nodes of a network are connected, then the network is also called connected. An irreducible matrix (section 2.4.1) is thus a stochastic matrix that describes a connected network.

- **Components of a network**. In a network, some subsets of nodes might be not connected to the others. A connected subset of nodes of a network is called a *component* whenever the addition of another node to this set makes it *not* connected.

- **Degree of a node**. The number of other nodes j that a node i is connected to in an undirected network is called the *degree* k_i of the node i. For instance, the degree of any node in a complete network of size N is $k_i = N-1$. In terms of the adjacency matrix, the degree is given by

  $$k_i = \sum_i A_{ij}. \tag{4.1.1}$$

- **Degree distribution**. When drawing a random node i of a network, its degree follows a probability distribution $p_k = \mathcal{P}(k_i = k)$, called the *degree distribution*.

For a given network, the degree distribution is given by the histogram $p_k = N_k/N$, where N_k is the number of nodes with degree k. The mathematical form of such distribution has been shown to play an important role in dynamic behaviors displayed by networks. For instance, the *epidemic size* of a given disease outbreak[1] changes if the population is connected as a Poissonian or as a power-law degree distribution [105, 106].

4.2 Random networks

When modeling real systems as networks, deciding which links are connected to each other is a very important step. For instance, it may be obvious from road maps how to describe the cities in a given region as a network, but if one is interested in obtaining general traffic properties of a non-specific place, "prototypical" networks should be designed. Suppose the following procedure: start with a set of N nodes; then connect each pair of nodes with probability p. The resulting graph is called a *random network*.

The *random network* model first appeared in 1951, in a paper by R. Solomonoff and A. Rapoport [107], who proved the existence of a *transition* in this scheme[2]. Around a decade later, in a series of papers, P. Erdös[3] and A. Rényi consistently studied the properties of this network, which is now also known as Erdös-Rényi network [97, 111].

4.2.1 Degree distribution

What are the properties of a random graph? For every node, there can be $N-1$ links, and each of them exists or not with the same probability p. This is the same as asking how many times a biased coin shows up a head in a total of $N-1$ tosses. If the success (tossing a head) is given with probability p, the probability of tossing head k times is given by the binomial distribution

$$p_k = \binom{N-1}{k} p^k (1-p)^{N-1-k}, \tag{4.2.1}$$

and this is the degree distribution of a random network with parameter p. Thus the average degree, i.e., the average number of links of a node, is

$$\langle k \rangle = \sum_{k=0}^{N-1} k p_k = p(N-1), \tag{4.2.2}$$

[1]The *epidemic size* is the number of infected individuals by the end of the outbreak.

[2]It is remarkable that the authors were interested in mathematical biology and that they provided, in the text, examples in neuroscience, epidemiology, and genetics [107].

[3]Paul Erdös was a very prolific Hungarian mathematician, with contributions to many different fields. His large network of collaborations motivated the definition of the *Erdös number*: the smallest number of links (path length) on the papers coauthoring network a scientist has between itself and Paul Erdös [108]. Mine is 5. You can check yours at `https://mathscinet.ams.org/mathscinet/freetools/collab-dist` [109]. A related definition is the *Bacon number* [110], which connects artists with the actor Kevin Bacon.

and thus the probability of connection p can be measured as a function of the average number of links, $p = \langle k \rangle / (N-1)$ [96].

When constructing a random network, different structures can be observed as a function of p. As $p = 0$, no node is connected to each other, but when $p = 1$, every node is connected to all other nodes. Thus the system must pass through a transition. It can be shown that when $\langle k \rangle = 1$, there is the emergence of a *giant component*, i.e., the size N_G of the largest component in the network is *macroscopic* when compared to the network size N, $(N_G/N > 0)$ as $N \to \infty$. It can also be shown that the random network still passes through another transition, for $\langle k \rangle = \ln N$, above which the network becomes connected [96].

4.3 Scalefree networks

However, real networks display features that are not obtained in the random network model. In 1967, Stanley Milgram showed that social networks are more "connected" than we would expect from random networks: people are much closer than that [112]. In social sciences, Milgram's result became known as the "six degrees of separation", and the concept of *small world* appeared. In 1998, Watts and Strogatz published their work on collective dynamics over a model of networks that would resemble the properties of Milgram's small world [113].

In 1998 and 1999, Albert-Laszlo Barabasi and Reka Albert found that the internet webpages network shows a power-law degree distribution, as also the wires of a computer chip, a power grid, and the Hollywood actors [96, 102, 114]. Real networks hence show more complex structures than pure random graphs, and that is why they are called *complex networks*.

Power-law degree distributions have the form

$$p_k \sim k^{-\gamma} \tag{4.3.1}$$

for $k \gg 1$ and such networks are called *scalefree* when $\gamma \leq 3$. Scalefree distributions with $\gamma > 2$ have finite mean but a divergent second moment. They are called *sparse* because the number of links L it has is $L \sim O(N)$. For instance, we can compare it with a random network. In such a graph, each one of the possible $N(N-1)/2$ links exists with probability p. Thus $\langle L \rangle = pN(N-1)/2 \sim O(N^2)$. Networks with $L \sim O(N^\alpha)$ with $\alpha > 1$ are said to be *dense*. Scalefree networks with $\gamma \leq 2$ are also dense [96].

4.3.1 Barabasi-Albert network

It is remarkable that power-law degree distributions were appearing in so different systems (from computer chips to Hollywood actors), and this fact got the eyes of Barabasi and Albert. They asked which mechanism could reproduce a power-law degree distribution. The answer came in June 1999 as a combination of *growth* and *preferential attachment* [96].

Growth means that networks are not constant in size: there is a continuous increase in the number of nodes and links. Preferential attachment means that new nodes are more likely to connect to already highly connected nodes than to not-so-connected ones. This

idea has actually been introduced before Barabasi and Albert, in a paper by Derek de Solla Price in 1976, where Price modeled citation networks [97, 115].

To raise a scalefree network, Barabasi and Albert proposed the following algorithm [102]:

1. Start with m_0 nodes;
2. Add one new node with $m \leq m_0$ links which connect the new node to the m old nodes;
3. The new vertice connects to the old node i with probability $\mathcal{P}(i) = k_i/\sum_j k_j$.
4. Repeat from step 2.

The resulting network is known as a Barabasi-Albert network and it shows a scalefree property, with degree distribution [95, 116]

$$p_k = \frac{2m(m+1)}{k(k+1)(k+2)} \sim k^{-3} \text{ for } k \gg 1. \tag{4.3.2}$$

These networks also show node degree correlations, i.e., nodes with different degrees are correlated, which does not happen in random networks. They also display *hubs*, nodes connected much more than the average degree [117]. This kind of heterogeneity can change dynamical behaviors over such structures, like speeding up epidemic spreads [105, 106].

Network science comprises a large set of works and results [95]. The number of real systems that can be modeled by its framework seems to be unlimited and there is still a lot to explore. I do not dare to say that this chapter has even scratched the skin of the large body network theory has become, but we should have started somewhere. The theory in the following chapters makes use of the concepts introduced here. Nonetheless, we have used network theory as a very convenient language, but it is much more than that: it is a tool that has evolved by itself, strongly grounded on abstract graph theory but solidly attached to concrete problems.

Part II

The Derrida-Higgs Model

Chapter 5

The Derrida-Higgs model

5.1 About the model

In 1991, Bernard Derrida and Paul G. Higgs introduced a model of population dynamics able to mimic the species formation in a sympatric community [94]. In this model, individuals are described by binary sequences, representing their genomes, and, according to a given mating rule, new generations are created. There is no spatial structure, thus geographical distance does not impose any restriction to mating, characterizing a sympatric community. This way, the population is represented by a set of points in the genetic space (defined by the binary sequences) that diffuse over time as mutations and recombination take place.

When restrictions to the reproduction based on the genetic distance between individuals are included, the individuals can cluster in the genetic space, forming reproductively isolated groups, which we recognize as different *species*. The emergence of these clusters depends on the parameter values, and we recognize the existence of two different phases: a low diversity phase, in which although there is genetic variability, it is not enough to produce more than a single species, and a high diversity phase, in which more than one species is observed.

There is no analytical description for the threshold parameter values between these two phases. In the limit of an infinite genome, Derrida and Higgs conjectured a very simple solution [94], which fails, as shown by de Aguiar in 2017 [118], when the genome is finite. The present Part of this thesis aims to describe the theory we developed on the model and to present the approximated heuristic solution to the *low-high* threshold that we are able to introduce. Although approximated, this solution is a remarkable new result on this very interesting model.

In what follows, we first define the model, then we show the original theory introduced in 1991, followed by the analytical theory we developed. After stating the problem, the heuristic solution and its ansatz are presented and compared to the simulation results. A theory for the stationary state of the system is also discussed. Finishing this part, we study an extension of the Derrida-Higgs model aimed at understanding the emergence of the barcode property of the mitochondrial DNA [119], based on the work of Princepe and de Aguiar [120], in which we carefully analyze its sympatric case, which is not discussed in their original work.

5.1.1 The Model

We consider a population of N individuals, each one with its own genome, described by a binary sequence of size B. The genome of an individual α at time t is represented by the string, $\mathbf{S}_t^\alpha = \left(s_{1,t}^\alpha, \ldots, s_{B,t}^\alpha\right)$ where the *alleles* s_i are $s_{i,t}^\alpha = \pm 1$ [94].

Now, it is possible to compare how close two individuals α and β are from each other by simply counting the number of alleles they have in common (or not). We define the genetic distance $d^{\alpha\beta}$ between α and β as the *Hamming distance* between the sequences $\mathbf{S}_t^\alpha$ and $\mathbf{S}_t^\beta$

$$d^{\alpha\beta} = \frac{1}{2}\sum_{i=1}^{B} |s_i^\alpha - s_i^\beta|, \tag{5.1.1}$$

(in which we are not showing the time index), i.e., the genetic distance counts how many distinct alleles exist between two individuals. If $d^{\alpha\beta} = 0$, the individuals are identical.

Another measure of proximity one can define is the genetic similarity,

$$q^{\alpha\beta} = \frac{1}{B}\sum_{i=1}^{B} s_i^\alpha s_i^\beta, \tag{5.1.2}$$

with $-1 \leq q^{\alpha\beta} \leq +1$ and $q^{\alpha\beta} = 1$ for identical individuals. It is possible to prove that these measures are equivalent,

$$\begin{aligned} d^{\alpha\beta} =& \frac{1}{2}\sum_{i=1}^{B} |s_i^\alpha - s_i^\beta| \\ =& \frac{1}{4}\sum_{i=1}^{B} \left(s_i^\alpha - s_i^\beta\right)^2 \\ =& \frac{1}{4}\sum_{i=1}^{B} \left((s_i^\alpha)^2 - 2s_i^\alpha s_i^\beta + (s_i^\beta)^2\right) \\ =& \frac{1}{4}\left(2B - 2Bq^{\alpha\beta}\right) = \frac{B}{2}\left(1 - q^{\alpha\beta}\right), \end{aligned}$$

therefore existing a biunivocal relation between $q^{\alpha\beta}$ and $d^{\alpha\beta}$

$$q^{\alpha\beta} = 1 - \frac{2}{B}d^{\alpha\beta}. \tag{5.1.3}$$

The Derrida-Higgs dynamics starts with a population of N identical hermaphrodite (no sexual differences) individuals. At time t, two individuals α and β are randomly chosen to be the parents of an individual γ from the next generation. The reproduction is sexual and then the genome $\mathbf{S}_{t+1}^\gamma$ is a combination of the parents' genomes. Each allele $s_{i,t+1}^\gamma$ from the offspring γ has probability $1/2$ of being equal to $s_{i,t}^\alpha$ of α and probability $1/2$ of being equal to $s_{i,t}^\beta$ of β. Nonetheless, every allele of γ can mutate with a given mutation probability r.

This process is repeated N times, in such a way that the generation at time $t+1$ is completely generated from the population at time t [94].

The probability r is calculated by considering a mutation rate μ, which defines the

master equations:[1]

$$\frac{d\rho_{(+)}}{dt} = \mu\left(\rho_{(-)} - \rho_{(+)}\right),$$
$$\frac{d\rho_{(-)}}{dt} = \mu\left(\rho_{(+)} - \rho_{(-)}\right),$$

with $\rho_{(\pm)}(t)$ being the probability that at time t a given allele has value ± 1. Its solution can be found with $\chi = \rho_{(+)} - \rho_{(-)}$ and $\rho_{(+)} + \rho_{(-)} = 1$,

$$\frac{d\chi}{dt} = -2\mu\chi \Rightarrow \chi(t) = \chi(t_0)e^{-2\mu(t-t_0)}. \tag{5.1.4}$$

As there are no generation overlap, every generation is an initial time t_0 in respect to the next generation, at time $t_0 + 1$. At t_0, every allele has a given value, in such a way that the probability distribution is a Kronecker delta. Let the allele be σ ($\sigma = \pm 1$), thus $\rho_{(\sigma)}(t_0) = 1$ and $\rho_{(-\sigma)}(t_0) = 0$. We aim to find the probability $r \equiv \rho_{(-\sigma)}(t_0 + 1)$. From the previous equations,

$$\rho_{(-\sigma)}(t_0 + 1) = \rho_{(+\sigma)}(t_0 + 1) - e^{-2\mu},$$
$$\rho_{(+\sigma)}(t_0 + 1) = 1 - \rho_{(-\sigma)}(t_0 + 1),$$

and then

$$r \equiv \rho_{(-\sigma)}(t_0 + 1) = \frac{1}{2}\left(1 - e^{-2\mu}\right), \tag{5.1.5}$$

and for $\mu \ll 1$, $r \approx \mu$.

So far, each individual has been described as a point in the genetic space $\{-1, 1\}^B$. A cloud of points thus describes a population. Due to mutation, this cloud spreads from one generation to another. On the other hand, genetic drift (due to the finite population size) decreases the genetic variability and controls the broadness of this cloud.

However, the absence of any restriction to reproduction is very non-realistic. Thus, we shall consider the inclusion of an *assortative parameter*, so that in order to reproduce, two individuals must have a minimum genetic similarity q_{min}. This restriction means that points very far from each other in the genetic space, according to their Hamming distance, cannot be "combined" to generate an offspring, i.e., they should be sufficiently close.

In this case, reproduction works as follows: a focal individual α from generation t is randomly chosen. Then, its partner β is randomly chosen from the set of individuals γ such that $q^{\alpha\gamma} \geq q_{min}$ (and therefore, if there is at least one compatible individual, α will succeed in reproducing). Fig. 5.1 shows the reproduction mechanism. An important observation is that this way of choosing the mating pair, i.e., by considering first a focal individual and then its partner, is different from choosing the whole pair at once.

In a nutshell: the model is a population dynamics of haploid and hermaphrodite individuals, evolving under sexual and assortative reproduction and subjected to mutation

[1] A *master equation* describes the rate of change of the probability of a given state, but when the time flow is continuous and the state space is discrete. [43]

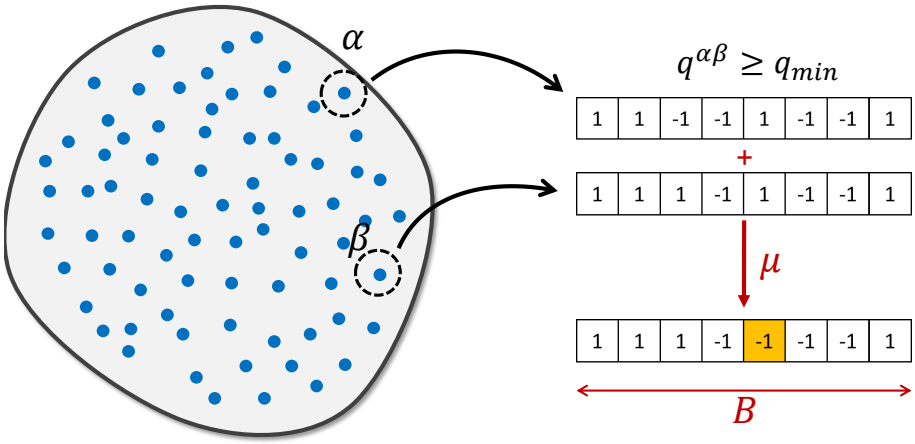


Figure 5.1: **The Derrida-Higgs model.** The figure shows the mechanism of reproduction in the model. At a time t, there is a population of N individuals (blue dots). A focal individual α is chosen and then its mating pair is chosen such that $q^{\alpha\beta} \geq q_{min}$. Their genomes (of size B) are then combined to generate the genome of their offspring, which can also have mutations with a rate μ. For instance, in the figure, the yellow allele has mutated, changing from $+1$ to -1.
Source: Figure produced by the author.

and random genetic drift, with no generational overlap.

The following section presents the first results obtained in this model.

5.1.2 The Derrida-Higgs theory

Once the reproduction threshold is defined considering the similarity between individuals, it is natural to study the distribution of similarities within the population. In the original 1991 paper [94], Derrida and Higgs calculated the similarity between two individuals α and β at generation $t+1$

$$q^{\alpha\beta} = \frac{1}{B}\sum_{i=1}^{B} s_i^{\alpha} s_i^{\beta}$$

in terms of their parents at generation t. Let p_1 and p_2 be the parents of α and p_1' and p_2' the parents of β. If α gets the allele i from p_1, then its mean value is $s_i^{\alpha} = (1-r)s_i^{p_1} - rs_i^{p_1}$. But as there is also a chance of getting it from p_2, with probability $1/2$, then

$$s_i^{\alpha} = \frac{1}{2}(1-r)\left(s_i^{p_1} + s_i^{p_2}\right) - \frac{1}{2}r\left(s_i^{p_1} + s_i^{p_2}\right) = \frac{e^{-2\mu}}{2}\left(s_i^{p_1} + s_i^{p_2}\right)$$

The same calculation can be performed for β, and thus the similarity is given by

$$\begin{aligned} q_{t+1}^{\alpha\beta} =& \frac{e^{-4\mu}}{4B}\sum_{i=1}^{B}(s_i^{p_1}+s_i^{p_2})(s_i^{p_1'}+s_i^{p_2'}) \\ =& \frac{e^{-4\mu}}{4}\left(q_t^{p_1p_1'}+q_t^{p_1p_2'}+q_t^{p_2p_1'}+q_t^{p_2p_2'}\right), \end{aligned} \tag{5.1.6}$$

which is exact if $B \to \infty$.

This expression defines an algorithm for calculating the evolution of the similarity distribution. Starting with a matrix $q_t^{\alpha\beta}$, we draw N pairs p_1 and p_2, following the assortativity rule and by first choosing the focal p_1 and then its mate p_2. With the N chosen pairs, and the equation (5.1.6), the matrix $q_{t+1}^{\alpha\beta}$ can be calculated.

To get more information from the previous equation, we calculate the average similarity over the *population*

$$\langle q_{t+1}^{\alpha\beta}\rangle_P = \frac{1}{N(N-1)}\sum_{\alpha\neq\beta} q_{t+1}^{\alpha\beta}, \tag{5.1.7}$$

since there are $N(N-1)$ pairs of individuals with $\alpha \neq \beta$. This sum can be written in terms of the pairs of parents (p_1, p_2) and (p_1', p_2')

$$\sum_{\alpha\neq\beta} q_{t+1}^{\alpha\beta} = \sum_{(p_1,p_2)\neq(p_1',p_2')} \frac{e^{-4\mu}}{4}\left(q_t^{p_1p_1'}+q_t^{p_1p_2'}+q_t^{p_2p_1'}+q_t^{p_2p_2'}\right).$$

Now, each term within parenthesis can be shown to be the same under indexes permutation, thus

$$\sum_{\alpha\neq\beta} q_{t+1}^{\alpha\beta} = e^{-4\mu}\sum_{(p_1,p_2)\neq(p_1',p_2')} q_t^{p_1p_1'}.$$

Whenever p_1' equals p_1, the similarity equals 1. Performing first the sum on (p_1', p_2'), p_1 is given, and thus there is a chance $1/N$ of $p_1' = p_1$, and since there are $N-1$ pairs (p_1', p_2') such that $(p_1', p_2') \neq (p_1, p_2)$, there are (on average) $(N-1)/N$ cases in which $p_1 = p_1'$ and $(N-1) - (N-1)/N$ cases in which they are different. In these cases, we can consider $q_t^{p_1p_1'} = \langle q_t^{\alpha\beta}\rangle_P$. This way,

$$\langle q_{t+1}^{\alpha\beta}\rangle_P = \frac{e^{-4\mu}}{N(N-1)}\sum_{(p_1,p_2)}\left(\frac{(N-1)}{N}+\left((N-1)-\frac{(N-1)}{N}\right)\langle q_t^{\alpha\beta}\rangle_P\right),$$

and since there are N pairs (p_1, p_2),

$$\langle q_{t+1}^{\alpha\beta}\rangle_P = e^{-4\mu}\left[\frac{1}{N}+\left(1-\frac{1}{N}\right)\langle q_t^{\alpha\beta}\rangle_P\right], \tag{5.1.8}$$

which is a recurrence equation for the average similarity in the case of infinite genome.

Imposing $\langle q_{t+1}^{\alpha\beta}\rangle_P = \langle q_t^{\alpha\beta}\rangle_P = q_{eq}$ in equation (5.1.8), we can find an equilibrium solution,

$$q_{eq} = \frac{1}{Ne^{4\mu}-(N-1)} \approx \frac{1}{1+4\mu N}, \tag{5.1.9}$$

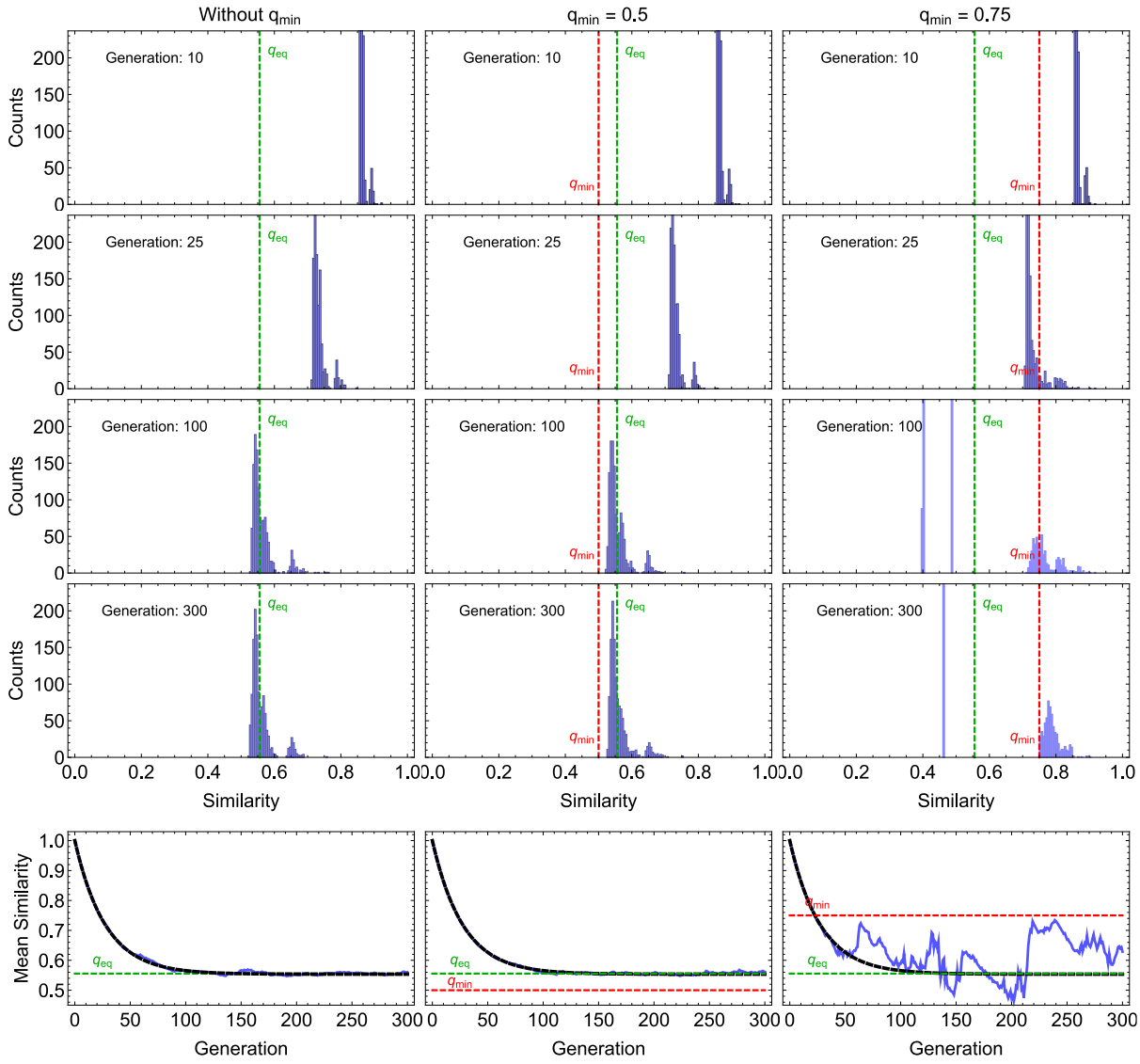


Figure 5.2: **The Derrida-Higgs model with $B \to \infty$.** The first column shows the evolution of the distribution of similarities of a population subjected to the Derrida-Higgs model without mating restrictions (in the absence of q_{min}) across the generations. The bottom panel presents the mean similarity evolution (in blue) showing that it reaches an equilibrium value. The middle column shows the model with $q_{min} < q_{eq}$ and the last column the model with $q_{min} > q_{eq}$, in which we observe that the distribution of similarities no longer achieves stationarity. In all panels, the black dashed curve is the theoretical evolution of the mean similarity without mating restrictions, the red line shows the value of q_{min} and the green line the value of q_{eq}. The three simulations (each column) in the figure considered $N = 50$ and $\mu = 0.004$.
Source: Figure produced by the author.

in which the approximation holds for $\mu \ll 1$. Its stability is easily shown by considering any perturbation $\langle q_t^{\alpha\beta} \rangle_P = q_{eq} + \delta q_t$ on the equilibrium value,

$$\langle q_{t+1}^{\alpha\beta} \rangle_P = e^{-4\mu} \left[\frac{1}{N} + \left(1 - \frac{1}{N} \right) (q_{eq} + \delta q_t) \right] = q_{eq} + \delta q_{t+1},$$

with

$$\delta q_{t+1} = e^{-4\mu}\left(1-\frac{1}{N}\right)\delta q_t \rightarrow \delta q_{t+k} = e^{-4\mu k}\left(1-\frac{1}{N}\right)^k \delta q_t, \tag{5.1.10}$$

and thus any perturbation on the equilibrium value exponentially decreases over the generations. Of course, equation (5.1.8) can then be solved to

$$\langle q_t^{\alpha\beta}\rangle_P = q_{eq} + (q_0 - q_{eq})\left[\left(1-\frac{1}{N}\right)e^{-4\mu}\right]^t. \tag{5.1.11}$$

Therefore, the similarity distribution evolves along the generations toward the equilibrium, as can be seen in the first column of Figure 5.2. In this figure, an identical population evolves according to the Derrida-Higgs model without restrictions to mating (absence of q_{min}) and the distribution of similarities is presented as a histogram. It is possible to see that this histogram gets broader over the generations and reaches a stationary state around q_{eq}. The bottom panel shows the evolution of the mean similarity (in blue) which follows Eq.(5.1.11) (in black), with $q_0 = 1$.

However, the recurrence (5.1.8) does not consider the reproduction threshold q_{min}. Due to the infinite genome size, we may consider the similarity distribution to be very narrow. Suppose now that this distribution is moving towards the equilibrium value and that $q_{min} < q_{eq}$. Thus, when the system reaches its equilibrium value, it remains there without being affected by the threshold, and the system does equilibrate, as shown in the middle column of Fig.5.2. On the other hand, if $q_{min} > q_{eq}$, then the distribution reaches the threshold before finding its natural equilibrium q_{eq}, and suddenly many individuals are not allowed to mate. In this case, the system should display a new behavior, as it can be seen in the last column of Fig.5.2.

The multiple peaks in the last column of Fig.5.2 can be well explained with the aid of network theory and they are indeed characteristic of the formation of *species*. Let us now introduce the definition of species in the model and how it relates to network theory.

5.1.3 The definition of a species

Species are defined in the model as a group of individuals which, according to their similarity, are able to reproduce. Indeed, if two individuals can mate, they belong to the same species, and if two individuals do not belong to the same species, then they cannot mate. However, two individuals that cannot mate also belong to the same species if there is a gene flow between them, e.g., if there is a third individual that can mate with both. This way, we define species according to the existence of gene flow between individuals, even if it is not direct. Let us define it now in mathematical terms.

Let the set $\mathcal{N}_t = \{1,\ldots,N\}$ be the population at time t. We define a path Γ_t in $\mathcal{N}_t$ as the subset $\Gamma_t = \left\{\alpha_1,\ldots,\alpha_n | \alpha_i \in \mathcal{N}_t; q_t^{\alpha_j\alpha_{j+1}} \geq q_{min}, \forall j \in [1,n-1]\right\}$. A species is an application $\mathcal{E}: \mathcal{N}_t \rightarrow \mathcal{I} \subset \mathbb{N}$, which has the following properties:

1. For any $\alpha \neq \beta \in \mathcal{N}_t$ t, if $q^{\alpha\beta} \geq q_{min}$ then $\mathcal{E}(\alpha) = \mathcal{E}(\beta)$;
2. For $\alpha,\beta,\gamma \in \mathcal{N}_t$ distinct, if $\mathcal{E}(\alpha) = \mathcal{E}(\beta)$ and $\mathcal{E}(\beta) = \mathcal{E}(\gamma)$, then $\mathcal{E}(\alpha) = \mathcal{E}(\gamma)$;
3. If there is no path Γ_t such that $\alpha \neq \beta \in \mathcal{N}_t$ belong both to it, then $\mathcal{E}(\alpha) \neq \mathcal{E}(\beta)$.

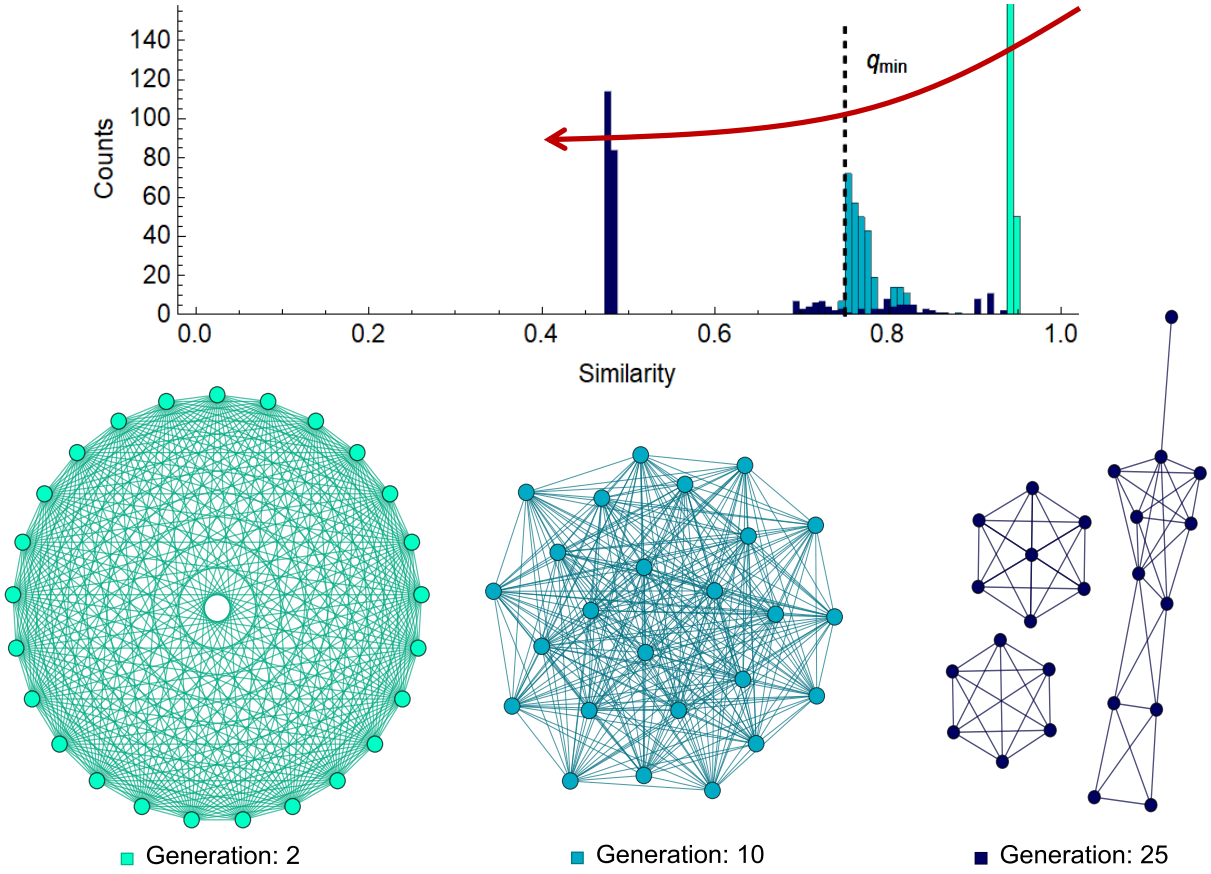


Figure 5.3: **The Derrida-Higgs model and network theory.** The upper plot shows the evolution of the similarity distribution, which evolves towards smaller values (the red arrow shows the evolution direction). Different blue shades show different generations, (time passing from the light to the darkest shade). The corresponding networks are shown below, starting with a complete network and reaching a state with many components. In the figure, the simulation parameters are $N = 25$, $\mu = 0.008$, $q_{min} = 0.75$ and $B \to \infty$. Source: Figure produced by the author.

These properties, although written in a mathematical way, only express what was said before. Property 1. means that if two individuals can reproduce, then they belong to the same species. Property 2. is a transitivity property, i.e., if two individuals cannot reproduce, but there is a third one that can reproduce with both, then they all belong to the same species, resulting in a definition of a species concerning the existence of gene flow. As a corollary of 1. and 2., all elements of a path Γ_t are of the same species. Finally, property 3. defines different species: if there is no path connecting two individuals, then they belong to different species.

Finally, the system is said to have passed through a speciation process if it is possible to find $\alpha \neq \beta \in \mathcal{N}_t$ such that $\mathcal{E}(\alpha) \neq \mathcal{E}(\beta)$.

5.1.4 The network description

The Derrida-Higgs model is easily visualized with the help of network theory. Let the individuals be the vertices of an undirected network. Two vertices are connected if and

only if the similarity between them is greater than the minimum value q_{min}, i.e., if they can mate. Thus, if we start the process with a clonal population, at the beginning there is a complete network, and at every time step a new network is constructed, in which some links may have been erased.

If there is a path connecting two vertices, then they belong to the same species, otherwise they do not. Indeed, species are defined as connected components of this network, and different components define different species.

Figure 5.3 shows the Derrida-Higgs model evolving as networks. The system starts with a clonal population. Then the similarity distribution starts to get broader while moving towards the equilibrium, which is smaller than q_{min}. When it reaches q_{min}, many connections are erased from the network, and after that, the network breaks up into different components. Each component is identified as a different species.

5.2 The finite genome problem

The appearance or not of more than one species defines if a given set of parameters leads to speciation or not, naturally raising the question for which parameters can one observe speciation in the system? Let the number of species at a time t be given by $\mathcal{S}_t$ and let it be a function of the parameters of the model $\mathcal{S}_t = \mathcal{S}_t(B, \mu, q_{min})$. The family $\{\mathcal{S}_t; t = 0, 1, \ldots\}$ defines a stochastic process, and due to the simulations, we may consider it reaches an stationary state with

$$\langle \mathcal{S}_{t\to\infty}(B, N, \mu, q_{min})\rangle = \mathcal{S}(B, N, \mu, q_{min}) \tag{5.2.1}$$

where $\langle\cdot\rangle$ is the ensemble average (we are also going to assume ergodicity for sufficiently long time scales). We then pose the question as **for which set of values B, N, μ and q_{min}, we can observe $\mathcal{S}(B, N, \mu, q_{min}) > 1$?**

Derrida and Higgs conjectured that when the genome size is infinite, speciation occurs if and only if [94],

$$q_{min} > q_{eq}. \tag{5.2.2}$$

On the other hand, in 2017 [118], de Aguiar showed that when the genome is finite, this relation is not satisfied: the similarity distribution reaches an stationary state approximately centered around q_{min} and the system does not undergo speciation unless the genome is very large [118]. However, *how large is large* is not known. We then parameterize the low-high diversity transition as a function of the genome size, defining the critical genome size B_c as

$$B \geq B_c = B_c(N, \mu, q_{min}) \Leftrightarrow S(B, N, \mu, q_{min}) > 1 \tag{5.2.3}$$

i.e., B_c is the smallest genome size that allows species formation when the other parameters N, μ and q_{min} are given.

To find an expression for $B_c(N, \mu, q_{min})$ is our goal.

5.3 Analytical Theory

As we have seen, the similarity distribution has a complex structure and its behavior on time changes when there is or not speciation. To understand its evolution is a pivotal point if one wants to describe the species formation in this model. Therefore, our next step is to introduce a complete formalism to deal with this distribution and to calculate the evolution of its moments.

We start by first constructing the probability distribution for the similarity between two individuals from generation $t+1$, given the similarities in the previous generation. From this distribution, we may be able to calculate the evolution of the mean similarity and its variance. These results are remarkably new. On the other hand, the underlying network structure turns out to be very complicated to treat, hindering a way to find an analytical description of the low-high diversity transition.

Nevertheless, when the system is treated in the absence of a similarity threshold q_{min}, we can find exact evolution equations for the first and second moments and these last results are essential to the *heuristic solution* we are going to introduce in the next chapter.

5.3.1 A Brief Summary

We are just about to start a hard and long mathematical endeavor to find the results we promised.o make it easier for you, reader, here is a summary of the results of this section, so if you would like to, you can skip the following lengthy calculations and go directly to the next section.

Notation

Given a population of N individuals, for every $\alpha \neq \beta$, if $q_t^{\alpha\beta} \geq q_{min}$, $A_{\alpha\beta} = 1$ and 0 otherwise; $N_\alpha = \sum_\beta A_{\alpha\beta}$, i.e., the matrix $\mathbb{A}$ with elements $(\mathbb{A})_{\alpha\beta} = A_{\alpha\beta}$ is the adjacency matrix of the underlying network and N_α is the degree of the vertice α. The individuals α, β, γ and δ are different individuals of generation $t+1$, whose parents are respectively (p_1, p_2), (p_1', p_2'), (p_1'', p_2'') and (p_1''', p_2'''). The genome size is B and the mutation rate μ. Also, the total number of pairs in the population is given by $\overline{N} = N(N-1)/2$

Results

1. **Similarity Distribution**. The probability distribution of a similarity $q_{t+1}^{\alpha\beta}$ is given by

$$\mathcal{P}(q_{t+1}^{\alpha\beta}) = \frac{1}{N^2} \sum_{\mathbf{S}^\alpha, \mathbf{S}^\beta} \sum_{p_1, p_2} \sum_{p_1', p_2'} \delta\left(q_{t+1}^{\alpha\beta}, \frac{\mathbf{S}^\alpha \cdot \mathbf{S}^\beta}{B}\right) \frac{A_{p_1 p_2} A_{p_1' p_2'}}{N_{p_1} N_{p_1'}} \times$$
$$\times \prod_{i=1}^{B} \left[\frac{1}{2} + \frac{s_{i,t+1}^{\alpha} e^{-2\mu}}{4}\left(s_{i,t}^{p_1} + s_{i,t}^{p_2}\right)\right] \left[\frac{1}{2} + \frac{s_{i,t+1}^{\beta} e^{-2\mu}}{4}\left(s_{i,t}^{p_1'} + s_{i,t}^{p_2'}\right)\right] \quad (5.3.1)$$

2. **Expected Similarity**. Given the similarity values at a time t, the expected similarity at the next generation is given by

$$\mathbb{E}(q_{t+1}^{\alpha\beta}) = \frac{e^{-4\mu}}{4N^2} \sum_{p_1,p_2} \sum_{p_1',p_2'} \frac{A_{p_1p_2}A_{p_1'p_2'}}{N_{p_1}N_{p_1'}} \left(q_t^{p_1p_1'} + q_t^{p_1p_2'} + q_t^{p_2p_1'} + q_t^{p_2p_2'} \right), \tag{5.3.2}$$

and when there is no restriction to mating

$$\langle q_{t+1}^{\alpha\beta} \rangle = e^{-4\mu} \left[\frac{1}{N} + \left(1 - \frac{1}{N} \right) \langle q_t^{\alpha\beta} \rangle \right].$$

3. **The second moment of the similarity distribution.** The expected value of the second moment of the similarity distribution $\mathcal{P}(q_{t+1}^{\alpha\beta})$ is given by

$$\begin{aligned} \mathbb{E}((q_{t+1}^{\alpha\beta})^2) = \frac{1}{N^2} \sum_{p_1,p_2} \sum_{p_1',p_2'} \frac{A_{p_1p_2}A_{p_1'p_2'}}{N_{p_1}N_{p_1'}} \Bigg[\frac{1}{B} + \frac{e^{-8\mu}}{16} \left[(q_t^{p_1p_1'} + q_t^{p_1p_2'} + q_t^{p_2p_1'} + q_t^{p_2p_2'}) \right]^2 \\ - \frac{e^{-8\mu}}{4B} \left(1 + q_t^{p_1p_2} + q_t^{p_1'p_2'} + q_t^{p_1p_2p_1'p_2'} \right) \Bigg], \end{aligned} \tag{5.3.3}$$

where the *second order overlap* is defined by

$$q^{\alpha\beta\gamma\delta} = \frac{1}{B} \sum_{i=1}^{B} s_i^\alpha s_i^\beta s_i^\gamma s_i^\delta. \tag{5.3.4}$$

4. **The second order overlap.** Once we have defined the second order overlap, we can calculate its expected value,

$$\begin{aligned} \mathbb{E}(q_{t+1}^{\alpha\beta\gamma\delta}) = \frac{e^{-8\mu}}{16N^4} \sum_{p_1,p_2} \sum_{p_1',p_2'} \sum_{p_1'',p_2''} \sum_{p_1''',p_2'''} \frac{A_{p_1p_2}A_{p_1'p_2'}A_{p_1''p_2''}A_{p_1'''p_2'''}}{N_{p_1}N_{p_1'}N_{p_1''}N_{p_1'''}} \\ \times \Big(q_t^{p_1p_1'p_1''p_1'''} + q_t^{p_1p_1'p_1''p_2'''} + q_t^{p_1p_1'p_2''p_1'''} + q_t^{p_1p_1'p_2''p_2'''} + q_t^{p_1p_2'p_1''p_1'''} + q_t^{p_1p_2'p_1''p_2'''} \\ + q_t^{p_1p_2'p_2''p_1'''} + q_t^{p_1p_2'p_2''p_2'''} + q_t^{p_2p_1'p_1''p_1'''} + q_t^{p_2p_1'p_1''p_2'''} + q_t^{p_2p_1'p_2''p_1'''} \\ + q_t^{p_2p_1'p_2''p_2'''} + q_t^{p_2p_2'p_1''p_1'''} + q_t^{p_2p_2'p_1''p_2'''} + q_t^{p_2p_2'p_2''p_1'''} + q_t^{p_2p_2'p_2''p_2'''} \Big), \end{aligned} \tag{5.3.5}$$

and in the absence of restrictions to mating,

$$\langle q_{t+1}^{\alpha\beta\gamma\delta} \rangle = \frac{e^{-8\mu}}{N^3} \left[(3N-2) + (N-1)(6N-8)\langle q_t^{\alpha\beta} \rangle + (N-1)(N-2)(N-3)\langle q_t^{\alpha\beta\gamma\delta} \rangle \right]. \tag{5.3.6}$$

5. **The variance evolution.** If there are no mating restrictions, the variance of

the similarity distribution evolves according to

$$\begin{aligned}
&\mathrm{Var}(q_{t+1}^{\alpha\beta}) \\
&= \frac{1}{B} - \frac{e^{-8\mu}}{4B}\left[\left(1+\frac{2}{N(N-1)}\right) + \left(2+\frac{4(N-2)}{N(N-1)}\right)\langle q_t^{\alpha\beta}\rangle + \frac{(N-2)(N-3)}{N(N-1)}\langle q_t^{\alpha\beta\gamma\delta}\rangle\right] \\
&+ \frac{e^{-8\mu}(N-2)^2}{4N^2(N-1)}\left[\left(1-\langle q_t^{\alpha\beta}\rangle\right)^2 + \left(N+2-\frac{2}{N}\right)\left(1-\frac{1}{\overline{\overline{N}}}\right)\mathrm{Var}(q_t^{\alpha\beta})\right. \\
&\left.+ 2\left(N-6+\frac{4}{N}\right)\left(1-\frac{1}{\overline{\overline{N}}}\right)\mathrm{Cov}(t)^{\alpha\beta\gamma}\right],
\end{aligned} \tag{5.3.7}$$

in which we have also defined the covariance

$$\mathrm{Cov}(t)^{\alpha\beta\gamma} = \mathrm{Cov}(q_t^{\alpha\beta}, q_t^{\beta\gamma}). \tag{5.3.8}$$

6. **The covariance evolution with a common individual.** The covariance between two similarities with one individual in common, $\mathrm{Cov}(t)^{\alpha\beta\gamma}$, in the absence of mating restrictions, evolves as

$$\begin{aligned}
\mathrm{Cov}(t+1)^{\alpha\beta\gamma} &= \frac{e^{-4\mu}}{B}\left[\frac{1}{N} + \left(1-\frac{1}{N}\right)\langle q_t^{\alpha\beta}\rangle\right] \\
&- \frac{e^{-8\mu}}{2B}\left[\frac{2}{N^2} + \frac{1}{N} + \left(1+\frac{4}{N}-\frac{8}{N^2}\right)\langle q_t^{\alpha\beta}\rangle + \left(1-\frac{2}{N}\right)\left(1-\frac{3}{N}\right)\langle q_t^{\alpha\beta\gamma\delta}\rangle\right] \\
&+ \frac{e^{-8\mu}(N-2)^2}{2N^3}\left(1-\frac{1}{\overline{\overline{N}}}\right)\left[\mathrm{Var}(q_t^{\alpha\beta}) + (N-4)\mathrm{Cov}(t)^{\alpha\beta\gamma}\right].
\end{aligned} \tag{5.3.9}$$

7. **The covariance with no common individual.** The covariance between similarities that do not share any common individual is zero,

$$\mathrm{Cov}(t)^{\alpha\beta\gamma\delta} = \mathbb{E}(q_{t+1}^{\alpha\beta} q_{t+1}^{\gamma\delta}) - \mathbb{E}(q_{t+1}^{\alpha\beta})\mathbb{E}(q_{t+1}^{\gamma\delta}) = 0. \tag{5.3.10}$$

It is important to emphasize that all these results are new and we compare them to simulations in Fig. 5.4. The evolution of the mean similarity without mating restrictions, although the same as found by Derrida and Higgs in 1991, was now calculated without considering an infinite number of alleles, thus extending the previously known results. Next, we present the mean, variance and covariance without the similarity threshold q_{min} at the important limit $B \to \infty$ and $N \gg 1$, which simplifies the complete solutions.

Results in the limit $B \to \infty$ and $N \gg 1$

1. **Expected Similarity**. In this limit, the result does not change,

$$\langle q_{t+1}^{\alpha\beta} \rangle = e^{-4\mu} \left[\frac{1}{N} + \left(1 - \frac{1}{N} \right) \langle q_t^{\alpha\beta} \rangle \right].$$

2. **The second order overlap.**

$$\langle q_{t+1}^{\alpha\beta\gamma\delta} \rangle = e^{-8\mu} \left[\frac{6}{N} \langle q_t^{\alpha\beta} \rangle + \left(1 - \frac{6}{N} \right) \langle q_t^{\alpha\beta\gamma\delta} \rangle \right]. \tag{5.3.11}$$

3. **The variance evolution.**

$$\mathrm{Var}(q_{t+1}^{\alpha\beta}) = \frac{e^{-8\mu}}{4} \left[\frac{1}{N} \left(1 - \langle q_t^{\alpha\beta} \rangle \right)^2 + \left(1 - \frac{1}{N} \right) \mathrm{Var}(q_t^{\alpha\beta}) + 2 \left(1 - \frac{9}{N} \right) \mathrm{Cov}(t)^{\alpha\beta\gamma} \right]. \tag{5.3.12}$$

4. **The covariance evolution with a common individual.**

$$\mathrm{Cov}(t+1)^{\alpha\beta\gamma} = \frac{e^{-8\mu}}{2} \left[\frac{1}{N} \mathrm{Var}(q_t^{\alpha\beta}) + \left(1 - \frac{8}{N} \right) \mathrm{Cov}(t)^{\alpha\beta\gamma} \right]. \tag{5.3.13}$$

5.3.2 The similarity distribution

Many of our calculations are going to follow the same route, even the calculations in Chapter 8. Thus, we summarized the procedure in the figure 5.5. Consider two individuals α and β from time $t+1$, whose parents are (p_1, p_2) and (p_1', p_2'), respectively. Assuming that α got its allele i from p_1, then the probability of its allele i to be equal to $\sigma = \pm 1$ is

$$\mathcal{P}\left(s_{i,t+1}^{\alpha} = \sigma | p_1 \right) = \frac{1}{2} \left[1 + (2r^c - 1) \sigma s_{i,t}^{p_1} \right], \tag{5.3.14}$$

in which r^c is the probability of not mutating, $r^c = 1 - r = \frac{1}{2}(1 + e^{-2\mu})$. However, if the parent from which the allele came from is not known, since the probability is 1/2 of coming from each one,

$$\begin{aligned} \mathcal{P}\left(s_{i,t+1}^{\alpha} = \sigma | (p_1, p_2) \right) &= \frac{1}{4} \left[1 + (2r^c - 1) \sigma s_{i,t}^{p_1} \right] + \frac{1}{4} \left[1 + (2r^c - 1) \sigma s_{i,t}^{p_2} \right] \\ &= \frac{1}{2} + \frac{\sigma e^{-2\mu}}{4} \left(s_{i,t}^{p_1} + s_{i,t}^{p_2} \right). \end{aligned} \tag{5.3.15}$$

Therefore, given the parents' genomes, the probability that α has genome $\mathbf{S}_{t+1}^{\alpha}$ is

$$\mathcal{P}\left(\mathbf{S}_{t+1}^{\alpha} | (p_1, p_2) \right) = \prod_{i=1}^{B} \left[\frac{1}{2} + \frac{s_{i,t+1}^{\alpha} e^{-2\mu}}{4} \left(s_{i,t}^{p_1} + s_{i,t}^{p_2} \right) \right], \tag{5.3.16}$$

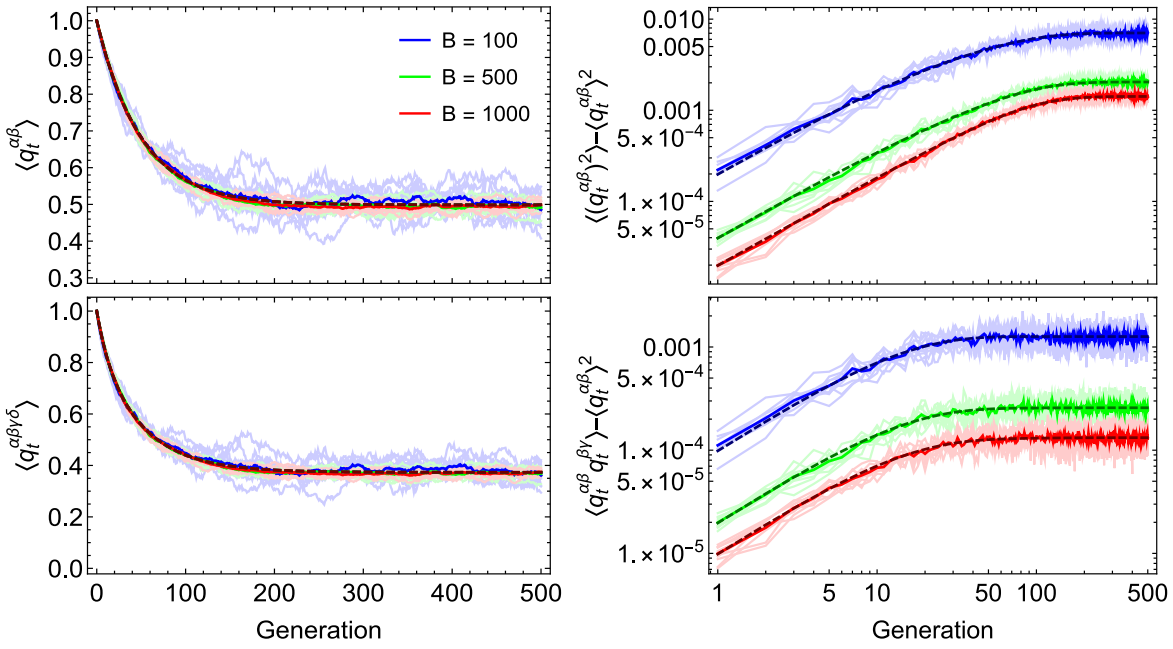


Figure 5.4: **Simulations and theory in the absence of assortative reproduction.** The panels show the evolution of the average similarity $\langle q_t^{\alpha\beta} \rangle$, variance $\mathrm{Var}(q_t^{\alpha\beta}) = \langle (q_t^{\alpha\beta})^2 \rangle - \langle q_t^{\alpha\beta} \rangle^2$, covariance $\mathrm{Cov}(t)^{\alpha\beta} = \langle q_t^{\alpha\beta} q_t^{\beta\gamma} \rangle - \langle q_t^{\alpha\beta} \rangle^2$ and the average second order overlap $\langle q_t^{\alpha\beta\gamma\delta} \rangle$ for different values of genome size B in the absence of q_{min}. The light curves in the background are the evolution of 10 different simulations, the darker continuous curve is the average of these curves and the dashed curve is the theoretical prediction. The average second order overlap was calculated from the simulations by considering a random sample of size N^2 from the set of all second order overlaps, which has size $\sim N^4$. In the figure, the simulation parameters are $N = 100$ and $\mu = 0.0025$.
Source: Figure produced by the author.

because the alleles are independent. An analogous equation can be calculated for β, whose parents are p'_1 and p'_2.

Now, given the genomes of α and β, the probability distribution for the similarity between them is

$$\mathcal{P}(q_{t+1}^{\alpha\beta} | \mathbf{S}_{t+1}^{\alpha}, \mathbf{S}_{t+1}^{\beta}) = \delta \left(q_{t+1}^{\alpha\beta}, \frac{\mathbf{S}_{t+1}^{\alpha} \cdot \mathbf{S}_{t+1}^{\beta}}{B} \right), \tag{5.3.17}$$

because the similarity is uniquely defined by their genomes. From the law of total probability,

$$\mathcal{P}(q^{\alpha\beta}) = \sum_{\mathbf{S}^{\alpha}, \mathbf{S}^{\beta}} \mathcal{P}(q^{\alpha\beta} | \mathbf{S}^{\alpha}, \mathbf{S}^{\beta}) \mathcal{P}(\mathbf{S}^{\alpha}, \mathbf{S}^{\beta}) \tag{5.3.18}$$

in which we are not showing the time index. $\mathcal{P}(\mathbf{S}^{\alpha}, \mathbf{S}^{\beta})$ is the probability of two individuals α and β to have simultaneously the genomes $\mathbf{S}^{\alpha}$ and $\mathbf{S}^{\beta}$, respectively. Let us now calculate this term by noticing that, if one knows their parents, there is a conditional independence between the genomes,

$$\mathcal{P}(\mathbf{S}^{\alpha}, \mathbf{S}^{\beta}) = \sum_{(p_1, p_2)} \sum_{(p'_1, p'_2)} \mathcal{P}(\mathbf{S}^{\alpha}, \mathbf{S}^{\beta} | (p_1, p_2), (p'_1, p'_2)) \mathcal{P}((p_1, p_2), (p'_1, p'_2)). \tag{5.3.19}$$

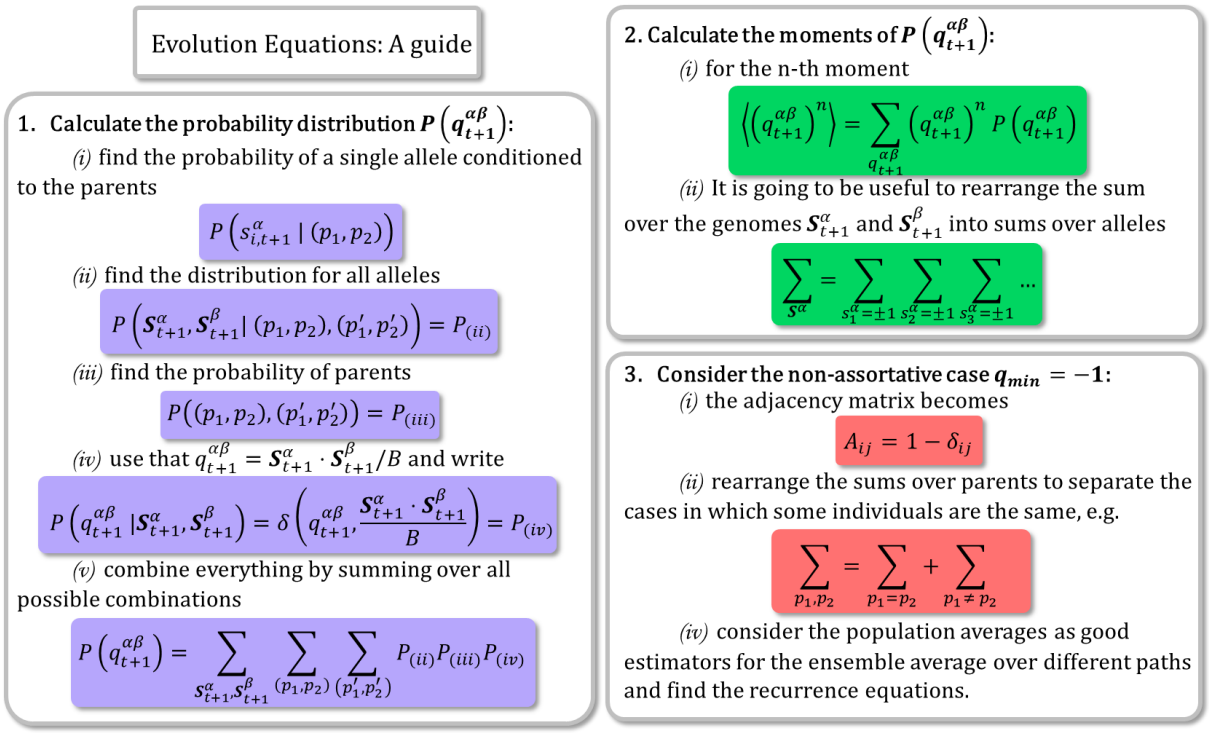


Figure 5.5: **Calculation Scheme.** In order to find analytical results, we follow the calculations as summarized in these panels.
Source: Figure produced by the author.

In this equation, the probability $\mathcal{P}(\mathbf{S}^{\alpha}, \mathbf{S}^{\beta})$ has been conditioned to the parents of α and β. $\mathcal{P}((p_1, p_2), (p_1', p_2'))$ is the probability of a given set of pairs of parents. However, under these conditions, the genomes α and β are independent,

$$\mathcal{P}(\mathbf{S}^{\alpha}, \mathbf{S}^{\beta} | (p_1, p_2), (p_1', p_2')) = \mathcal{P}(\mathbf{S}^{\alpha} | (p_1, p_2)) \mathcal{P}(\mathbf{S}^{\beta} | (p_1', p_2')) \tag{5.3.20}$$

and once the pairs of parents are drawn independently,

$$\mathcal{P}((p_1, p_2), (p_1', p_2')) = \mathcal{P}((p_1, p_2)) \mathcal{P}((p_1', p_2')). \tag{5.3.21}$$

The probabilities $\mathcal{P}(\mathbf{S}^{\alpha} | (p_1, p_2))$ and $\mathcal{P}(\mathbf{S}^{\beta} | (p_1', p_2'))$ have already been calculated (Eq.(5.3.16) and analogous for β). We shall now calculate $\mathcal{P}((p_1, p_2))$ and $\mathcal{P}((p_1', p_2'))$. Let p_1 be the focal individual. Thus, it is drawn at random from the entire population (of size N), with probability $1/N$. Now, p_2 must be drawn among the compatible individuals. Let $\mathbb{N}_{p_1}$ the set of these individuals. Then, for $p_2 \in \mathbb{N}_{p_1}$, since the draw is uniform, it may be chosen with probability $1/N_{p_1}$, where $N_{p_1} = \#(\mathbb{N}_{p_1})$ is the cardinality of the set $\mathbb{N}_{p_1}$, i.e., the number of individuals which are *connected* to p_1. For an individual p_2 not connected to p_1, the probability of being drawn is zero.

This way, we are able to calculate $\mathcal{P}((p_1, p_2))$ as

$$\begin{aligned}\mathcal{P}((p_1,p_2)) =& \mathcal{P}((p_1,p_2)|p_1 \text{ is focal})\mathcal{P}(p_1 \text{ is focal}) + \mathcal{P}((p_1,p_2)|p_2 \text{ is focal})\mathcal{P}(p_2 \text{ is focal}) \\ =& \frac{A_{p_1p_2}}{N_{p_1}}\frac{1}{N} + \frac{A_{p_1p_2}}{N_{p_2}}\frac{1}{N} \\ =& \frac{A_{p_1p_2}}{N}\left(\frac{1}{N_{p_1}} + \frac{1}{N_{p_2}}\right), \end{aligned} \tag{5.3.22}$$

in which $A_{p_1p_2}$ is the element p_1, p_2 of the adjacency matrix of the network defined by the similarities at time t, $\mathbb{A}_t \in \mathbb{M}_{N\times N}$. $A_{p_1p_2}$ is equal to 1 if p_1 and p_2 are connected and 0 otherwise.

Joining all these results together,

$$\begin{aligned}\mathcal{P}(q_{t+1}^{\alpha\beta}) =& \frac{1}{N^2}\sum_{\mathbf{S}^\alpha,\mathbf{S}^\beta}\sum_{(p_1,p_2)}\sum_{(p_1',p_2')}\delta\left(q_{t+1}^{\alpha\beta}, \frac{\mathbf{S}^\alpha\cdot\mathbf{S}^\beta}{B}\right)A_{p_1p_2}A_{p_1'p_2'}\left(\frac{1}{N_{p_1}}+\frac{1}{N_{p_2}}\right)\left(\frac{1}{N_{p_1'}}+\frac{1}{N_{p_2'}}\right)\times \\ &\times\prod_{i=1}^{B}\left[\frac{1}{2}+\frac{s_{i,t+1}^{\alpha}e^{-2\mu}}{4}\left(s_{i,t}^{p_1}+s_{i,t}^{p_2}\right)\right]\left[\frac{1}{2}+\frac{s_{i,t+1}^{\beta}e^{-2\mu}}{4}\left(s_{i,t}^{p_1'}+s_{i,t}^{p_2'}\right)\right]. \end{aligned} \tag{5.3.23}$$

Notice that in this equation, the sums regarding the parents are performed over the pairs of individuals, not over the individuals themselves. We can change it by noticing that the sums are the same if one interchanges p_1 by p_2,

$$\sum_{(p_1,p_2)} \longrightarrow \frac{1}{2}\sum_{p_1,p_2}.$$

The case $p_1 = p_2$ could be a problem, but it is automatically solved once $A_{p_1p_2}$ is zero when $p_1 = p_2$. Obviously, the same works for the sums over (p_1', p_2'). Moreover,

$$\begin{aligned}&\sum_{p_1,p_2}\frac{A_{p_1p_2}}{N_{p_2}}\prod_{i=1}^{B}\left[\frac{1}{2}+\frac{s_{i,t+1}^{\alpha}e^{-2\mu}}{4}\left(s_{i,t}^{p_1}+s_{i,t}^{p_2}\right)\right] \\ =&\sum_{p_2,p_1}\frac{A_{p_2p_1}}{N_{p_2}}\prod_{i=1}^{B}\left[\frac{1}{2}+\frac{s_{i,t+1}^{\alpha}e^{-2\mu}}{4}\left(s_{i,t}^{p_2}+s_{i,t}^{p_1}\right)\right] \\ =&\sum_{p_1,p_2}\frac{A_{p_1p_2}}{N_{p_1}}\prod_{i=1}^{B}\left[\frac{1}{2}+\frac{s_{i,t+1}^{\alpha}e^{-2\mu}}{4}\left(s_{i,t}^{p_1}+s_{i,t}^{p_2}\right)\right]\end{aligned}$$

in which in the second line, we have inverted the order of the sums and used that $A_{p_1p_2} = A_{p_2p_1}$, while in the third line, we have changed the index names, $p_1 \to p_2$ and $p_2 \to p_1$.

Thus,

$$\begin{aligned}&\sum_{p_1,p_2} A_{p_1p_2}\left(\frac{1}{N_{p_1}}+\frac{1}{N_{p_2}}\right)\prod_{i=1}^{B}\left[\frac{1}{2}+\frac{s^{\alpha}_{i,t+1}e^{-2\mu}}{4}\left(s^{p_1}_{i,t}+s^{p_2}_{i,t}\right)\right]\\ =&2\sum_{p_1,p_2}\frac{A_{p_1p_2}}{N_{p_1}}\prod_{i=1}^{B}\left[\frac{1}{2}+\frac{s^{\alpha}_{i,t+1}e^{-2\mu}}{4}\left(s^{p_1}_{i,t}+s^{p_2}_{i,t}\right)\right]\end{aligned}$$

in such a way that $\mathcal{P}(q^{\alpha\beta}_{t+1})$ can be written as

$$\begin{aligned}\mathcal{P}(q^{\alpha\beta}_{t+1}) =&\frac{1}{N^2}\sum_{\mathbf{S}^{\alpha},\mathbf{S}^{\beta}}\sum_{p_1,p_2}\sum_{p'_1,p'_2}\delta\left(q^{\alpha\beta}_{t+1},\frac{\mathbf{S}^{\alpha}\cdot\mathbf{S}^{\beta}}{B}\right)\frac{A_{p_1p_2}A_{p'_1p'_2}}{N_{p_1}N_{p'_1}}\times\\ &\times\prod_{i=1}^{B}\left[\frac{1}{2}+\frac{s^{\alpha}_{i,t+1}e^{-2\mu}}{4}\left(s^{p_1}_{i,t}+s^{p_2}_{i,t}\right)\right]\left[\frac{1}{2}+\frac{s^{\beta}_{i,t+1}e^{-2\mu}}{4}\left(s^{p'_1}_{i,t}+s^{p'_2}_{i,t}\right)\right]\end{aligned}\tag{5.3.24}$$

This equation allows us now to calculate the similarity probability distribution at time $t+1$ given all the genetic information about the system at time t. The time index will not appear in the text sometimes, but we must keep in mind that everything regarding α and β is calculated (or observed) at time $t+1$ while everything that regards their parents p_1, p_2 and p'_1, p'_2, is observed at time t.

5.3.3 The evolution of the Mean Similarity

Once we have calculated the similarity distribution, we can calculate its moments. Let us start by calculating the expected value of $q^{\alpha\beta}_{t+1}$, given by

$$\mathbb{E}(q^{\alpha\beta}_{t+1})=\sum_{q} q\mathcal{P}(q^{\alpha\beta}_{t+1}=q).$$

Because of the δ function, the sum over q is easy to perform and the term $q\delta(q,\mathbf{S}^{\alpha}\cdot\mathbf{S}^{\beta}/B)$ changes to $\mathbf{S}^{\alpha}\cdot\mathbf{S}^{\beta}/B$, decoupling the sums over $\mathbf{S}^{\alpha}$ and $\mathbf{S}^{\beta}$ from the others,

$$\mathbb{E}(q^{\alpha\beta}_{t+1})=\frac{1}{N^2}\frac{1}{B}\sum_{\mathbf{S}^{\alpha},\mathbf{S}^{\beta}}\sum_{p_1,p_2}\sum_{p'_1,p'_2}\sum_{j=1}^{B}s^{\alpha}_j s^{\beta}_j\frac{A_{p_1p_2}A_{p'_1p'_2}}{N_{p_1}N_{p'_1}}\prod_{i=1}^{B}F_{i,t}(\alpha,p_1,p_2)\prod_{i=1}^{B}F_{i,t}(\beta,p'_1,p'_2)$$

where we have defined

$$F_{i,t}(\gamma,a,b)\equiv\frac{1}{2}+\frac{s^{\gamma}_{i,t+1}e^{-2\mu}}{4}\left(s^{a}_{i,t}+s^{b}_{i,t}\right)=\mathcal{P}(s^{\gamma}_{i,t+1}|(a,b))\tag{5.3.25}$$

the probability of the allele $s^{\gamma}_{i,t+1}$ of γ given the alleles of its parents a and b. Rearranging the sums,

$$\mathbb{E}(q^{\alpha\beta}_{t+1}) = \frac{1}{N^2}\frac{1}{B}\sum_{p_1,p_2}\sum_{p'_1,p'_2}\frac{A_{p_1p_2}A_{p'_1p'_2}}{N_{p_1}N_{p'_1}}\sum_j\left(\sum_{\mathbf{S}^\alpha} s_j^\alpha\prod_{i=1}^{B}F_{i,t}(\alpha,p_1,p_2)\right)\left(\sum_{\mathbf{S}^\beta} s_j^\beta\prod_{i=1}^{B}F_{i,t}(\beta,p'_1,p'_2)\right)$$

The terms in parenthesis can be calculated as follows,

$$\begin{aligned}\sum_{\mathbf{S}^\alpha} s_j^\alpha\prod_{i=1}^{B}F_{i,t}(\alpha,p_1,p_2) &= \sum_{s_1^\alpha,\ldots,s_B^\alpha} s_j^\alpha\prod_{i=1}^{B}F_{i,t}(\alpha,p_1,p_2)\\ &= \sum_{s_1^\alpha,\ldots,s_B^\alpha} F_{1,t}\ldots F_{j-1,t}s^\alpha_{j,t}F_{j,t}F_{j+1,t}\ldots F_{B,t}\\ &= \left(\sum_{s_1^\alpha}F_{1,t}\right)\ldots\left(\sum_{s_j^\alpha}s_j^\alpha F_{j,t}\right)\ldots\left(\sum_{s_{B,t}}F_{B,t}\right)\end{aligned}$$

These last terms are easily calculated,

$$\sum_{s_k^\alpha}F_{k,t} = \sum_{s_k^\alpha=\pm1}F_{k,t} = 1$$

and

$$\sum_{s_k^\alpha}s_k^\alpha F_{k,t} = \sum_{s_k^\alpha=\pm1}s_k^\alpha F_{k,t} = \frac{e^{-2\mu}}{2}(s_k^{p_1}+s_k^{p_2}).$$

Thus,

$$\sum_{\mathbf{S}^\alpha} s_j^\alpha\prod_{i=1}^{B}F_{i,t}(\alpha,p_1,p_2) = \frac{e^{-2\mu}}{2}(s_j^{p_1}+s_j^{p_2}). \tag{5.3.26}$$

Now, the expected value is given by

$$\mathbb{E}(q^{\alpha\beta}_{t+1}) = \frac{1}{N^2}\frac{1}{B}\sum_{p_1,p_2}\sum_{p'_1,p'_2}\frac{A_{p_1p_2}A_{p'_1p'_2}}{N_{p_1}N_{p'_1}}\sum_{j=1}^{B}\frac{e^{-4\mu}}{4}(s_j^{p_1}+s_j^{p_2})(s_j^{p'_1}+s_j^{p'_2}),$$

and remembering that the terms relative to the parents are calculated at time t, we end up with

$$\mathbb{E}(q^{\alpha\beta}_{t+1}) = \frac{e^{-4\mu}}{4N^2}\sum_{p_1,p_2}\sum_{p'_1,p'_2}\frac{A_{p_1p_2}A_{p'_1p'_2}}{N_{p_1}N_{p'_1}}\left(q_t^{p_1p'_1}+q_t^{p_1p'_2}+q_t^{p_2p'_1}+q_t^{p_2p'_2}\right). \tag{5.3.27}$$

In Appendix A.1, we show that this equation can be written with an interesting matrix

notation.

5.3.4 The mean without assortative reproduction

In the general case, equation (5.3.27) is not easy to deal with, but in the absence of assortative reproduction, i.e., without q_{min} (or $q_{min} = -1$), this equation is quite simple. Since there are no limitations to the reproduction, any individual can mate with any other and then this case is equivalent to a complete network without loops: $A_{p_1p_2} = A_{p'_1p'_2} = 1$ for any $p_1 \neq p_2$ and $p'_1 \neq p'_2$, and $N_{p_1} = N_{p'_1} = N - 1$. Thus,

$$\mathbb{E}(q_{t+1}^{\alpha\beta}) = \frac{e^{-4\mu}}{4N^2(N-1)^2} \sum_{p_1,p_2} \sum_{p'_1,p'_2} A_{p_1p_2} A_{p'_1p'_2} \left(q_t^{p_1p'_1} + q_t^{p_1p'_2} + q_t^{p_2p'_1} + q_t^{p_2p'_2} \right)$$

And now, by changing the order and indexes names,

$$\begin{aligned} \mathbb{E}(q_{t+1}^{\alpha\beta}) &= \frac{e^{-4\mu}}{N^2(N-1)^2} \sum_{p_1,p_2} \sum_{p'_1,p'_2} A_{p_1p_2} A_{p'_1p'_2} q_t^{p_1p'_1} \\ &= \frac{e^{-4\mu}}{N^2(N-1)^2} \sum_{p_1,p'_1} \left(\sum_{p_2} A_{p_1p_2} \right) \left(\sum_{p'_2} A_{p'_1p'_2} \right) q_t^{p_1p'_1} \\ &= \frac{e^{-4\mu}}{N^2} \sum_{p_1,p'_1} q_t^{p_1p'_1} \end{aligned} \tag{5.3.28}$$

once $\sum_j A_{ij} = N_j = N - 1$. The expression above calculates the expected value of the similarity distribution at time $t + 1$ given the similarity values at time t and can be simplified as

$$\begin{aligned} \mathbb{E}(q_{t+1}^{\alpha\beta}) &= \frac{e^{-4\mu}}{N^2} \sum_{p_1,p'_1} q_t^{p_1p'_1} \\ &= \frac{e^{-4\mu}}{N^2} \left(N + \sum_{p_1 \neq p'_1} q_t^{p_1p'_1} \right). \end{aligned}$$

In this equation, we can recognize the average similarity within the population, given by $\langle q_t^{\alpha\beta} \rangle_P = \frac{1}{N(N-1)} \sum_{p_1 \neq p'_1} q_t^{p_1p'_1}$, then

$$\mathbb{E}(q_{t+1}^{\alpha\beta}) = \frac{e^{-4\mu}}{N^2} \left(N + N(N-1) \langle q_t^{\alpha\beta} \rangle_P \right) = e^{-4\mu} \left[\frac{1}{N} + \left(1 - \frac{1}{N} \right) \langle q_t^{\alpha\beta} \rangle_P \right]. \tag{5.3.29}$$

This equation is interesting because it relates the expected value for the similarity between a pair of individuals in the next step with the observed average similarity of the present population. On the other hand, this is not a recurrence equation. In order to find it, one can take the ensemble average of both sides by averaging over all the possible trajectories

$\mathcal{T}_t$ up to time t, and once the equation is linear, it is possible to write

$$\langle q_{t+1}^{\alpha\beta} \rangle = e^{-4\mu} \left[\frac{1}{N} + \left(1 - \frac{1}{N} \right) \langle q_t^{\alpha\beta} \rangle \right]. \tag{5.3.30}$$

This last equation is identical to Eq.(5.1.8), (although there we wrote it as population averages, instead of ensemble averages) showing that the evolution of the mean similarity in the case $B < \infty$ and in the absence of assortative reproduction is the same as the evolution in the case $B \to \infty$, therefore showing the same equilibrium point, Eq.(5.1.9).

5.3.5 The evolution of the variance

Now, in order to calculate the variance of the distribution,

$$\mathrm{Var}(q_{t+1}^{\alpha\beta}) = \mathbb{E}((q_{t+1}^{\alpha\beta})^2) - \mathbb{E}(q_{t+1}^{\alpha\beta})^2 \tag{5.3.31}$$

we must calculate the second moment

$$\mathbb{E}((q_{t+1}^{\alpha\beta})^2) = \sum_q q^2 \mathcal{P}(q_{t+1}^{\alpha\beta} = q) \tag{5.3.32}$$

to which we have

$$\begin{aligned}
\mathbb{E}((q_{t+1}^{\alpha\beta})^2) &= \frac{1}{N^2} \sum_{\mathbf{S}^\alpha, \mathbf{S}^\beta} \sum_{p_1,p_2} \sum_{p_1',p_2'} \left(\frac{\mathbf{S}^\alpha \cdot \mathbf{S}^\beta}{B} \right)^2 \frac{A_{p_1p_2} A_{p_1'p_2'}}{N_{p_1} N_{p_1'}} \prod_{i=1}^{B} F_{i,t}(\alpha, p_1, p_2) F_{i,t}(\beta, p_1', p_2') \\
&= \frac{1}{N^2 B^2} \sum_{\mathbf{S}^\alpha, \mathbf{S}^\beta} \sum_{p_1,p_2} \sum_{p_1',p_2'} \left(\sum_{j,k} s_j^\alpha s_j^\beta s_k^\alpha s_k^\beta \right) \frac{A_{p_1p_2} A_{p_1'p_2'}}{N_{p_1} N_{p_1'}} \prod_{i=1}^{B} F_{i,t}(\alpha, p_1, p_2) F_{i,t}(\beta, p_1', p_2') \\
&= \frac{1}{N^2 B^2} \sum_{p_1,p_2} \sum_{p_1',p_2'} \frac{A_{p_1p_2} A_{p_1'p_2'}}{N_{p_1} N_{p_1'}} \\
&\times \sum_{\mathbf{S}^\alpha, \mathbf{S}^\beta} \left(\sum_{j=1}^{B} s_j^\alpha s_j^\beta s_j^\alpha s_j^\beta + \sum_{j \neq k} s_j^\alpha s_j^\beta s_k^\alpha s_k^\beta \right) \prod_{i=1}^{B} F_{i,t}(\alpha, p_1, p_2) F_{i,t}(\beta, p_1', p_2') \\
&= \frac{1}{N^2 B^2} \sum_{p_1,p_2} \sum_{p_1',p_2'} \frac{A_{p_1p_2} A_{p_1'p_2'}}{N_{p_1} N_{p_1'}} \\
&\times \sum_{\mathbf{S}^\alpha, \mathbf{S}^\beta} \left[B \prod_{i=1}^{B} F_{i,t}(\alpha) F_{i,t}(\beta) + \sum_{j \neq k} \left(s_j^\alpha s_k^\alpha \prod_{i=1}^{B} F_{i,t}(\alpha) \right) \left(s_j^\beta s_k^\beta \prod_{i=1}^{B} F_{i,t}(\beta) \right) \right]
\end{aligned} \tag{5.3.33}$$

In the expression above, the first term in brackets, after summing on $\mathbf{S}^\alpha$ and $\mathbf{S}^\beta$, equals to B (as we have already calculated). The second term is also easy to calculate if we follow the same procedure for the expected value $\mathbb{E}(q_{t+1}^{\alpha\beta})$, the difference now is that two terms

in the product do not equal to 1, instead of only one,

$$\begin{aligned}
&\sum_{\mathbf{S}^\alpha,\mathbf{S}^\beta}\sum_{j\neq k}\left(s_j^\alpha s_k^\alpha\prod_{i=1}^{B}F_{i,t}(\alpha)\right)\left(s_j^\beta s_k^\beta\prod_{i=1}^{B}F_{i,t}(\beta)\right)\\
&=\sum_{j\neq k}\frac{e^{-8\mu}}{16}(s_j^{p_1}+s_j^{p_2})(s_j^{p_1'}+s_j^{p_2'})(s_k^{p_1}+s_k^{p_2})(s_k^{p_1'}+s_k^{p_2'})\\
&=\frac{e^{-8\mu}}{16}\sum_{j=1}^{B}(s_j^{p_1}+s_j^{p_2})(s_j^{p_1'}+s_j^{p_2'})\sum_{k=1,k\neq j}^{B}(s_k^{p_1}+s_k^{p_2})(s_k^{p_1'}+s_k^{p_2'})\\
&=\frac{e^{-8\mu}}{16}\sum_{j=1}^{B}(s_j^{p_1}+s_j^{p_2})(s_j^{p_1'}+s_j^{p_2'})\left[-(s_j^{p_1}+s_j^{p_2})(s_j^{p_1'}+s_j^{p_2'})+\sum_{k=1}^{B}(s_k^{p_1}+s_k^{p_2})(s_k^{p_1'}+s_k^{p_2'})\right]\\
&=\frac{e^{-8\mu}}{16}\left[B^2(q_t^{p_1p_1'}+q_t^{p_1p_2'}+q_t^{p_2p_1'}+q_t^{p_2p_2'})^2-\sum_{j=1}^{B}(s_j^{p_1}+s_j^{p_2})^2(s_j^{p_1'}+s_j^{p_2'})^2\right]\\
&=\frac{e^{-8\mu}}{16}\left[B(q_t^{p_1p_1'}+q_t^{p_1p_2'}+q_t^{p_2p_1'}+q_t^{p_2p_2'})\right]^2-\frac{e^{-8\mu}}{4}\left[B+Bq_t^{p_1p_2}+Bq_t^{p_1'p_2'}+\sum_{j=1}^{B}s_j^{p_1}s_j^{p_2}s_j^{p_1'}s_j^{p_2'}\right].
\end{aligned}\tag{5.3.34}$$

Defining the *second order overlap* among individuals α, β, γ and δ as

$$q^{\alpha\beta\gamma\delta}=\frac{1}{B}\sum_{i=1}^{B}s_i^\alpha s_i^\beta s_i^\gamma s_i^\delta \tag{5.3.35}$$

we can write,

$$\begin{aligned}
\mathbb{E}((q_{t+1}^{\alpha\beta})^2)=\frac{1}{N^2}\sum_{p_1,p_2}\sum_{p_1',p_2'}\frac{A_{p_1p_2}A_{p_1'p_2'}}{N_{p_1}N_{p_1'}}\Bigg[&\frac{1}{B}+\frac{e^{-8\mu}}{16}\left[(q_t^{p_1p_1'}+q_t^{p_1p_2'}+q_t^{p_2p_1'}+q_t^{p_2p_2'})\right]^2\\
&-\frac{e^{-8\mu}}{4B}\left(1+q_t^{p_1p_2}+q_t^{p_1'p_2'}+q_t^{p_1p_2p_1'p_2'}\right)\Bigg].
\end{aligned}\tag{5.3.36}$$

5.3.6 The variance without assortative reproduction

As for the mean similarity, the equation for the variance is not easy to treat in the general case, but when there is no q_{min}, due to the network structure, it is possible to calculate it. However, it is not simple as the mean and we must treat the sums very carefully,

$$\begin{aligned}
\mathbb{E}((q_{t+1}^{\alpha\beta})^2)=&\frac{1}{B}-\frac{e^{-8\mu}}{4B}\left(1+\frac{2}{N}\sum_{p_1,p_2}\frac{A_{p_1p_2}}{(N-1)}q_t^{p_1p_2}+\frac{1}{N^2}\sum_{p_1,p_2}\sum_{p_1',p_2'}\frac{A_{p_1p_2}A_{p_1'p_2'}}{(N-1)^2}q_t^{p_1p_2p_1'p_2'}\right)\\
&+\frac{e^{-8\mu}}{16N^2}\sum_{p_1,p_2}\sum_{p_1',p_2'}\frac{A_{p_1p_2}A_{p_1'p_2'}}{(N-1)^2}\left((q_t^{p_1p_1'})^2+(q_t^{p_1p_2'})^2+(q_t^{p_2p_1'})^2+(q_t^{p_2p_2'})^2\right.\\
&\left.+2q_t^{p_1p_1'}q_t^{p_1p_2'}+2q_t^{p_1p_1'}q_t^{p_2p_1'}+2q_t^{p_1p_1'}q_t^{p_2p_2'}+2q_t^{p_1p_2'}q_t^{p_2p_1'}+2q_t^{p_1p_2'}q_t^{p_2p_2'}+2q_t^{p_2p_1'}q_t^{p_2p_2'}\right)
\end{aligned}$$

$$= \frac{1}{B} - \frac{e^{-8\mu}}{4B}\left(1 + \frac{2}{N}\sum_{p_1,p_2}\frac{A_{p_1p_2}}{(N-1)}q_t^{p_1p_2} + \frac{1}{N^2}\sum_{p_1,p_2}\sum_{p_1',p_2'}\frac{A_{p_1p_2}A_{p_1'p_2'}}{(N-1)^2}q_t^{p_1p_2p_1'p_2'}\right)$$
$$+ \frac{e^{-8\mu}}{4N^2}\left(\sum_{p_1,p_1'}(q_t^{p_1p_1'})^2 + 2\sum_{p_1}\sum_{p_1',p_2'}\frac{A_{p_1'p_2'}}{(N-1)}q_t^{p_1p_1'}q_t^{p_1p_2'} + \sum_{p_1,p_2}\sum_{p_1',p_2'}\frac{A_{p_1p_2}A_{p_1'p_2'}}{(N-1)^2}q_t^{p_1p_1'}q_t^{p_2p_2'}\right). \tag{5.3.37}$$

The double sums are easy to calculate,

$$\mathbb{E}((q_{t+1}^{\alpha\beta})^2) = \frac{1}{B} - \frac{e^{-8\mu}}{4B}\left(1 + 2\langle q_t^{p_1p_2}\rangle_P + \frac{1}{N^2}\sum_{p_1,p_2}\sum_{p_1',p_2'}\frac{A_{p_1p_2}A_{p_1'p_2'}}{(N-1)^2}q_t^{p_1p_2p_1'p_2'}\right)$$
$$+ \frac{e^{-8\mu}}{4N^2}\left(N + N(N-1)\langle(q_t^{p_1p_1'})^2\rangle_P + 2\sum_{p_1}\sum_{p_1',p_2'}\frac{A_{p_1'p_2'}}{(N-1)}q_t^{p_1p_1'}q_t^{p_1p_2'}\right.$$
$$\left. + \sum_{p_1,p_2}\sum_{p_1',p_2'}\frac{A_{p_1p_2}A_{p_1'p_2'}}{(N-1)^2}q_t^{p_1p_1'}q_t^{p_2p_2'}\right), \tag{5.3.38}$$

where we have used that $\sum_{p_1,p_2} A_{p_1p_2}q_t^{p_1p_2} = N(N-1)\langle q_t^{p_1p_2}\rangle_P$ and $\sum_{p_1,p_1'}(q_t^{p_1p_2})^2 = N + N(N-1)\langle(q_t^{p_1p_2})^2\rangle_P$. However, the sums over 3 and 4 indexes are not so simple. The best way to deal with them is to open the sums in all possible combinations of indexes. The complete expressions for these "opened sums" are presented in Appendix A.2 and the final forms for each sum equal to

$$\sum_{p_1,p_2}\sum_{p_1',p_2'}A_{p_1p_2}A_{p_1'p_2'}q_t^{p_1p_2p_1'p_2'} = 2N(N-1) + 4N(N-1)(N-2)\langle q_t^{p_1p_1'}\rangle_P$$
$$+ N(N-1)(N-2)(N-3)\langle q_t^{p_1p_2p_1'p_2'}\rangle_P \tag{5.3.39}$$

$$\sum_{p_1}\sum_{p_1',p_2'}A_{p_1'p_2'}q_t^{p_1p_1'}q_t^{p_1p_2'} = 2N(N-1)\langle q_t^{p_1p_1'}\rangle_P + N(N-1)(N-2)\langle q_t^{p_1p_1'}q_t^{p_1p_2'}\rangle_P \tag{5.3.40}$$

$$\sum_{p_1,p_2}\sum_{p_1',p_2'}A_{p_1p_2}A_{p_1'p_2'}q_t^{p_1p_1'}q_t^{p_2p_2'} = N(N-1) + N(N-1)\langle(q_t^{p_1p_1'})^2\rangle_P$$
$$+ 2N(N-1)(N-2)\left(\langle q_t^{p_1p_1'}\rangle_P + \langle q_t^{p_1p_1'}q_t^{p_1p_2'}\rangle_P\right)$$
$$+ N(N-1)(N-2)(N-3)\langle q_t^{p_1p_1'}q_t^{p_2p_2'}\rangle_P, \tag{5.3.41}$$

where expressions for the population averages we have identified are shown in Appendix A.3. We emphasize that, in our notation, whenever a similarity (or second order overlap) is between angular brackets (i.e., averaged), different indexes are in fact different. Also, the final result expressed in equation (5.3.39) considers some simple properties of the second order overlap quantity which are going to be treated in the following section.

With these expressions,

$$\begin{aligned}
&\mathbb{E}((q_{t+1}^{\alpha\beta})^2) \\
&= \frac{1}{B} - \frac{e^{-8\mu}}{4B}\left[\left(1+\frac{2}{N(N-1)}\right) + \left(2+\frac{4(N-2)}{N(N-1)}\right)\langle q_t^{p_1p_1'}\rangle_P + \frac{(N-2)(N-3)}{N(N-1)}\langle q_t^{p_1p_2p_1'p_2'}\rangle_P\right] \\
&+ \frac{e^{-8\mu}}{4N}\left[\frac{N}{(N-1)} + \frac{(6N-8)}{(N-1)}\langle q_t^{p_1p_1'}\rangle_P + \left(N-1+\frac{1}{N-1}\right)\langle (q_t^{p_1p_1'})^2\rangle_P\right. \\
&\left.+ \frac{2N(N-2)}{(N-1)}\langle q_t^{p_1p_1'}q_t^{p_1p_2'}\rangle_P + \frac{(N-2)(N-3)}{(N-1)}\langle q_t^{p_1p_1'}q_t^{p_2p_2'}\rangle_P\right].
\end{aligned} \tag{5.3.42}$$

In order to finish the calculation for the variance, we should subtract $\mathbb{E}[q_{t+1}^{\alpha\beta}]^2$. From equation (5.3.28),

$$\mathbb{E}(q_{t+1}^{\alpha\beta})^2 = \left(\frac{e^{-4\mu}}{N^2}\sum_{p_1,p_1'} q_t^{p_1p_1'}\right)^2 = \frac{e^{-8\mu}}{N^4}\sum_{p_1,p_1'}\sum_{p_2,p_2'} q_t^{p_1p_1'}q_t^{p_2p_2'} \tag{5.3.43}$$

and using the expansions in Appendix A.2,

$$\begin{aligned}
\mathbb{E}(q_{t+1}^{\alpha\beta})^2 = \frac{e^{-8\mu}}{N^3}&\left[N + 2N(N-1)\langle q_t^{p_1p_1'}\rangle + 2(N-1)\langle (q_t^{p_1p_1'})^2\rangle\right. \\
&\left.+4(N-1)(N-2)\langle q_t^{p_1p_1'}q_t^{p_1p_2'}\rangle + (N-1)(N-2)(N-3)\langle q_t^{p_1p_1'}q_t^{p_2p_2'}\rangle\right].
\end{aligned} \tag{5.3.44}$$

Hence,

$$\begin{aligned}
&\mathrm{Var}(q_{t+1}^{\alpha\beta}) \\
&= \frac{1}{B} - \frac{e^{-8\mu}}{4B}\left[\left(1+\frac{2}{N(N-1)}\right) + \left(2+\frac{4(N-2)}{N(N-1)}\right)\langle q_t^{p_1p_1'}\rangle_P + \frac{(N-2)(N-3)}{N(N-1)}\langle q_t^{p_1p_2p_1'p_2'}\rangle_P\right] \\
&+ \frac{e^{-8\mu}(N-2)^2}{4N^2(N-1)}\left[1 - 2\langle q_t^{p_1p_1'}\rangle_P + \left(N+2-\frac{2}{N}\right)\langle (q_t^{p_1p_1'})^2\rangle_P\right. \\
&\left.+ 2\left(N-6+\frac{4}{N}\right)\langle q_t^{p_1p_1'}q_t^{p_1p_2'}\rangle_P - (N-3)\left(3-\frac{2}{N}\right)\langle q_t^{p_1p_1'}q_t^{p_2p_2'}\rangle_P\right]
\end{aligned} \tag{5.3.45}$$

Now, let us consider

$$\mathrm{Var}(q_t^{\alpha\beta})_P = \left(\frac{\overline{N}}{\overline{N}-1}\right)\left(\langle (q_t^{\alpha\beta})^2\rangle_P - \langle q_t^{\alpha\beta}\rangle_P^2\right) \Rightarrow \langle (q_t^{\alpha\beta})^2\rangle_P = \left(1-\frac{1}{\overline{N}}\right)\mathrm{Var}(q_t^{\alpha\beta})_P + \langle q_t^{\alpha\beta}\rangle_P^2 \tag{5.3.46}$$

$$\mathrm{Cov}(t)_P^{\alpha\beta\gamma} = \left(\frac{\overline{N}}{\overline{N}-1}\right)\left(\langle q_t^{\alpha\beta}q_t^{\alpha\gamma}\rangle_P - \langle q_t^{\alpha\beta}\rangle_P^2\right) \Rightarrow \langle q_t^{\alpha\beta}q_t^{\alpha\gamma}\rangle_P = \left(1-\frac{1}{\overline{N}}\right)\mathrm{Cov}(t)_P^{\alpha\beta\gamma} + \langle q_t^{\alpha\beta}\rangle_P^2 \tag{5.3.47}$$

$$\mathrm{Cov}(t)_P^{\alpha\beta\gamma\delta} = \left(\frac{\overline{N}}{\overline{N}-1}\right)\left(\langle q_t^{\alpha\beta}q_t^{\gamma\delta}\rangle_P - \langle q_t^{\alpha\beta}\rangle_P^2\right) \Rightarrow \langle q_t^{\alpha\beta}q_t^{\gamma\delta}\rangle_P = \left(1-\frac{1}{\overline{N}}\right)\mathrm{Cov}(t)_P^{\alpha\beta\gamma\delta} + \langle q_t^{\alpha\beta}\rangle_P^2 \tag{5.3.48}$$

with $\overline{N} = N(N-1)/2$ the number of pairs. We have just introduced the covariance

$\mathrm{Cov}(t)_P^{\alpha\beta\gamma}$ between similarities sharing one individual in common and $\mathrm{Cov}(t)_P^{\alpha\beta\gamma\delta}$ between similarities sharing no individual in common, calculated over the populations. Including these definitions in equation (5.3.45),

$$\begin{aligned}
&\mathrm{Var}(q_{t+1}^{\alpha\beta}) \\
&= \frac{1}{B} - \frac{e^{-8\mu}}{4B}\left[\left(1+\frac{2}{N(N-1)}\right) + \left(2+\frac{4(N-2)}{N(N-1)}\right)\langle q_t^{p_1p_1'}\rangle_P + \frac{(N-2)(N-3)}{N(N-1)}\langle q_t^{p_1p_2p_1'p_2'}\rangle_P\right] \\
&+ \frac{e^{-8\mu}(N-2)^2}{4N^2(N-1)}\left[\left(1-\langle q_t^{p_1p_1'}\rangle_P\right)^2 + \left(N+2-\frac{2}{N}\right)\left(1-\frac{1}{\overline{\overline{N}}}\right)\mathrm{Var}(q_t^{\alpha\beta})\right. \\
&\left.+ 2\left(N-6+\frac{4}{N}\right)\left(1-\frac{1}{\overline{\overline{N}}}\right)\mathrm{Cov}(t)^{\alpha\beta\gamma} - (N-3)\left(3-\frac{2}{N}\right)\left(1-\frac{1}{\overline{\overline{N}}}\right)\mathrm{Cov}(t)^{\alpha\beta\gamma\delta}\right]
\end{aligned} \tag{5.3.49}$$

In this expression, considering the population quantities as *good estimators* for the theoretical values, we finally have a recurrence equation for the evolution of the variance of the similarity distribution in the absence of mating restrictions. The recurrence equation then takes the form presented in the beginning of this section (there we have also considered the result $\mathrm{Cov}(t)^{\alpha\beta\gamma\delta} = 0$). Now, we shall calculate the covariances $\mathrm{Cov}(t)^{\alpha\beta\gamma\delta}$ and $\mathrm{Cov}(t)^{\alpha\beta\gamma}$ as also the mean second order overlap $\mathbb{E}(q_t^{\alpha\beta\gamma\delta})$.

5.3.7 The Second Order Overlap

When calculating the variance of the similarity distribution, we defined the second order overlap $q^{\alpha\beta\gamma\delta}$ between the individuals α, β, γ and δ,

$$q_t^{\alpha\beta\gamma\delta} = \frac{1}{B}\sum_{i=1}^{B} s_{i,t}^{\alpha}s_{i,t}^{\beta}s_{i,t}^{\gamma}s_{i,t}^{\delta}. \tag{5.3.50}$$

We aim now to study its properties and evolution.

First, it is easy to see that when two individuals are the same, the second order overlap equals the *first order overlap* (i.e., the similarity) between the remaining two individuals,

$$q^{\alpha\alpha\gamma\delta} = \frac{1}{B}\sum_{i=1}^{B} s_i^{\alpha}s_i^{\alpha}s_i^{\gamma}s_i^{\delta} = \frac{1}{B}\sum_{i=1}^{B} s_i^{\gamma}s_i^{\delta} = q^{\gamma\delta}. \tag{5.3.51}$$

Also, when there are two pairs of common individuals, the second overlap equals 1,

$$q^{\alpha\alpha\gamma\gamma} = \frac{1}{B}\sum_{i=1}^{B} s_i^{\alpha}s_i^{\alpha}s_i^{\gamma}s_i^{\gamma} = \frac{1}{B}\sum_{i=1}^{B} 1 = 1. \tag{5.3.52}$$

These properties tell us that the interesting case to be considered is when the 4 individuals are different. Then, when writing any average of $q_t^{\alpha\beta\gamma\delta}$ we are gonna be referring to this case of interest.

To calculate the expected value $\mathbb{E}(q_{t+1}^{\alpha\beta\gamma\delta})$, we must know the distribution $\mathcal{P}(q_{t+1}^{\alpha\beta\gamma\delta})$, which is easily calculated once we remember that the genomes of different individuals

are independent once we know their parents, as we have done for the mean. Thus, the distribution is completely analogous to the one we have found for $q_{t+1}^{\alpha\beta}$, and is given by

$$\begin{aligned}\mathcal{P}(q_{t+1}^{\alpha\beta\gamma\delta}) = \frac{1}{N^4}\sum_{p_1,p_2}\sum_{p_1',p_2'}\sum_{p_1'',p_2''}\sum_{p_1''',p_2'''}\frac{A_{p_1p_2}A_{p_1'p_2'}A_{p_1''p_2''}A_{p_1'''p_2'''}}{N_{p_1}N_{p_1'}N_{p_1''}N_{p_1'''}}\\ \times\sum_{\mathbf{S}^\alpha,\mathbf{S}^\beta,\mathbf{S}^\gamma,\mathbf{S}^\delta}\delta\left(q_{t+1}^{\alpha\beta\gamma\delta},\frac{1}{B}\sum_{j=1}^{B}s_j^\alpha s_j^\beta s_j^\gamma s_j^\delta\right)\\ \times\prod_{i=1}^{B}F_{i,t}(\alpha,p_1,p_2)F_{i,t}(\beta,p_1',p_2')F_{i,t}(\gamma,p_1'',p_2'')F_{i,t}(\delta,p_1''',p_2''')\end{aligned} \tag{5.3.53}$$

where we have introduced the parents (p_1'', p_2'') and (p_1''', p_2''') of γ and δ, respectively. Once all individuals are different, the expected value is easily obtained as

$$\begin{aligned}&\mathbb{E}(q_{t+1}^{\alpha\beta\gamma\delta})\\ &=\sum_q q\mathcal{P}(q_{t+1}^{\alpha\beta\gamma\delta}=q)\\ &=\frac{e^{-8\mu}}{16N^4}\sum_{p_1,p_2}\sum_{p_1',p_2'}\sum_{p_1'',p_2''}\sum_{p_1''',p_2'''}\frac{A_{p_1p_2}A_{p_1'p_2'}A_{p_1''p_2''}A_{p_1'''p_2'''}}{N_{p_1}N_{p_1'}N_{p_1''}N_{p_1'''}}\\ &\times\frac{1}{B}\sum_{j=1}^{B}(s_j^{p_1}+s_j^{p_2})(s_j^{p_1'}+s_j^{p_2'})(s_j^{p_1''}+s_j^{p_2''})(s_j^{p_1'''}+s_j^{p_2'''})\\ &=\frac{e^{-8\mu}}{16N^4}\sum_{p_1,p_2}\sum_{p_1',p_2'}\sum_{p_1'',p_2''}\sum_{p_1''',p_2'''}\frac{A_{p_1p_2}A_{p_1'p_2'}A_{p_1''p_2''}A_{p_1'''p_2'''}}{N_{p_1}N_{p_1'}N_{p_1''}N_{p_1'''}}\\ &\times\Big(q_t^{p_1p_1'p_1''p_1'''}+q_t^{p_1p_1'p_1''p_2'''}+q_t^{p_1p_1'p_2''p_1'''}+q_t^{p_1p_1'p_2''p_2'''}+q_t^{p_1p_2'p_1''p_1'''}+q_t^{p_1p_2'p_1''p_2'''}+q_t^{p_1p_2'p_2''p_1'''}+q_t^{p_1p_2'p_2''p_2'''}\\ &+q_t^{p_2p_1'p_1''p_1'''}+q_t^{p_2p_1'p_1''p_2'''}+q_t^{p_2p_1'p_2''p_1'''}+q_t^{p_2p_1'p_2''p_2'''}+q_t^{p_2p_2'p_1''p_1'''}+q_t^{p_2p_2'p_1''p_2'''}+q_t^{p_2p_1'p_2''p_2'''}+q_t^{p_2p_2'p_2''p_2'''}\Big)\end{aligned} \tag{5.3.54}$$

which in the absence of assortative reproduction reduces to

$$\mathbb{E}(q_{t+1}^{\alpha\beta\gamma\delta}) = \frac{e^{-8\mu}}{N^4}\sum_{p_1,p_1',p_1'',p_1'''}q_t^{p_1p_1'p_1''p_1'''}. \tag{5.3.55}$$

Again, to deal with the four indexes sum, we use the result in Appendix A.2 and find

$$\mathbb{E}(q_{t+1}^{\alpha\beta\gamma\delta}) = \frac{e^{-8\mu}}{N^3}\left[(3N-2)+(N-1)(6N-8)\langle q_t^{p_1p_1'}\rangle_P+(N-1)(N-2)(N-3)\langle q_t^{p_1p_1'p_1''p_1'''}\rangle_P\right] \tag{5.3.56}$$

$$\approx e^{-8\mu}\left[\frac{6}{N}\langle q_t^{p_1p_1'}\rangle_P+\left(1-\frac{6}{N}\right)\langle q_t^{p_1p_1'p_1''p_1'''}\rangle_P\right] \tag{5.3.57}$$

in which the approximation holds for $N \gg 1$. The factor $6/N$ is the probability that two of the parents p_1, p_1', p_1'' and p_1''' are actually the same individual.

5.3.8 The similarity covariance

The covariance appeared when we were calculating the expected value of $(q_{t+1}^{\alpha\beta})$. And it is, indeed, a very important calculation. When changing from the genetic description of the individuals to the similarity description of the population, we are neglecting individual genome values and considering only a measure of their pairwise distance. Although it is a very natural change of description, once the dynamical constraint is given in terms of the genetic similarity, the probability of genomes of different individuals are independent of each other (as it can be shown with equations (5.3.19), (5.3.20) and (5.3.21)), while genetic similarities between different pairs of individuals may not be.

So let us start by defining the covariance between $q_t^{\alpha\beta}$ and $q_t^{\gamma\delta}$, with $\alpha \neq \beta$ and $\gamma \neq \delta$ is defined as

$$\begin{aligned}\mathrm{Cov}(q_t^{\alpha\beta}, q_t^{\gamma\delta}) &= \mathbb{E}\left[(q_t^{\alpha\beta} - \mathbb{E}(q_t^{\alpha\beta}))(q_t^{\gamma\delta} - \mathbb{E}(q_t^{\gamma\delta}))\right] \\ &= \mathbb{E}(q_t^{\alpha\beta} q_t^{\gamma\delta}) - \mathbb{E}(q_t^{\alpha\beta})\mathbb{E}(q_t^{\gamma\delta}) \\ &= \mathbb{E}(q_t^{\alpha\beta} q_t^{\gamma\delta}) - \mathbb{E}(q_t^{\alpha\beta})^2. \end{aligned} \tag{5.3.58}$$

When both pairs of individuals are the same, $(\alpha, \beta) = (\gamma, \delta)$, the covariance equals the variance of that quantity,

$$\mathrm{Cov}(q_t^{\alpha\beta}, q_t^{\alpha\beta}) = \mathrm{Cov}(q_t^{\alpha\beta}, q_t^{\beta\alpha}) = \mathrm{Var}(q_t^{\alpha\beta}) \tag{5.3.59}$$

and we have already performed this calculation. So now we are going to consider cases with all distinct individuals and with only one in common,

$$\mathrm{Cov}(q_t^{\alpha\beta}, q_t^{\gamma\delta}) = \mathrm{Cov}(t)^{\alpha\beta\gamma\delta} \tag{5.3.60}$$

$$\mathrm{Cov}(q_t^{\alpha\beta}, q_t^{\gamma\beta}) = \mathrm{Cov}(t)^{\alpha\beta\gamma} \tag{5.3.61}$$

in which α, β, γ and δ are all different.

Second Moment with one individual in common

Let us start with the calculation of $\mathbb{E}(q_{t+1}^{\alpha\beta} q_{t+1}^{\gamma\beta}) = \sum_{q_{t+1}^{\alpha\beta}, q_{t+1}^{\gamma\beta}} q_{t+1}^{\alpha\beta} q_{t+1}^{\gamma\beta} \mathcal{P}(q_{t+1}^{\alpha\beta}, q_{t+1}^{\gamma\beta})$, what introduces the task of calculating the joint distribution $\mathcal{P}(q_{t+1}^{\alpha\beta}, q_{t+1}^{\gamma\beta})$. Following the same procedure of section 5.3.2, once we learned that the genomes from different individuals can be treated as independent when conditioned to the individuals' parents, it is not difficult to see that this distribution can be written as

$$\begin{aligned}\mathcal{P}(q_{t+1}^{\alpha\beta}, q_{t+1}^{\gamma\beta}) = \frac{1}{N^3} \sum_{p_1,p_2} \sum_{p_1',p_2'} \sum_{p_1'',p_2''} \frac{A_{p_1p_2} A_{p_1'p_2'} A_{p_1''p_2''}}{N_{p_1} N_{p_1'} N_{p_1''}} \sum_{\mathbf{S}^\alpha,\mathbf{S}^\beta,\mathbf{S}^\gamma} \delta\left(q_{t+1}^{\alpha\beta}, \frac{\mathbf{S}^\alpha \cdot \mathbf{S}^\beta}{B}\right) \delta\left(q_{t+1}^{\gamma\beta}, \frac{\mathbf{S}^\gamma \cdot \mathbf{S}^\beta}{B}\right) \\ \times \prod_{i=1}^{B} F_{i,t}(\alpha, p_1, p_2) F_{i,t}(\beta, p_1', p_2') F_{i,t}(\gamma, p_1'', p_2'')\end{aligned} \tag{5.3.62}$$

where we introduced the parents (p_1'', p_2'') of the individual γ. Then,

$$\begin{aligned}
\mathbb{E}(q_{t+1}^{\alpha\beta} q_{t+1}^{\gamma\beta}) &= \sum_{q_{t+1}^{\alpha\beta}, q_{t+1}^{\gamma\beta}} q_{t+1}^{\alpha\beta} q_{t+1}^{\gamma\beta} \mathcal{P}(q_{t+1}^{\alpha\beta}, q_{t+1}^{\gamma\beta}) \\
&= \frac{1}{N^3} \sum_{p_1,p_2} \sum_{p_1',p_2'} \sum_{p_1'',p_2''} \frac{A_{p_1p_2} A_{p_1'p_2'} A_{p_1''p_2''}}{N_{p_1} N_{p_1'} N_{p_1''}} \sum_{\mathbf{S}^\alpha, \mathbf{S}^\beta, \mathbf{S}^\gamma} \left(\frac{\mathbf{S}^\alpha \cdot \mathbf{S}^\beta}{B} \right) \left(\frac{\mathbf{S}^\gamma \cdot \mathbf{S}^\beta}{B} \right) \\
&\times \prod_{i=1}^{B} F_{i,t}(\alpha, p_1, p_2) F_{i,t}(\beta, p_1', p_2') F_{i,t}(\gamma, p_1'', p_2'') \\
&= \frac{1}{N^3 B^2} \sum_{p_1,p_2} \sum_{p_1',p_2'} \sum_{p_1'',p_2''} \frac{A_{p_1p_2} A_{p_1'p_2'} A_{p_1''p_2''}}{N_{p_1} N_{p_1'} N_{p_1''}} \sum_{\mathbf{S}^\alpha, \mathbf{S}^\beta, \mathbf{S}^\gamma} \left(\sum_{j,k} s_j^\alpha s_j^\beta s_k^\gamma s_k^\beta \right) \\
&\times \prod_{i=1}^{B} F_{i,t}(\alpha, p_1, p_2) F_{i,t}(\beta, p_1', p_2') F_{i,t}(\gamma, p_1'', p_2'') \\
&= \frac{1}{N^3 B^2} \sum_{p_1,p_2} \sum_{p_1',p_2'} \sum_{p_1'',p_2''} \frac{A_{p_1p_2} A_{p_1'p_2'} A_{p_1''p_2''}}{N_{p_1} N_{p_1'} N_{p_1''}} \\
&\times \sum_{\mathbf{S}^\alpha, \mathbf{S}^\beta, \mathbf{S}^\gamma} \sum_{j=1}^{B} \left(s_j^\alpha s_j^\gamma + \sum_{k \neq j} s_j^\alpha s_j^\beta s_k^\gamma s_k^\beta \right) \prod_{i=1}^{B} F_{i,t}(\alpha) F_{i,t}(\beta) F_{i,t}(\gamma). \qquad (5.3.63)
\end{aligned}$$

The last line of this equation is calculated with the same technique as before,

$$\begin{aligned}
&\sum_{\mathbf{S}^\alpha, \mathbf{S}^\beta, \mathbf{S}^\gamma} \sum_{j=1}^{B} \left(s_j^\alpha s_j^\gamma + \sum_{k \neq j} s_j^\alpha s_j^\beta s_k^\gamma s_k^\beta \right) \prod_{i=1}^{B} F_{i,t}(\alpha) F_{i,t}(\beta) F_{i,t}(\gamma) \\
&= \sum_{j=1}^{B} \frac{e^{-4\mu}}{4} (s_j^{p_1} + s_j^{p_2})(s_j^{p_1''} + s_j^{p_2''}) + \sum_{j=1}^{B} \sum_{k \neq j} \frac{e^{-8\mu}}{16} (s_j^{p_1} + s_j^{p_2})(s_j^{p_1'} + s_j^{p_2'})(s_k^{p_1''} + s_k^{p_2''})(s_k^{p_1'} + s_k^{p_2'}) \\
&= \sum_{j=1}^{B} \frac{e^{-4\mu}}{4} (s_j^{p_1} + s_j^{p_2})(s_j^{p_1''} + s_j^{p_2''}) + \sum_{j=1}^{B} \frac{e^{-8\mu}}{16} (s_j^{p_1} + s_j^{p_2})(s_j^{p_1'} + s_j^{p_2'}) \sum_{k=1}^{B} (s_k^{p_1''} + s_k^{p_2''})(s_k^{p_1'} + s_k^{p_2'}) \\
&- \sum_{j=1}^{B} \frac{e^{-8\mu}}{16} (s_j^{p_1} + s_j^{p_2})(s_j^{p_1'} + s_j^{p_2'})(s_j^{p_1''} + s_j^{p_2''})(s_j^{p_1'} + s_j^{p_2'}) \\
&= B \frac{e^{-4\mu}}{4} (q_t^{p_1p_1''} + q_t^{p_1p_2''} + q_t^{p_2p_1''} + q_t^{p_2p_2''}) \\
&+ B^2 \frac{e^{-8\mu}}{16} (q_t^{p_1p_1'} + q_t^{p_1p_2'} + q_t^{p_2p_1'} + q_t^{p_2p_2'})(q_t^{p_1''p_1'} + q_t^{p_1''p_2'} + q_t^{p_2''p_1'} + q_t^{p_2''p_2'}) \\
&- B \frac{e^{-8\mu}}{8} \left[(q_t^{p_1p_1''} + q_t^{p_1p_2''} + q_t^{p_2p_1''} + q_t^{p_2p_2''}) + (q_t^{p_1p_1''p_1'p_2'} + q_t^{p_1p_2''p_1'p_2'} + q_t^{p_2p_1''p_1'p_2'} + q_t^{p_2p_2''p_1'p_2'}) \right]. \qquad (5.3.64)
\end{aligned}$$

Now, we can write

$$\mathbb{E}(q_{t+1}^{\alpha\beta} q_{t+1}^{\gamma\beta}) = \frac{1}{N^3} \sum_{p_1,p_2} \sum_{p_1',p_2'} \sum_{p_1'',p_2''} \frac{A_{p_1p_2} A_{p_1'p_2'} A_{p_1''p_2''}}{N_{p_1} N_{p_1'} N_{p_1''}}$$

$$\times \left\{ \frac{e^{-8\mu}}{16}(q_t^{p_1p_1'} + q_t^{p_1p_2'} + q_t^{p_2p_1'} + q_t^{p_2p_2'})(q_t^{p_1''p_1'} + q_t^{p_1''p_2'} + q_t^{p_2''p_1'} + q_t^{p_2''p_2'}) \right.$$
$$+ \frac{1}{B}\left[\frac{e^{-4\mu}}{4} - \frac{e^{-8\mu}}{8}\right](q_t^{p_1p_1''} + q_t^{p_1p_2''} + q_t^{p_2p_1''} + q_t^{p_2p_2''})$$
$$\left. -\frac{e^{-8\mu}}{8B}(q_t^{p_1p_1''p_1'p_2'} + q_t^{p_1p_2''p_1'p_2'} + q_t^{p_2p_1''p_1'p_2'} + q_t^{p_2p_2''p_1'p_2'}) \right\}. \tag{5.3.65}$$

Covariance with one individual in common and no assortative reproduction

As we did for the variance, we shall now consider the case without assortative reproduction. Although in this case there is a sum over six indexes, a maximum of only four indexes appear in each term,

$$\begin{aligned}
\mathbb{E}(q_{t+1}^{\alpha\beta}q_{t+1}^{\gamma\beta}) &= \frac{1}{N^3(N-1)^3}\sum_{p_1,p_2}\sum_{p_1',p_2'}\sum_{p_1'',p_2''} A_{p_1p_2}A_{p_1'p_2'}A_{p_1''p_2''} \\
&\times \left\{ \frac{e^{-8\mu}}{2}(q_t^{p_1p_1'}q_t^{p_1'p_1''} + q_t^{p_1p_1'}q_t^{p_2'p_2''}) + \frac{1}{B}\left[e^{-4\mu} - \frac{e^{-8\mu}}{2}\right] q_t^{p_1p_1''} - \frac{e^{-8\mu}}{2B}q_t^{p_1p_1''p_1'p_2'} \right\} \\
&= \frac{1}{B}\left[e^{-4\mu} - \frac{e^{-8\mu}}{2}\right]\left[\frac{1}{N} + \left(1 - \frac{1}{N}\right)\langle q_t^{p_1p_1''}\rangle_P\right] \\
&+ \frac{e^{-8\mu}}{2N^3}\sum_{p_1,p_1',p_1''} q_t^{p_1p_1'}q_t^{p_1'p_1''} + \frac{e^{-8\mu}}{2N^3(N-1)}\sum_{p_1,p_1',p_2',p_1''} A_{p_1'p_2'}q_t^{p_1p_1'}q_t^{p_2'p_1''} \\
&- \frac{e^{-8\mu}}{2BN^3(N-1)}\sum_{p_1,p_1',p_2',p_1''} A_{p_1'p_2'}q_t^{p_1p_1''p_1'p_2'}
\end{aligned} \tag{5.3.66}$$

and using the expansions in Appendix A.2 we find, for the sums,

$$\sum_{p_1,p_1',p_1''} q_t^{p_1p_1'}q_t^{p_1'p_1''} = N(N-1)\left[\frac{1}{N-1} + 2\langle q_t^{p_1p_1'}\rangle_P + \langle (q_t^{p_1p_1'})^2\rangle_P + (N-2)\langle q_t^{p_1p_1'}q_t^{p_1'p_1''}\rangle_P\right] \tag{5.3.67}$$

$$\begin{aligned}
\sum_{p_1,p_1',p_2',p_1''} A_{p_1'p_2'}q_t^{p_1p_1'}q_t^{p_2'p_1''} &= N(N-1)\left[1 + 2(N-1)\langle q_t^{p_1p_1'}\rangle_P + \langle (q_t^{p_1p_1'})^2\rangle_P \right. \\
&\left. + (N-2)(N-3)\langle q_t^{p_1p_1'}q_t^{p_2p_2'}\rangle_P + 3(N-2)\langle q_t^{p_1p_1'}q_t^{p_1'p_1''}\rangle_P\right]
\end{aligned} \tag{5.3.68}$$

$$\sum_{p_1,p_1',p_2',p_1''} A_{p_1'p_2'}q_t^{p_1p_1''p_1'p_2'} = N(N-1)\left[2 + (5N-8)\langle q_t^{p_1p_1'}\rangle_P + (N-2)(N-3)\langle q_t^{p_1p_2p_1'p_2'}\rangle_P\right] \tag{5.3.69}$$

and joining all these results together, and subtracting $\mathbb{E}(q_{t+1}^{\alpha\beta})^2$ (given by Eq.(5.3.44)) we get

$$\begin{aligned}
\mathrm{Cov}(t+1)^{\alpha\beta\gamma} &= \frac{e^{-4\mu}}{B}\left[\frac{1}{N}+\left(1-\frac{1}{N}\right)\langle q_t^{p_1p_1'}\rangle_P\right] \\
&- \frac{e^{-8\mu}}{2B}\left[\frac{2}{N^2}+\frac{1}{N}+\left(1+\frac{4}{N}-\frac{8}{N^2}\right)\langle q_t^{p_1p_1'}\rangle_P+\left(1-\frac{2}{N}\right)\left(1-\frac{3}{N}\right)\langle q_t^{p_1p_2p_1'p_2'}\rangle_P\right] \\
&+ \frac{e^{-8\mu}(n-2)^2}{2N^3}\left[\langle (q_t^{p_1p_1'})^2\rangle_P+(N-4)\langle q_t^{p_1p_1'}q_t^{p_1p_2'}\rangle_P^2-(N-3)\langle q_t^{p_1p_1'}q_t^{p_2p_2'}\rangle_P\right] \\
&= \frac{e^{-4\mu}}{B}\left[\frac{1}{N}+\left(1-\frac{1}{N}\right)\langle q_t^{p_1p_1'}\rangle_P\right] \\
&- \frac{e^{-8\mu}}{2B}\left[\frac{2}{N^2}+\frac{1}{N}+\left(1+\frac{4}{N}-\frac{8}{N^2}\right)\langle q_t^{p_1p_1'}\rangle_P+\left(1-\frac{2}{N}\right)\left(1-\frac{3}{N}\right)\langle q_t^{p_1p_2p_1'p_2'}\rangle_P\right] \\
&+ \frac{e^{-8\mu}(n-2)^2}{2N^3}\left(1-\frac{1}{\overline{\overline{N}}}\right)\left[\mathrm{Var}(q_t^{p_1p_1'})+(N-4)\mathrm{Cov}(t)^{\alpha\beta\gamma}-(N-3)\mathrm{Cov}(t)^{\alpha\beta\gamma\delta}\right].
\end{aligned} \tag{5.3.70}$$

Considering again that the population values are good estimators for the theoretical values, we find the recurrence equation presented at the beginning of the section.

Covariance with no individual in common

Calculating this moment requires the joint distribution $\mathcal{P}(q_{t+1}^{\alpha\beta}q_{t+1}^{\gamma\delta})$, which can be calculated analogously as before, being easy to see that

$$\begin{aligned}
\mathcal{P}(q_{t+1}^{\alpha\beta},q_{t+1}^{\gamma\delta}) &= \frac{1}{N^4}\sum_{p_1,p_2}\sum_{p_1',p_2'}\sum_{p_1'',p_2''}\sum_{p_1''',p_2'''}\frac{A_{p_1p_2}A_{p_1'p_2'}A_{p_1''p_2''}A_{p_1'''p_2'''}}{N_{p_1}N_{p_1'}N_{p_1''}N_{p_1'''}} \\
&\times \sum_{\mathbf{S}^\alpha,\mathbf{S}^\beta,\mathbf{S}^\gamma,\mathbf{S}^\delta}\delta\left(q_{t+1}^{\alpha\beta},\frac{\mathbf{S}^\alpha\cdot\mathbf{S}^\beta}{B}\right)\delta\left(q_{t+1}^{\gamma\delta},\frac{\mathbf{S}^\gamma\cdot\mathbf{S}^\delta}{B}\right) \\
&\times \prod_{i=1}^{B}F_{i,t}(\alpha,p_1,p_2)F_{i,t}(\beta,p_1',p_2')F_{i,t}(\gamma,p_1'',p_2'')F_{i,t}(\delta,p_1''',p_2'''),
\end{aligned} \tag{5.3.71}$$

where we introduced the parents (p_1''',p_2''') of the individual δ. Then,

$$\begin{aligned}
\mathbb{E}(q_{t+1}^{\alpha\beta}q_{t+1}^{\gamma\delta}) &= \sum_{q_{t+1}^{\alpha\beta},q_{t+1}^{\gamma\delta}}q_{t+1}^{\alpha\beta}q_{t+1}^{\gamma\delta}\mathcal{P}(q_{t+1}^{\alpha\beta},q_{t+1}^{\gamma\beta}) \\
&= \frac{1}{N^4}\sum_{p_1,p_2}\sum_{p_1',p_2'}\sum_{p_1'',p_2''}\sum_{p_1''',p_2'''}\frac{A_{p_1p_2}A_{p_1'p_2'}A_{p_1''p_2''}A_{p_1'''p_2'''}}{N_{p_1}N_{p_1'}N_{p_1''}N_{p_1'''}}\sum_{\mathbf{S}^\alpha,\mathbf{S}^\beta,\mathbf{S}^\gamma,\mathbf{S}^\delta}\left(\frac{\mathbf{S}^\alpha\cdot\mathbf{S}^\beta}{B}\right)\left(\frac{\mathbf{S}^\gamma\cdot\mathbf{S}^\delta}{B}\right) \\
&\times \prod_{i=1}^{B}F_{i,t}(\alpha,p_1,p_2)F_{i,t}(\beta,p_1',p_2')F_{i,t}(\gamma,p_1'',p_2'')F_{i,t}(\delta,p_1''',p_2''') \\
&= \frac{1}{N^2}\sum_{p_1,p_2}\sum_{p_1',p_2'}\frac{A_{p_1p_2}A_{p_1'p_2'}}{N_{p_1}N_{p_1'}}\sum_{\mathbf{S}^\alpha,\mathbf{S}^\beta}\left(\frac{\mathbf{S}^\alpha\cdot\mathbf{S}^\beta}{B}\right)\prod_{i=1}^{B}F_{i,t}(\alpha,p_1,p_2)F_{i,t}(\beta,p_1',p_2')
\end{aligned}$$

$$\times \frac{1}{N^2} \sum_{p_1'',p_2''} \sum_{p_1''',p_2'''} \frac{A_{p_1''p_2''}A_{p_1'''p_2'''}}{N_{p_1''}N_{p_1'''}} \sum_{\mathbf{S}^\gamma,\mathbf{S}^\delta} \left(\frac{\mathbf{S}^\gamma \cdot \mathbf{S}^\delta}{B}\right) \prod_{i=1}^{B} F_{i,t}(\gamma, p_1'', p_2'') F_{i,t}(\delta, p_1''', p_2''')$$
$$= \mathbb{E}(q_{t+1}^{\alpha\beta})\mathbb{E}(q_{t+1}^{\gamma\delta}). \tag{5.3.72}$$

This result shows that the covariance of similarities that do not share common individuals is zero,

$$\mathrm{Cov}(t)^{\alpha\beta\gamma\delta} = \mathbb{E}(q_{t+1}^{\alpha\beta}q_{t+1}^{\gamma\delta}) - \mathbb{E}(q_{t+1}^{\alpha\beta})\mathbb{E}(q_{t+1}^{\gamma\delta}) = 0. \tag{5.3.73}$$

5.4 On higher-order overlaps

The Derrida-Higgs model, as we are trying to describe, can be understood as an indexed family of random vectors $\boldsymbol{Q}_t$ defined as

$$\boldsymbol{Q}_t \equiv \{q_t^{12}, q_t^{13}, \ldots, q_t^{1N}, q_t^{23}, q_t^{24}, \ldots, q_t^{2N}, \ldots, q_t^{N-1,N}\} = \{q_t^{ij} | 1 \leq i < j \leq N\} \tag{5.4.1}$$

i.e., the upper triangle of the similarity matrix $\mathbb{Q}_t$, whose entries follow a distribution with mean and covariance which we have calculated in the previous section. In fact, the probability distributions we have calculated so far are marginals of the distribution $\mathcal{P}(\boldsymbol{Q}_{t+1})$. For instance,

$$\mathcal{P}(q_{t+1}^{\alpha\beta}) = \sum_{q_{t+1}^{12}} \cdots \sum_{q_{t+1}^{\alpha 1}} \cdots \sum_{q_{t+1}^{\alpha,\beta-1}} \sum_{q_{t+1}^{\alpha,\beta+1}} \cdots \sum_{q_{t+1}^{N-1,N}} \mathcal{P}(\boldsymbol{Q}_{t+1}) \tag{5.4.2}$$

with the property of being the same regardless of α and β (with $\alpha \neq \beta$). Taking into account the procedure we have introduced in order to calculate $\mathcal{P}(q_{t+1}^{\alpha\beta})$, it is not hard to calculate $\mathcal{P}(\boldsymbol{Q}_{t+1})$. Let $p_1(\alpha_i)$ be the focal parent of the individual α_i, and $p_2(\alpha_i)$ its second parent (p_1 and p_2 are taken from the generation t while α_i is from generation $t+1$). Then $\mathcal{P}(\boldsymbol{Q}_{t+1})$ takes the form

$$\mathcal{P}(\boldsymbol{Q}_{t+1}) = \frac{1}{N^N} \sum_{p_1(\alpha_1),p_2(\alpha_1)} \cdots \sum_{p_1(\alpha_N),p_2(\alpha_N)} \left(\prod_{i=1}^{N} \frac{A_{p_1(\alpha_i)p_2(\alpha_i)}}{N_{p_1(\alpha_i)}}\right)$$
$$\times \sum_{\mathbf{S}^1} \cdots \sum_{\mathbf{S}^N} \left[\prod_{\{q_{t+1}^{ij} | 1 \leq i < j \leq N\}} \delta\left(q_{t+1}^{ij}, \frac{\mathbf{S}^i \cdot \mathbf{S}^j}{B}\right)\right] \left[\prod_{k=1}^{B} \left(\prod_{l=1}^{N} F_{k,t}(\alpha_l, p_1(\alpha_l), p_2(\alpha_l))\right)\right]. \tag{5.4.3}$$

On the other hand, as we have seen, second moments of $\mathcal{P}(\boldsymbol{Q}_{t+1})$ also depend on the second order overlap, and it is not hard to see that a third moment of this distribution would also depend on the *third order overlap*, and so on. For instance, let us consider the third moment $\mathbb{E}(q_{t+1}^{\alpha\beta}q_{t+1}^{\gamma\beta}q_{t+1}^{\gamma\alpha})$, whose involved probability is

$\mathcal{P}(q_{t+1}^{\alpha\beta}, q_{t+1}^{\gamma\beta}, q_{t+1}^{\gamma\alpha})$

$$= \frac{1}{N^3} \sum_{p_1,p_2} \sum_{p_1',p_2'} \sum_{p_1'',p_2''} \frac{A_{p_1p_2}A_{p_1'p_2'}A_{p_1''p_2''}}{N_{p_1}N_{p_1'}N_{p_1''}} \sum_{\mathbf{S}^\alpha,\mathbf{S}^\beta,\mathbf{S}^\gamma} \delta\left(q_{t+1}^{\alpha\beta}, \frac{\mathbf{S}^\alpha\cdot\mathbf{S}^\beta}{B}\right)\delta\left(q_{t+1}^{\gamma\beta}, \frac{\mathbf{S}^\gamma\cdot\mathbf{S}^\beta}{B}\right)\delta\left(q_{t+1}^{\gamma\alpha}, \frac{\mathbf{S}^\gamma\cdot\mathbf{S}^\alpha}{B}\right)$$
$$\times \prod_{i=1}^{B} F_{i,t}(\alpha,p_1,p_2)F_{i,t}(\beta,p_1',p_2')F_{i,t}(\gamma,p_1'',p_2''), \tag{5.4.4}$$

and thus, when calculating the moment, we find the term

$$\begin{aligned}
&\frac{1}{B^3}\sum_{\mathbf{S}^\alpha,\mathbf{S}^\beta,\mathbf{S}^\gamma}\sum_{j,k,l} s_j^\alpha s_j^\beta s_k^\beta s_k^\gamma s_l^\gamma s_l^\alpha \prod_{i=1}^{B} F_{i,t}(\alpha,p_1,p_2)F_{i,t}(\beta,p_1',p_2')F_{i,t}(\gamma,p_1'',p_2'')\\
=&\frac{1}{B^3}\left(\frac{e^{-2\mu}}{2}\right)^6 \sum_{j,k,l}(s_j^{p_1}+s_j^{p_2})(s_j^{p_1'}+s_j^{p_2'})(s_k^{p_1'}+s_k^{p_2'})(s_k^{p_1''}+s_k^{p_2''})(s_l^{p_1''}+s_l^{p_2''})(s_l^{p_1}+s_l^{p_2})\\
=&\frac{1}{B^3}\left(\frac{e^{-2\mu}}{2}\right)^6 \left[\ldots + 8\sum_k s_k^{p_1}s_k^{p_2}s_k^{p_1'}s_k^{p_2'}s_k^{p_1''}s_k^{p_2''}\right],
\end{aligned} \tag{5.4.5}$$

where we can recognize the third order overlap,

$$q^{p_1p_2p_1'p_2'p_1''p_2''} \equiv \frac{1}{B}\sum_k s_k^{p_1}s_k^{p_2}s_k^{p_1'}s_k^{p_2'}s_k^{p_1''}s_k^{p_2''}. \tag{5.4.6}$$

Thus, it is not possible to completely change from the genome description to *only* the first order overlap description, once the evolution of its distribution depends on higher order overlaps.

5.4.1 The definition

For completeness, let us introduce a general definition for the overlap. Let n individuals $\{\alpha_1,\alpha_2,\ldots,\alpha_n\}$ all with their own genome $\{s_1^{\alpha_k},\ldots,s_B^{\alpha_k}\}$. The $j-$order overlap of the $2j$ individuals $\{\alpha_{i_1},\ldots,\alpha_{i_{2j}}\}$ with $\{i_1,\ldots,i_{2j}\}\subset\{1,\ldots,n\}$ is defined by

$$q^{(j)}(i_1,\ldots,i_{2j}) \equiv \frac{1}{B}\sum_{k=1}^{B} s_k^{\alpha_{i_1}}s_k^{\alpha_{i_2}}\ldots s_k^{\alpha_{i_{2j}}}. \tag{5.4.7}$$

Notice that we have changed the notation from what we have used so far, since carrying the individuals as upper indexes can be quite messy for higher order overlaps.

5.4.2 Properties

Let us now introduce some of its properties.

Identity property

If all individuals have the same genome, then the overlap (of any order) equals 1,

$$q^{(j)}(i_1,\ldots,i_{2j}) = \frac{1}{B}\sum_{k=1}^{B} s_k^{\alpha_{i_1}} s_k^{\alpha_{i_2}} \ldots s_k^{\alpha_{i_{2j}}} = \frac{1}{B}\sum_{k=1}^{B} 1 = 1. \tag{5.4.8}$$

Permutation symmetry

The overlap (of any order) does not change under permutations of individuals,

$$\begin{aligned} q^{(j)}(\ldots\alpha_l,\ldots,\alpha_m\ldots) &= \frac{1}{B}\sum_{k=1}^{B} \ldots s_k^{\alpha_l}\ldots s_k^{\alpha_m}\ldots \\ &= \frac{1}{B}\sum_{k=1}^{B} \ldots s_k^{\alpha_m}\ldots s_k^{\alpha_l}\ldots \\ &= q^{(j)}(\ldots\alpha_m,\ldots,\alpha_l\ldots). \end{aligned} \tag{5.4.9}$$

The reduced order property

If two individuals of the set $\{\alpha_{i_1},\ldots,\alpha_{i_{2j}}\}$ are the same, then the $j-$order overlap equals the $(j-1)-$order overlap of the same set without these two individuals. Suppose $\alpha_l = \alpha_m$. Then,

$$\begin{aligned} q^{(j)}(\ldots,\alpha_{l-1},\alpha_l,\alpha_{l+1},\ldots\ \alpha_{m-1},\alpha_m,\alpha_{l+1}\ldots) &= \frac{1}{B}\sum_{k=1}^{B} \ldots s_k^{\alpha_{l-1}} s_k^{\alpha_l} s_k^{\alpha_{l+1}} \ldots s_k^{\alpha_{m-1}} s_k^{\alpha_m} s_k^{\alpha_{m+1}} \ldots \\ &= \frac{1}{B}\sum_{k=1}^{B} \left(\ldots s_k^{\alpha_{l-1}} s_k^{\alpha_{l+1}} \ldots s_k^{\alpha_{m-1}} s_k^{\alpha_{m+1}} \ldots\right) s_k^{\alpha_l} s_k^{\alpha_m} \\ &= \frac{1}{B}\sum_{k=1}^{B} \left(\ldots s_k^{\alpha_{l-1}} s_k^{\alpha_{l+1}} \ldots s_k^{\alpha_{m-1}} s_k^{\alpha_{m+1}} \ldots\right) \\ &= q^{(j-1)}(\ldots,\alpha_{l-1},\alpha_{l+1},\ldots\ \alpha_{m-1},\alpha_{l+1}\ldots). \end{aligned} \tag{5.4.10}$$

Indeed, if $\{\alpha_{i_1},\ldots,\alpha_{i_{2j}}\}$ has m pairs of equal individuals, then its $j-$order overlap equals the $(j-m)-$order of the same set excluding the m pairs.

First order mean evolution

In the absence of q_{min}, we can approximate the evolution of the $j-$order overlap as follows. Considering $\{\alpha_1,\ldots,\alpha_{2j}\}$ of $2j$ different individuals in a population of size N, and extending the result of equation (5.3.55), we can write

$$\mathbb{E}(q^{(j)}(\alpha_1,\ldots,\alpha_{2j})) = e^{-4j\mu}\left(\frac{1}{N}\right)^{2j} \sum_{p_1(\alpha_1),\ldots,p_1(\alpha_{2j})} q^{(j)}(p_1(\alpha_1),\ldots,p_1(\alpha_{2j})). \tag{5.4.11}$$

We can approximate the last two factors of this expression if we remember that two individuals of the set can be the same $\{(p_1(\alpha_1), \ldots, p_1(\alpha_{2j})\}$ and then use the reduced order property. If one is going to form a set of entries, there can be a pair of the same individual on $2j(2j-1)/2!$ different positions. Choosing the entries at random (individuals) from a total of N, given an individual, the chance of choosing another one equal to the first is $1/N$. Then,

$$\mathbb{E}(q_{t+1}^{(j)}) \approx e^{-4j\mu}\left[\frac{j(2j-1)}{N}\langle q_t^{(j-1)}\rangle_P + \left(1 - \frac{j(2j-1)}{N}\right)\langle q_t^{(j)}\rangle_P\right], \tag{5.4.12}$$

which approximates the evolution up to order $1/N$.

5.5 The one-parent model

In 1991, Derrida and Higgs also worked on what they called the *one-parent model* [94, 121], which considers an asexual version of what we have done so far. We include it here for completeness, since a similar process appears in the next chapters. The mitochondrion is an organelle in eukaryotic cells that has its own genetic material, and it is transmitted without recombination from the mother to the offspring [122], which is similar to an asexual replication process, and we work on this model in Chapter 8. Also, the epidemic model with viral evolution we introduce in Chapter 10 poses a similar mechanism.

In the one-parent model, a finite size population of N individuals evolve under asexual reproduction and has no generational overlap. A focal individual is chosen at random, with uniform probability $1/N$ and its offspring is going to be a part of the next generation. The genome of every individual α is, as before, a binary sequence $\mathbf{S}^\alpha$ of B alleles $s_i^\alpha = \pm 1$. Whenever an individual replicates, its offspring heirs the same alleles as its parents, plus mutations at rate μ. As the reader can see, besides the absence of genome recombination, there are no further differences in the description of the model, despite its results being interestingly not the same.

As the recombination mixes the genetic material of the individuals, its absence creates genetic lineages that may randomly be conserved or not due to random genetic drift. Suppose a population with $N = 2$ individuals and with genetic similarity q_t. If they both reproduce, the genetic similarity in the next step would be close to $q_{t+1} \approx q_t(1 - 4\mu)$. On the other hand, if only one of the individuals reproduces twice, the similarity would be $q_{t+1} \approx 1 - 4\mu$. Since there is no recombination, the effect of random drift can be much more drastic and large deviations from the mean are much more frequent. Hence the expected similarity evolution is not representative of any given realization of the process. We call such property as non *self-averaging* [121].

Another way to understand the non-triviality of this process is to think about two different lineages evolving in a population. Every descent of a lineage has high similarity with each other, but they have low similarity with respect to individuals of the other lineage. What would happen if randomly one lineage disappear? The similarity would increase! This would not happen in a sexual population because two different lineages recombine, they are not isolated. The similarity distribution for the asexual case would show different peaks for each lineage and a third one for the similarity between both lineages. As soon as a lineage disappears, there would remain only one peak in the distribution, which would break into other peaks as time passes by and the genetic drift

acts. The picture displayed by the one-parent model is very similar to the dynamics at the species level in the high diversity phase of the Derrida-Higgs model (Chapter 5), and this map could result in a very interesting theory, although we follow a different route in the following chapters.

Notwithstanding, it is still possible to follow the same math procedure we have introduced to describe the ensemble similarity probability distribution in this case.

5.5.1 The similarity distribution

Because there is no recombination, every individual γ at time $t+1$ comes from only one parent p_1 from time t. Hence, the probability of a genome $\mathbf{S}^{\gamma}_{t+1}$ is

$$\begin{aligned}\mathcal{P}(\mathbf{S}^{\gamma}_{t+1}) &= \sum_{p_1} \mathcal{P}(\mathbf{S}^{\gamma}_{t+1}|p_1)\mathcal{P}(p_1) \\ &= \frac{1}{N}\sum_{p_1}\prod_{i=1}^{B}\left[\frac{1}{2}(1+e^{-2\mu}s^{\gamma}_{i,t+1}s^{p_1}_{i,t})\right], \end{aligned} \tag{5.5.1}$$

where we have used equation (5.3.14) and that $\mathcal{P}(p_1) = 1/N$. Now, using equations (5.3.17) to (5.3.20), but considering that individuals have only one parent, p_1 for α and p'_1 for β, for the similarity distribution between α and β we find:

$$\mathcal{P}(q^{\alpha\beta}_{t+1}) = \frac{1}{N^2}\sum_{p_1,p'_1}\delta\left(q^{\alpha\beta}_{t+1}, \frac{\mathbf{S}^{\alpha}_{t+1}\cdot\mathbf{S}^{\beta}_{t+1}}{B}\right)\prod_{i=1}^{B}\left[\frac{1}{2}(1+e^{-2\mu}s^{\alpha}_{i,t+1}s^{p_1}_{i,t})\right]\left[\frac{1}{2}(1+e^{-2\mu}s^{\beta}_{i,t+1}s^{p'_1}_{i,t})\right]. \tag{5.5.2}$$

A similar result is going to appear in Chapter 8 for the similarity between mitochondrial genetic material.

The moments of the distribution are calculated analogously as before. For instance, the first moment is given by

$$\begin{aligned}\mathbb{E}(q^{\alpha\beta}_{t+1}) &= \sum_{q} q\mathcal{P}(q^{\alpha\beta}_{t+1} = q) \\ &= \frac{e^{-4\mu}}{N^2}\sum_{p_1,p'_1} q^{p_1,p'_1}_{t}, \end{aligned} \tag{5.5.3}$$

which is the same result for the sexual case in the absence of assortative reproduction, but it worths to emphasize that now it is valid only for the ensemble average.

Chapter 6

The Heuristic Approximation to the Transition

6.1 Some clues from simulations

As we have learned from the previous chapter, the Derrida-Higgs model cannot be described only by its similarity distribution: higher moments of the first order overlap distribution (i.e., the genetic similarity) involve higher order overlaps. Thus, we first performed a large number of simulations in order to get some information about the system. With this information, we shall propose a heuristic formula for the transition between the regimes of single and multiple species.

Although approximate, as we will show, the behavior of this formula makes the heuristic solution important for computational reasons. Since the Derrida-Higgs model is a model for species formation, knowing that species can be formed after a given transition curve (which is a characteristic of this solution) helps to guide simulations by making sure that many species are going to be formed.

6.1.1 Before and after the transition

An important aspect of the simulations is that, starting with a clonal population, when the similarity distribution is still *far* from the similarity threshold q_{min}, its dynamics still behaves as if the threshold did not exist. It happens because since the distribution is narrow and still moving towards smaller similarity values, no pair of individuals has similarity smaller than q_{min} and the underlying network is still complete, i.e., every individual can mate with any other individual.

However, things change when the distribution reaches q_{min}. Many individuals are not able to mate anymore, but the system may still reach some stability (discussed in the following subsection). On the other hand, if the system does not find any stability around q_{min}, after passing through a transient behaviour, many peaks appear in the similarity distribution although its average moves again towards the equilibrium q_{eq}. In Fig. 6.1, the evolution of the system for 3 different values of genome size B is shown. When the genome is small enough, the similarity distribution is stationary when it reaches q_{min},

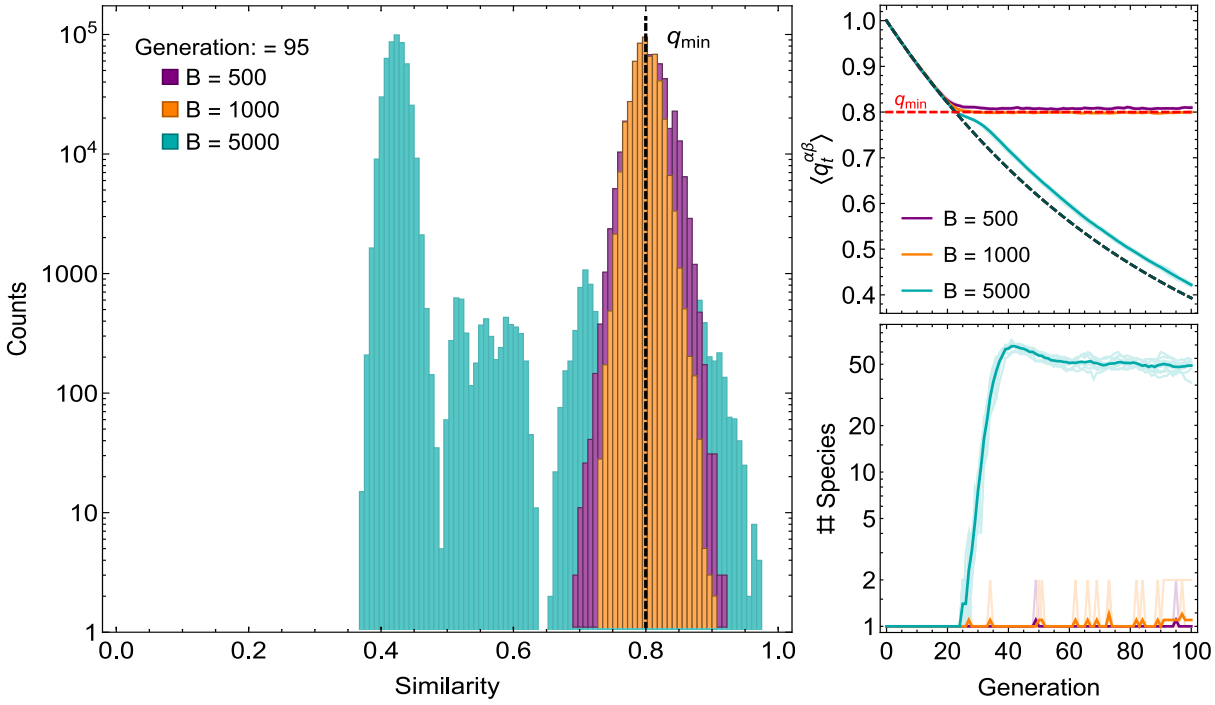


Figure 6.1: **The similarity distribution for different genome sizes.** The big left panel shows the similarity distribution at a given generation ($t = 95$) for three different genome sizes. For a small enough B, the distribution becomes stationary and no species are formed, unless for small fluctuations. When B is large enough, many peaks appear, which is characteristic of the species formation and the average similarity keeps evolving towards q_{eq}. The top right panel shows the average (continuous darker curves) of 10 simulations (shown as lighter curves at the background). The dashed line shows the theoretical value in the case without q_{min}. The bottom right panel shows the species formation for the same set of simulations as the upper panel. In the figure, the simulation parameters are $N = 1000$, $\mu = 0.0025$ and $q_{min} = 0.8$.
Source: Figure produced by the author.

while it shows a complex structure when the genome is much larger and species are formed.

Multiple peaks in the similarity distribution are a signature of the species emergence in the system. The *intraspecific* similarity tends to be high, while the *interspecific* similarity decreases towards zero, since the species evolve uncorrelatedly. Notwithstanding, species are appearing and *disappearing* constantly from the system, and this turnover dynamics prevents the similarity distribution from reaching stationarity, as shown in Fig. 6.2.

6.1.2 Two equilibrium values

Another important aspect of the dynamics is that, when the genome size is not large enough to break the population into different species, the similarity distribution becomes stationary around q_{min}. The system then shows the existence of a new mean similarity equilibrium value, which is approximately q_{min}, as it can be seen in the left panels of Fig.6.3. In this situation, not all the pairs of individuals are able to reproduce, but the remaining pairs, i.e., the remaining network structure, is enough to generate a new popula-

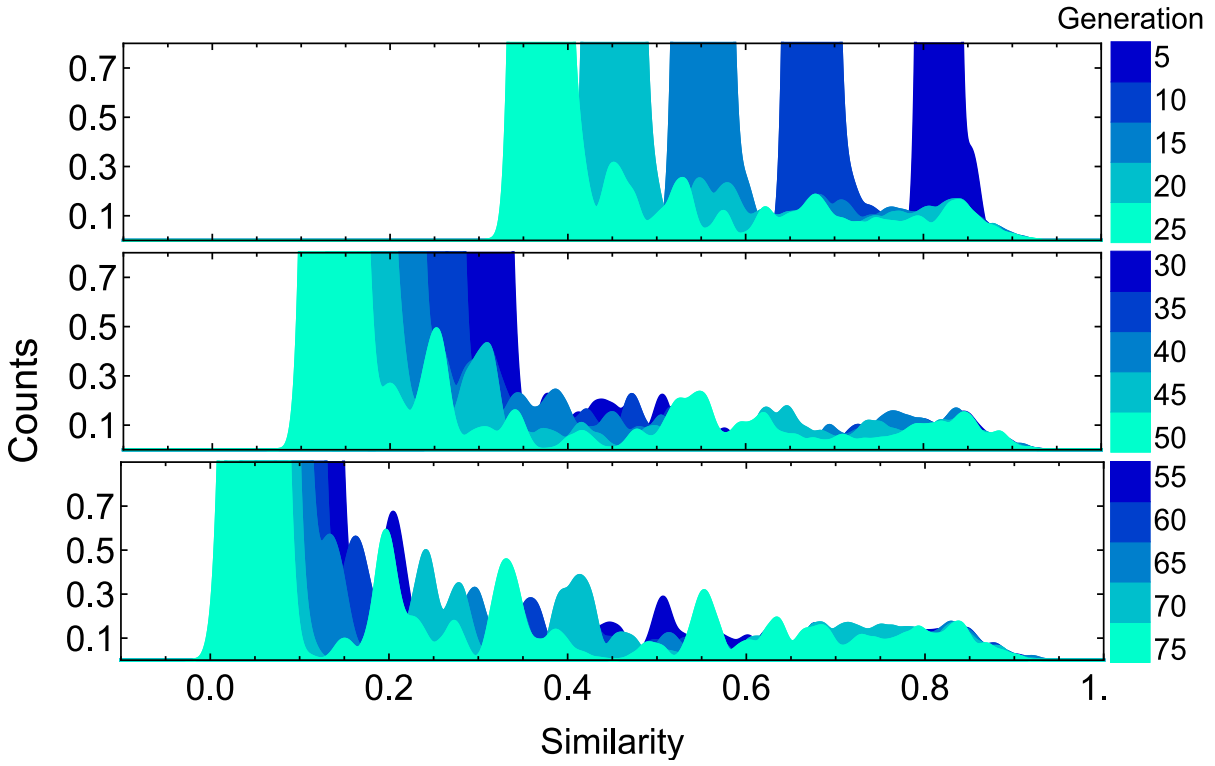


Figure 6.2: **The non-stationarity of the similarity distribution.** The figure shows the similarity distribution in a case with species formation for many generations, showing how the distribution does not reach stationarity along its evolution. In the figure, the simulation parameters are $N = 500$, $B = 5000$, $\mu = 0.01$ and $q_{min} = 0.8$ and the histogram is smoothed out and normalized to 1.
Source: Figure produced by the author.

tion which has the same structure. One could also raise questions about the metastability of this state, but simulations for small enough genome sizes do indicate that this can be a stable equilibrium.

As the genome size increases, this equilibrium starts to become unstable, and more than one species starts to emerge in the system. The right panel of Fig.6.3 shows the evolution of the number of species as a function of the genome size, the greater the value of B, the faster is the species formation.

6.1.3 Important scales

We here define two important scales of the system, which are useful to guide simulations. The first scale is the time up to q_{min}. Starting the system with a clonal population, as we have seen, as long as the distribution has not reached the similarity threshold, the system behaves as a complete network, therefore we can estimate the time τ the distribution takes to reach q_{min} with equation (5.1.11),

$$\tau = \frac{\ln\left[\frac{q_{min}-q_{eq}}{1-q_{eq}}\right]}{\ln\left[\left(1-\frac{1}{N}\right)e^{-4\mu}\right]}, \tag{6.1.1}$$

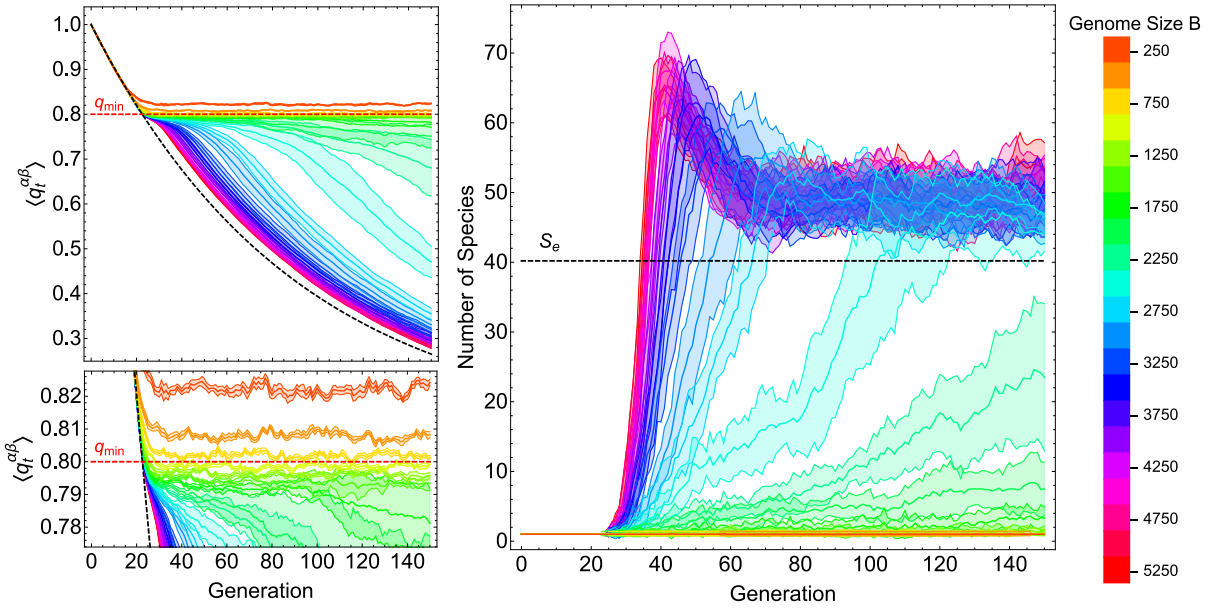


Figure 6.3: **Speciation as a function of B.** The top left panel shows the evolution of the average similarity for many values of B. Each curve is the average of 10 simulations and the interval around them corresponds to the standard deviation. The black dashed line is the theoretical evolution in the absence of q_{min}. The bottom left panel zooms into the upper plot around q_{min}, showing how the evolution finds equilibrium for small values of B. The right panel shows the corresponding evolution of the number of species that are formed in each case, and the black dashed line shows the value of S_e calculated as Eq.(6.1.3). In the figure, the simulation parameters are $N = 1000$, $\mu = 0.0025$ and $q_{min} = 0.8$.
Source: Figure produced by the author.

in which we have considered $q_0 = 1$.

The second scale is the number of species that the system would have (when species appear). A simple estimative is to consider that new species may find its equilibrium similarity distribution centered on q_{min}, which introduces a *characteristic species size* N^*,

$$q_{min} = q_{eq} = \frac{1}{N^* e^{4\mu} - (N^* - 1)} \Rightarrow N^* = \frac{1}{e^{4\mu} - 1}\left(\frac{1}{q_{min}} - 1\right), \tag{6.1.2}$$

and then, an estimative for the number of species S_e can be given by

$$S_e = \frac{N}{N^*} = N(e^{4\mu} - 1)\left(\frac{1}{q_{min}} - 1\right)^{-1}. \tag{6.1.3}$$

However, the simulations have shown more species than this number, which can be simply understood as an effect of the distribution of abundances being asymetrical, i.e., the species have different sizes (called their abundance) and there can be more rare species than highly abundant ones. On the other hand, this is a nice scale for the number of species, once even when the species formation is slow, we can expect the system to reach at least S_e, if given enough time. In the right panel of Fig.6.3, the value S_e is shown in

comparison to the number of species in the system.

6.1.4 Attractive region

Given all the observations so far, now we mention the main ingredient for the solution we are about to propose for the low-high diversity transition. We observe that the region around q_{min}, regarding the evolution of the mean similarity, always slows down the dynamics, even for very large genome sizes, and as we have described before, for small genome sizes, the distribution finds stationarity there. Therefore, since after the transition the average similarity keeps evolving as if without q_{min} (Section 6.1.1), we propose the existence of an *attractive region* around q_{min}.

Since the dynamics is discrete, it is possible that in a given time step, the distribution gets *trapped* inside this region. The size of this region must then decrease as the genome size increases, making it easier for the distribution to *jump over* it. When it happens, there is species formation.

In the bottom left panel of Fig.6.3, a region around q_{min} is zoomed in, and we can see the evolutions of the average similarity stopping as close to q_{min} as B increases, indicating that this attractive region decreases its size for larger values of B, up to a value it cannot keep the system within it, where species start to be formed.

6.2 The ansatz and the solution

The results of the previous section allow us to describe a heuristic approximation to find the critical genome size B_c. Let the attractive region around q_{min} have size Δ. As we have described, the species formation would happen when the mean similarity evolution jumps over this region. Thus, we want the temporal variation of the average to be greater than Δ,

$$|\langle q_{t+1}^{\alpha\beta}\rangle - \langle q_t^{\alpha\beta}\rangle| > \Delta, \tag{6.2.1}$$

when $\langle q_t^{\alpha\beta}\rangle \approx q_{min} + \Delta$. Once that before and after the transition, the mean similarity approximately behaves as a complete network, we can use equation (5.3.30) to find $\langle q_{t+1}^{\alpha\beta}\rangle$. Hence,

$$\Delta < \frac{q_{min}/q_{eq} - 1}{N - 1} \equiv \delta q, \tag{6.2.2}$$

in order to have speciation.

6.2.1 The size Δ

Equation (6.2.2) is in the correct form of what we would expect according to our observations: if the attractive region is small enough, there should exist species formation. Now, we must only find an expression for the size Δ. As an ansatz, we propose

$$\Delta = \sqrt{\sigma_B^2} - \lim_{B\to\infty} \sqrt{\sigma_B^2}, \tag{6.2.3}$$

where

$$\sigma_B^2 = \mathrm{Var}(q_\tau^{\alpha\beta}), \tag{6.2.4}$$

i.e., the variance of the similarity distribution, considering a complete network, calculated when $\langle q_t^{\alpha\beta}\rangle = q_{min}$.

The first important aspect of this ansatz is that when $B \to \infty$, $\Delta \to 0$ and then Eq. (6.2.2) gives $q_{min} > q_{eq}$, which is the speciation condition conjectured by Derrida and Higgs in this limit.

The recurrence equation for the variance takes the form

$$\mathrm{Var}(q_{t+1}) = \frac{a_1}{B} + a_2\mathrm{Var}(q_t),$$

with a_1 and a_2 constants, which leads to

$$\sigma_B^2 = \Lambda_1 + \frac{\Lambda_2}{B}, \tag{6.2.5}$$

where Λ_1 and Λ_2 do not depend on B. In this notation,

$$\Delta = \sqrt{\Lambda_1 + \frac{\Lambda_2}{B}} - \sqrt{\Lambda_1}$$

and therefore, with Eq. (6.2.3),

$$B > \frac{\Lambda_2}{\delta q^2 + 2\delta q\sqrt{\Lambda_1}},$$

which defines the critical genome size

$$\boxed{B_c = \frac{\Lambda_2}{\delta q^2 + 2\delta q\sqrt{\Lambda_1}}}, \tag{6.2.6}$$

in which δq has been defined in equation (6.2.2) and Λ_1 and Λ_2 are given by solving equation (5.3.49) up to time τ. In the results we are going to present, we solve the system considering $\lfloor\tau\rfloor$ (Floor function, i.e., the largest integer smaller or equal to τ) and $\lceil\tau\rceil$ (Ceiling function, i.e., the smallest integer larger or equal to τ). This is not a hard computational task, but to find it analytically is not so simple. In order to find the variance, we must solve the system of recurrence equations

$$\begin{aligned}
\langle q_{t+1}^{\alpha\beta}\rangle &= f_1(\langle q_t^{\alpha\beta}\rangle),\\
\langle q_{t+1}^{\alpha\beta\gamma\delta}\rangle &= f_2(\langle q_t^{\alpha\beta}\rangle, \langle q_{t+1}^{\alpha\beta\gamma\delta}\rangle),\\
\mathrm{Var}(q_{t+1}^{\alpha\beta}) &= f_3(\langle q_t^{\alpha\beta}\rangle, \langle q_{t+1}^{\alpha\beta\gamma\delta}\rangle, \mathrm{Var}(q_t^{\alpha\beta}), \mathrm{Cov}(t)^{\alpha\beta\gamma}),\\
\mathrm{Cov}(t+1)^{\alpha\beta\gamma} &= f_4(\langle q_t^{\alpha\beta}\rangle, \langle q_{t+1}^{\alpha\beta\gamma\delta}\rangle, \mathrm{Var}(q_t^{\alpha\beta}), \mathrm{Cov}(t)^{\alpha\beta\gamma}),
\end{aligned}$$

which, although analytically solvable (once linear), it is sufficiently messy to hinder an

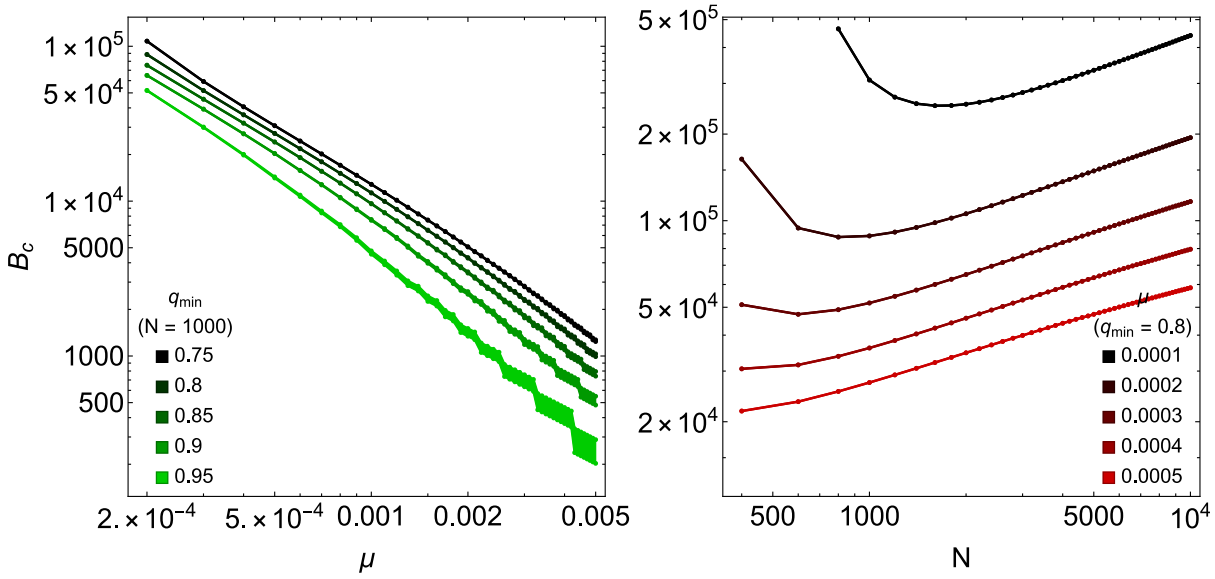


Figure 6.4: **Heuristic critical genome size B_c.** The figure shows the value of B_c given by Eq.(6.2.6) for different parameter values. In the left panel, B_c is calculated as a function of μ for fixed N and different q_{min} values, and in the right panel, B_c is given as a function of N, with fixed q_{min} and different μ values. When calculating the solution up to time τ (Eq.(6.1.1)), τ can be considered $\lfloor\tau\rfloor$ (Floor function) or $\lceil\tau\rceil$ (Ceiling function) and the results for B_c can be different. This difference is shown as bars instead of points in the plots, the top of the bar corresponding to the ceiling result and its bottom to the floor result.
Source: Figure produced by the author.

easy mathematical form for B_c as a function of the parameters N, μ and q_{min}. Fig. 6.4 shows B_c numerically calculated in different regions of the parameter space. In the next section, we compare this solution with simulations.

6.2.2 Comparison with simulation

To compare this solution with simulations is not an easy task. First, remember that the B_c is defined when the expected number of species at time $t \to \infty$ is greater than one. However, since this is a heuristic approximation, we do not have a characteristic time-scale to be sure that we have simulated the system for a sufficiently long time. Of course, we can expect to observe many species after a very small transient time if $B \gg B_c$ or only one species when $B \ll B_c$. But close to the transition, many species may be formed very slowly or not be formed at all. Thus, simulations must be consistently run up to a well chosen time horizon. On the other hand, the simulations can take a long time to finish, since we calculate the similarity of every pair of individuals, the simulation time scales at least with N^2.

We chose the number of generations as 3τ (with τ calculated here as $\lceil\tau\rceil$). The first τ generations account for reaching the threshold; the following τ generations account for a transient behavior which we try to avoid. We calculate the average number of species $\langle S\rangle_\tau$ over the last τ generations. Of course, the more generations we simulate, the more precise the result we get is, but the computational time is an important limiting factor.

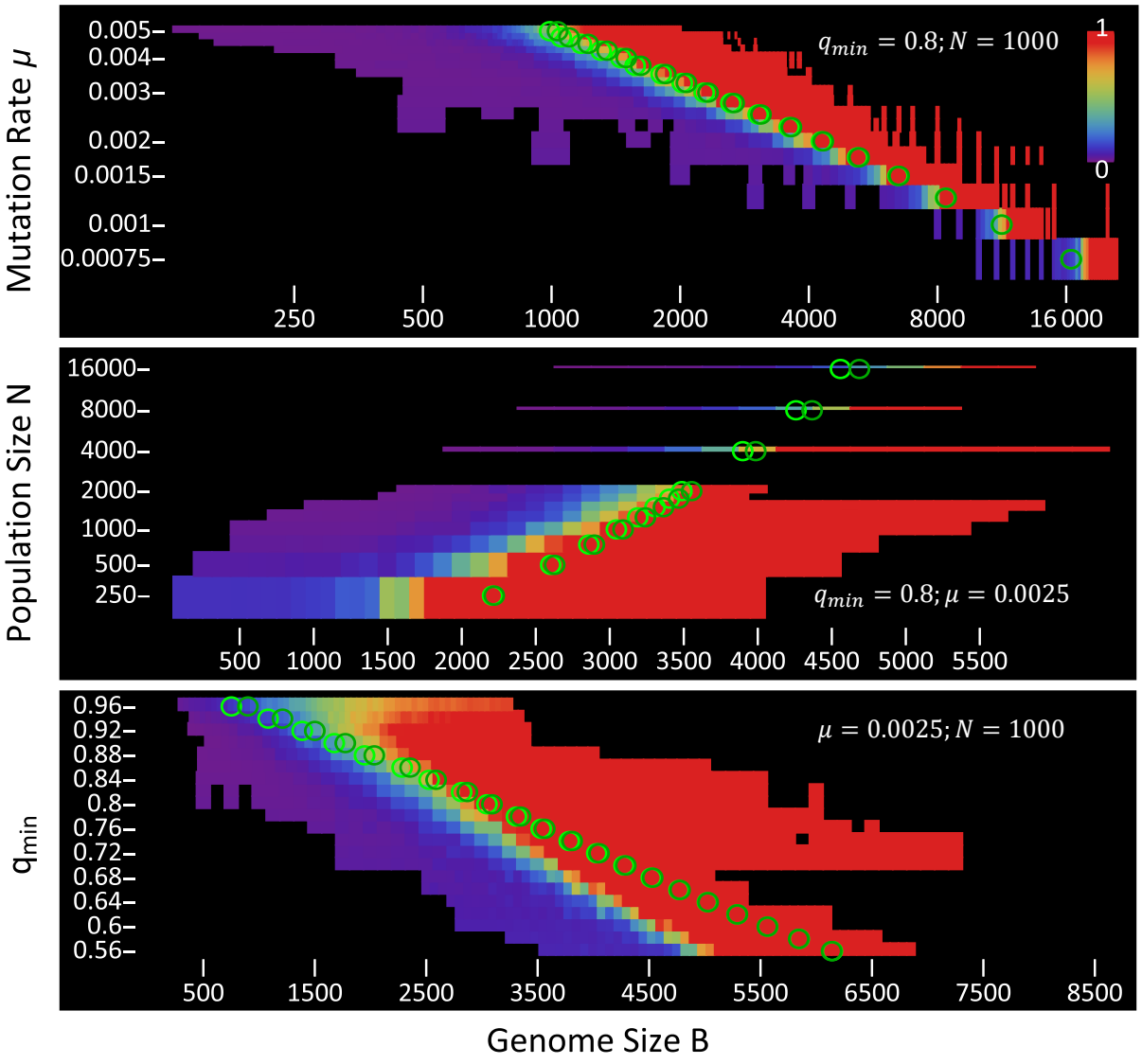


Figure 6.5: **Heuristic solution and simulations.** The figure shows the heuristic solution (green circles; calculations with $\lfloor \tau \rfloor$ (light green) and $\lceil \tau \rceil$ (dark green)) and the result from simulations, shown as a heatmap for different regions of the parameter set. The colors show the normalized number of species $\langle S \rangle_\tau / S_e$ calculated as described in the present section, ranging from 0 to 1 (whenever the normalized number of species is greater than 1, it is still plotted as 1). Thus, the purple region shows no species formation, while in the red region species have appeared beyond the subestimated value S_e. The black region was not simulated.
Source: Figure produced by the author.

The other important factor is to understand the meaning of *many* species. In order to do that, we compare the average number of species $\langle S \rangle_\tau$ that we find with the estimated value S_e. Thus, for $\langle S \rangle_\tau / S_e$ greater or close to 1, we can say that the system has really gone through speciation. But when this value is close to zero ($\approx 1/S_e$), then the system did not form species. Values between $1/S_e$ and 1 are characteristic of the transient behavior when the equilibrium richness value is still being reached.

We plot the value $\langle S \rangle_\tau / S_e$ as a color code in the heatmaps of Fig. 6.5, over which we also plot the solution B_c. We see that for large values of q_{min}, our solution seems to

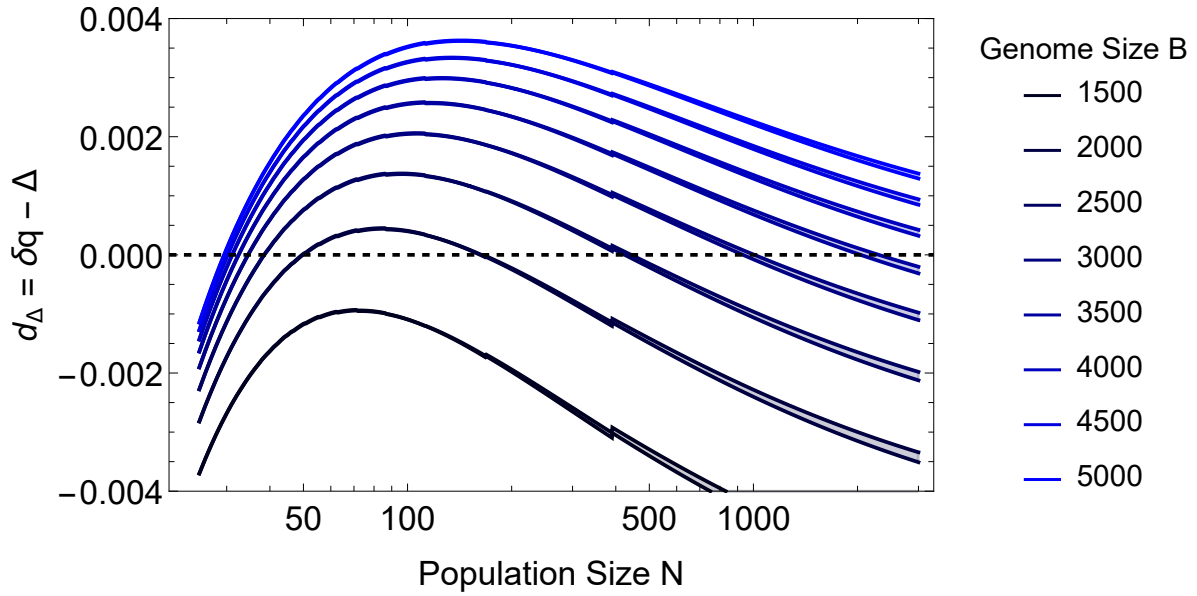


Figure 6.6: **Another way of visualizing the solution.** In the figure, the quantity d_Δ is calculated as a function of N and B. For small B, it is always negative, thus there is no species formation for no value of N. But as B increases, it is possible to see a positive region, in which speciation is possible. The bandwidth in each curve is due to the differences of $\lfloor \tau \rfloor$ (upper value) and $\lceil \tau \rceil$ (lower value). In the figure, the simulation parameters are $\mu = 0.0025$ and $q_{min} = 0.8$.
Source: Figure produced by the author.

agree very well with the transition region, thus being a quite good approximation for the critical genome size. The actual critical genome size, as we defined it, is smaller than the heuristic approximation, since it should be at the *beginning* of the transition region, not over or after it. In this sense, the heuristic approximation seems a good upper bound for the critical genome size. For smaller values of q_{min}, this interpretation still holds, since it overestimates the transition region. Therefore, although approximate, this solution is very useful for computational reasons.

6.2.3 Visualizing the solution

An interesting way to *visualize* the heuristic solution is to analyze the difference $d_\Delta \equiv \delta q - \Delta$ as a function of the population size. Whenever this quantity is positive, the parameters enable species formation. Fig. 6.6 shows this curve as a function of N for different parameters. Suppose a population with a large size N_0 has $d_\Delta(N_0) < 0$. Then, according to the heuristic solution, it would not form different species. However, because the system is stochastic, it may happen that this population breaks into two species, each one with a different size, e.g., N_1 and N_2 ($N_1 + N_2 = N_0$), which are of course smaller than N_0. Because of the shape of the $d_\Delta(N)$ curve, it may happen that $d_\Delta(N_1)$ or $d_\Delta(N_2)$ is positive, meaning that the new species may speciate again, driving the system to the high diversity phase. This example shows how the metastability close to the transition can originate from fluctuations in the population size.

Chapter 7

The high-diversity phase

7.1 Introducing the challenge

So far, we have worked on the development of an analytical description of the transition from the low diversity phase to the high diversity phase of the Derrida-Higgs model. We ran into the problem of having a non-complete description of the genetic similarity distribution in terms of only its values, being necessary to know the higher order overlap values. We then developed a heuristic solution for the transition which works in good agreement with the simulation when the similarity threshold is high. The low diversity phase, before the non-trivial equilibrium (or when $q_{eq} > q_{min}$) is very well understood and we were even able to calculate the evolution of the first moments of the first order overlap distribution. However, nothing has been said concerning the high-diversity phase, and this is the aim of the present chapter.

We introduced the finite genome problem as the question "when does the number of species $S(B, N, \mu, q_{min})$ is greater than 1?" without asking the actual value of $S(B, N, \mu, q_{min})$. In other words, **what is the species richness of the system?** Another interesting biological question concerns the distribution of *species sizes*, i.e., how many species have a given number of individuals (in the network description, it is the same as asking the number of nodes in each component). We then address the question **what is the species abundance distribution of the system?**

7.1.1 Species abundance distribution and species richness

The *abundance* of a species is the number of individuals it has in a given community. The same species can have different abundances in different communities, as a result of different ecological interactions [123]. On the other hand, an interesting ecological pattern whose shape does not seem to change among different communities is the *species abundance distribution* (SAD) which counts the number of species (in a specific community) with a given abundance [124]. This curve (displayed as a histogram) typically shows a large number of rare species and a small number of very abundant ones, and this pattern has been observed in many communities studied so far [124, 125]. A richer variety of

shapes only appears when the distribution is displayed in a log-scale[1] [124].

Because the data in a SAD is not labeled, it is possible to compare SADs coming from different communities even if they do not have species in common, making it a powerful measure in ecology [124]. Theoretically, this curve can be calculated in many different ways [124], from purely statistical models [128] to mechanistic ones [129], all of them showing the same pattern of a few large species and many rare ones. Despite this enormous mathematical toolkit, there is no systematic agreement of a model to different data favoring a description over another [124].

In what follows, we are going to work with the *relative abundance distribution* (RAD), which is the SAD normalized by the number of species, written as $\mathcal{P}_R(n)$, meaning the proportion of species in a community that have abundance n. In the Derrida-Higgs model, the community is easily defined by the whole population at a given instant of time. We are interested in the long-time limit, in which the number of species has reached an equilibrium average value (that is why we neglect time indexes now).

Of course,

$$\sum_{n=1}^{\infty} \mathcal{P}_R(n) = 1, \tag{7.1.1}$$

and calling the number of species $\mathcal{S} = \mathcal{S}(B, \mu, N, q_{min})$, we have

$$S \sum_{n=1}^{\infty} n\mathcal{P}_R(n) = N, \tag{7.1.2}$$

and thus, the equilibrium average *Species Richness* is given by

$$\mathcal{S} = \frac{N}{\langle n \rangle}, \tag{7.1.3}$$

where we defined $\langle n \rangle = \sum_n n\mathcal{P}_R(n)$.

Therefore, if we were able to calculate the Species Abundance Distribution that emerges from the Derrida-Higgs dynamics, we would also be able to calculate the average equilibrium species richness. We will see that our formulation of the problem also leads to the proportion of zero size species $\mathcal{P}_R(n = 0)$, i.e., the proportion of species of the previous generation that went extinct in the next generation.

7.1.2 The equilibrium general rules

The observed equilibrium in the Derrida-Higgs model in the high-diversity phase regards the observed species, i.e., the clusters in the genome space, not the genome values or its pairwise distances. In other words, we neglect the genetic dynamics and focus only on the network structure.

According to the Derrida-Higgs rules, the dynamics of the high-diversity phase is described as follows:

[1]To display species abundance distributions in log-scale is not a trivial procedure. Different binning methods may lead to different histograms which would better be adjusted to different curves [126, 127].

1. At a given time t, there are $\mathcal{S}$ species, of sizes $\{m_1, m_2, \ldots, m_{\mathcal{S}}\}$, with $\sum_i m_i = N$;

2. N individuals are drawn with replacement from the population, and each one generates an offspring. Thus, in the absence of speciation events (i.e., no network component breaks into different components), the probability that a species of size $m \geq 2$ at time t has size n at time $t+1$, is given by a binomial distribution

$$\mathcal{P}(n) = \binom{N}{n} \left(\frac{m}{N}\right)^n \left(1 - \frac{m}{N}\right)^{N-n}. \tag{7.1.4}$$

3. However, the offspring of a given species at time t may belong to different species at time $t+1$. When it happens, we consider that there was a speciation event. The probability of a species of size m having its offspring split into different species is $p_s(m)$, named *probability of speciation*.

   This is a simplification, since it should also depend on the lifetime of a given species. However, at the equilibrium, an effective value that would incorporate this timescale should emerge and this is what we are considering here.

4. Also, when a speciation event occurs, the size n and the number r of new species are given by the probabilities $\rho_n(n|m, \overline{n})$ and $\rho_r(r|m, \overline{n})$, where $\overline{n}$ is the number of offspring of the ancestral species.

5. Since the Derrida-Higgs model evolves through sexual reproduction, species of size 1 go extinct in the next time step.

This set of rules defines an algorithm. Starting with a single species of abundance N, we can recursively apply these rules and check the existence of a stationary abundance distribution. Of course, we must know a priori the probabilities $p_s(m)$, $\rho_n(n|m, \overline{n})$ and $\rho_r(r|m, \overline{n})$, but once known, it is possible to write the above rules as a *Markov chain* process and to find an analytical description of the stationary SAD, which is our goal in the following sections.

7.2 Estimating the probabilities

The probabilities we introduced in the previous section are very important quantities to the dynamics and defining them properly is our goal now.

7.2.1 Numerical investigations

In order to get some information from the simulations, we used the following algorithm.

1. We start with the $N-$dimensional vectors $v_{counts} = \{0, \ldots, 0\}$ and $v_{events} = \{\{\}, \ldots, \{\}\}$.

2. For every species of size m that exists at time t we update $v_{counts}(m) = v_{counts}(m)+1$.

3. If a species of size m breaks into r new species of sizes $\{n_1, \ldots, n_r\}$ we save this information as

$$v_{events}(m) = \text{AppendTo}\left[v_{events}(m), \{n_1, \ldots, n_r\}\right].$$

4. We repeat these steps for the whole simulation after the equilibration time and for many simulations, updating always the same vectors v_{counts} and v_{abund}.

These two vectors contain all the information we need to try to infer the probabilities we are looking for. First, $v_{counts}(m)$ gives the number of times a species of size m has been observed in the simulation. Then, the number of lists contained within every $v_{events}(m)$ counts the number of times a species of size m has undergone speciation. From this, we can infer the probability of speciation as

$$\overline{p}_s(m) = \frac{\text{Lenght}\left[v_{events}(m)\right]}{v_{counts}(m)}. \tag{7.2.1}$$

The distribution of sizes of each list $\{n_1, \ldots, n_r\}$ within $v_{events}(m)$ infers the distribution $\rho_r(r|m,\overline{n})$ where the $\overline{n}$ is simply the sum $n_1 + \ldots + n_r$. Also, the distribution of all n_i within each list infers the distribution $\rho_n(n|m,\overline{n})$.

For instance, suppose that for a given m^* we find $v_{counts}(m^*) = 100$ and $v_{events}(m^*) = \{\{n_1^{(1)}, n_2^{(1)}, n_3^{(1)}\}, \{n_1^{(2)}, n_2^{(2)}\}, \{n_1^{(3)}, n_2^{(3)}\}, \{n_1^{(4)}, n_2^{(4)}, n_3^{(4)}\}, \{n_1^{(5)}, n_2^{(5)}, n_3^{(5)}, n_4^{(5)}\}\}$. Suppose also that in each case, $n_1^{(i)} + \ldots + n_r^{(i)} = \overline{n}$ has the same value. Thus, species of size m^* has undergone speciation 5 times. In this case, the probability of speciation is inferred as $\overline{p}_s(m^*) = 1/20$ and we have $\rho_r(2|m^*,\overline{n}) = \rho_r(3|m^*,\overline{n}) = 2/5$ and $\rho_r(4|m^*,\overline{n}) = 1/5$. The distribution $\rho_n(n|m^*,\overline{n})$ is given by the normalized histogram of

$$\{n_1^{(1)}, n_2^{(1)}, n_3^{(1)}, n_1^{(2)}, n_2^{(2)}, n_1^{(3)}, n_2^{(3)}, n_1^{(4)}, n_2^{(4)}, n_3^{(4)}, n_1^{(5)}, n_2^{(5)}, n_3^{(5)}, n_4^{(5)}\}$$

Since this algorithm does not separate any time scale, it should capture the effective probability of speciation. However, in order to find good inferred values for these probabilities for all values of m in a given range, we must run the simulations for $N \gg m$, once after the radiation period, large species ($m \sim N$) are extremely rare.

Notwithstanding, N^* (Eq.(6.1.2)) is a good scale for the species size in the Derrida-Higgs dynamics, and since it does not depend on the population size, increasing N in order to calculate $\overline{p}_s(m)$ for large m is also not a good approach because the only effect would be to increase the number of species, not their sizes. Thus, although simple, this algorithm does not provide a fast way to extract these probabilities from the simulations.

The numerical results of this algorithm are going to be discussed in the following sections together with the theoretical propositions.

7.2.2 The size and number of new species

When a species of m individuals and $\overline{n}$ offspring breaks into different species, there is a distribution $\rho_r(r|m,\overline{n})$ of how many r new species appear as also a distribution $\rho_n(n|m,\overline{n})$ for their abundances n.

Numerical investigations (left panel of Fig. 7.1) show that the majority of all speciation events results in only two species and thus we consider

$$\rho_r(r|m,\overline{n}) = \delta_{r,2}, \tag{7.2.2}$$

and neglect the other events in which $r > 2$.

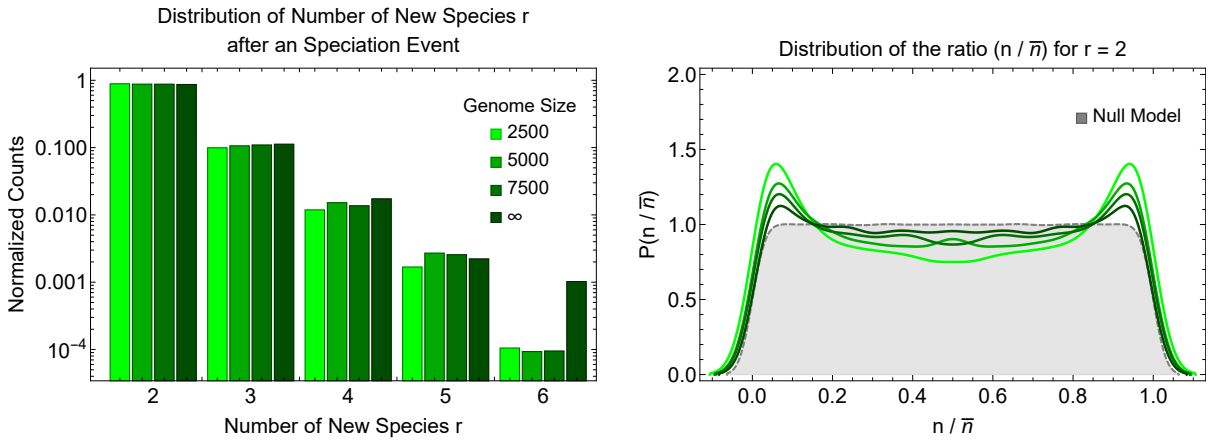


Figure 7.1: **Numerical investigations on speciation events.** The left panel shows the distribution of how many species r appear after a speciation event for different genome sizes. Around 90% of the speciation events result in only 2 new species. When analyzing the sizes of these two species, we find the panel at right, which shows the distribution of $n/\overline{n}$, i.e., the ratio between the sizes of the new species and the number of offspring of the ancestral species. The gray curve is a null model in which n is uniformly chosen from the set $\{1, \ldots, \overline{n}\}$. We see that although there is a small asymmetry in this distribution, it decreases for greater genome values and it can be fairly approximated by the null model. In the figure, the simulation parameters are $N = 400$, $\mu = 0.0025$ and $q_{min} = 0.8$. More information about this data can be found in Appendix A.4.
Source: Figure produced by the author.

Now, since we are going to consider only speciation events with $r = 2$, we look for a probability $\rho_n(n|m, \overline{n})$ that is symmetric around $\overline{n}/2$. With some generality, let us suppose that this probability does not depend on the initial size m of the ancestral species. Again, numerical investigations (right panel of Fig. 7.1) show that it is not a bad approximation to consider a uniform probability distribution, mainly for larger values of genome size,

$$\rho_n(n|m, \overline{n}) = \rho_n(n|\overline{n}) = \frac{1}{\overline{n} - 1}, \tag{7.2.3}$$

where the normalization is given by $\sum_{n=1}^{\overline{n}-1} \rho_n(n|m, \overline{n}) = 1$ (if there is speciation, the final size cannot be exactly the same as the number of offsprings, $\rho_n(n = \overline{n}|m, \overline{n}) = 0$).

7.2.3 The probability of speciation

The probability of speciation $p_s(m)$ of a species of size m is not easy to calculate, and we lack of an expression for it. We are going to propose an expression that lies on the heuristic solution to the transition, but as we see from numerical investigations, it does not describe the true value of $p_s(m)$. On the other hand, this proposition ends in a very interesting relative abundance distribution, since it is not so distant from the RADs observed in the Derrida-Higgs model.

From the heuristic solution, the quantity $d_\Delta = \delta q - \Delta$ (defined in Section 6.2.3) can be a good way of measuring how likely a species is to undergo speciation, and thus one

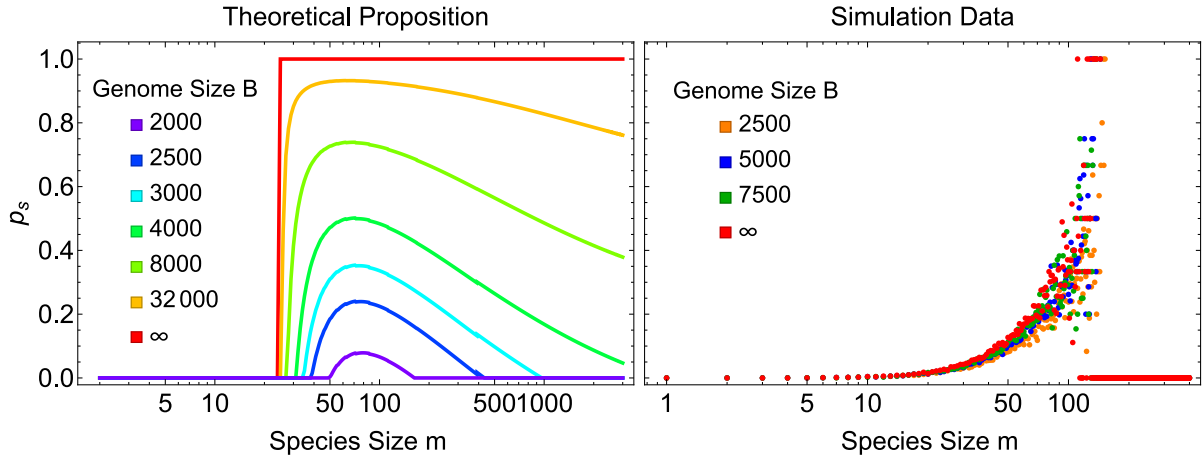


Figure 7.2: **The probability of speciation.** The left panel shows the proposition of equation (7.2.5) for the probability of speciation for different values of genome size. p_s is calculated as the average obtained by using $\lfloor \tau \rfloor$ and $\lceil \tau \rceil$. The right panel shows the probability inferred with the algorithm of Section 7.2.1. In the figure, the simulation parameters are $N = 400$, $\mu = 0.0025$ and $q_{min} = 0.8$. The set of simulations is the same as those from Fig.7.1 and more information can be found in in Appendix A.4.
Source: Figure produced by the author.

may consider

$$p_s(m) \sim \delta q - \Delta, \tag{7.2.4}$$

and once the probability should be maximum when $B \to \infty$ (when $\Delta = 0$),

$$p_s(m) = \max\left(0, 1 - \frac{\Delta}{\delta q}\right), \tag{7.2.5}$$

where δq and Δ are calculated with $N = m$.

Fig. 7.2 shows at left our proposition for $p_s(m)$ as a function of the species size m in comparison with the one inferred from simulations (with the algorithm of section 7.2.1). It is possible to see how different they are, and because of that, we also use the data obtained from simulations to calculate the SAD in this process. In order to do that, we found that a quadratic curve am^2 adjusts very well the results. The coefficient a was found by adjusting the curve over points that are statistically significant. We consider it significant if it was inferred by more than 0.05% of the data. Figure 7.3 displays this procedure, by showing the data cutoff and the inferred probability, which we consider as

$$p_s(m) = \min\left(1, am^2\right). \tag{7.2.6}$$

As we are going to see, the SADs resulting from this expression are in good agreement with the simulations, showing that the assumptions on ρ_r and ρ_n are good assumptions and that the probability of speciation is of greater importance in the present theory.

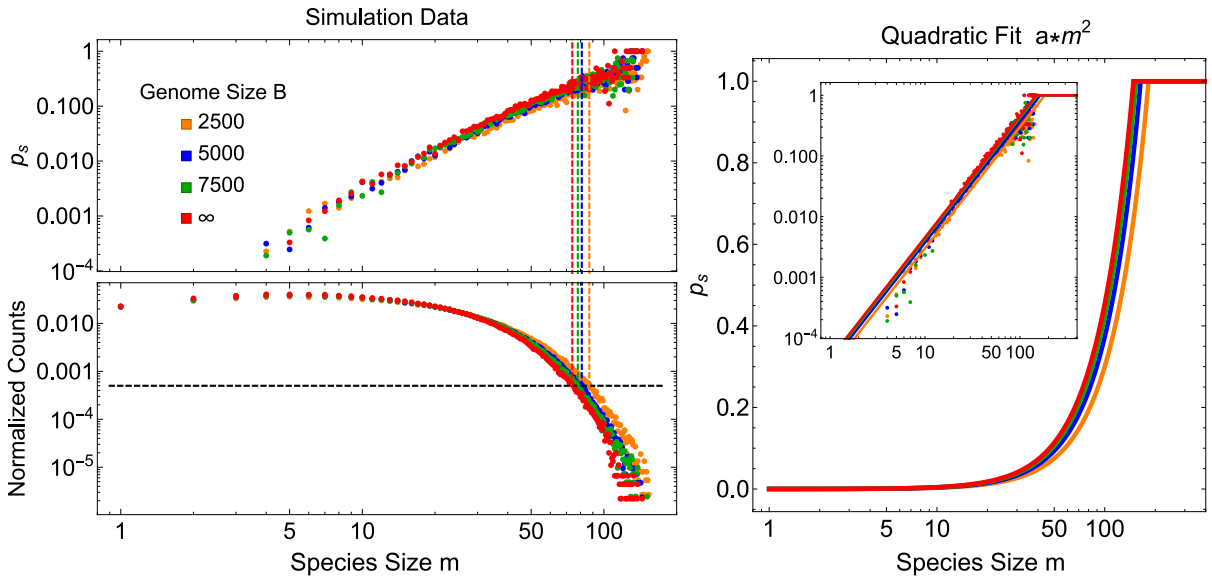


Figure 7.3: **Probability of speciation from simulations.** The upper left panel shows the probability of speciation inferred from simulations in a log-log scale. Below it, the normalized number of data used to inferr the probability. In black, the cutoff of 5×10^{-4} and the vertical dashed lines show the cutoff on the data. The points at left of these lines were used to adjust a quadratic curve, shown in the right panel. In the figure, the simulation parameters are $N = 400$, $\mu = 0.0025$ and $q_{min} = 0.8$. The set of simulations is the same as those from Fig.7.1 and more information can be found in in Appendix A.4. Source: Figure produced by the author.

7.3 A Markov chain model

So far we have described the high-diversity phase in an algorithmic way. We aim now to give an analytical description of this process. As we are going to see, it can be easily described as a Markov chain, since given the Derrida-Higgs rules, the system at time t completely determines the system at time $t + 1$. Although the whole process has proven to be very hard to describe analytically, with the rules of section 7.1.2, it is not hard to describe it if we consider the system as an occupation of abundance states, i.e., the states are the possible abundance sizes of a species and how many species are in each state describes the system.

Let us start with the probability of finding a species of size n at time $t + 1$, which is given by

$$\mathcal{P}_{t+1}(n) = \sum_{m=1}^{N} \mathbb{P}(m \to n)\mathcal{P}_t(m) \tag{7.3.1}$$

where $\mathbb{P}(m \to n)$ is the transition probability from a species of size m to the size n. Our goal now is to calculate the transition probabilities, which can be very tricky due to the boundary conditions.

7.3.1 The transition matrix

The probability of changing from one size to another has two components: the random sampling with replacement and the speciation. We are not considering *hybridization* events, i.e., when individuals from different species evolve in a way of being able to mate again, thus making their species become the same. Although we know hybridizations can occur in the model [130], we shall see that our description is enough for replicating the emerging RADs. Hence,

$$\mathbb{P}(m \to n) = \mathbb{P}(m \to n|\text{no speciation})(1 - p_s(m)) + \mathbb{P}(m \to n|\text{speciation})p_s(m). \quad (7.3.2)$$

When there is no speciation, there is only the random sampling with replacement, which is given by the binomial distribution of Eq.(7.1.4), which is valid only for $2 \leq m \leq N-2$,

$$\mathbb{P}(m \to n|\text{no speciation}) = \binom{N}{n}\left(\frac{m}{N}\right)^n\left(1 - \frac{m}{N}\right)^{N-n}, \quad (7.3.3)$$

because species of size 1 do not reproduce, going extinct in the next time step, which also sets the complementary boundary condition: if a species has size $N-1$, it means that the remaining species are going extinct and then

$$\mathbb{P}(m = N \to n|\text{no speciation}) = \mathbb{P}(m = N-1 \to n|\text{no speciation}) = \delta_{n,N}, \quad (7.3.4)$$

and

$$\mathbb{P}(m = 1 \to n|\text{no speciation}) = \delta_{n,0},$$

which is actually more general

$$\mathbb{P}(m = 1 \to n) = \delta_{n,0} \quad (7.3.5)$$

But in the presence of speciation, the probability is more complicates. According to the dynamics, first, the individuals are sampled from the previous population. Then, the offspring belong to different species. First, N individuals are sampled from the population; $\overline{n}$ are from the specific species one is looking at; then, only n individuals (smaller than $\overline{n}$) compose the next species.

Species with size $m = N-1$ and $m = N$ are responsible for all the offspring in the next generation, so it only changes its size to something smaller than N due to speciation. Thus

$$\mathbb{P}(m = N-1 \to n|\text{speciation}) = \rho_n(n|m = N-1, \overline{n} = N) \quad (7.3.6)$$

and

$$\mathbb{P}(m = N \to n|\text{speciation}) = \rho_n(n|m = N, \overline{n} = N). \quad (7.3.7)$$

Also, in the presence of speciation,

$$\mathbb{P}(m = N-1 \to N|\text{speciation}) = \mathbb{P}(m = N \to N|\text{speciation}) = 0.$$

Actually, the previous equation is more general, also being valid for any value $2 \leq m \leq N-2$ and also for $n = 0$:

$$\mathbb{P}(m \to N|\text{speciation}) = \mathbb{P}(m \to 0|\text{speciation}) = 0, \tag{7.3.8}$$

since if a species goes to zero size, it has been extinct and if it goes to the whole population size, it means that it did not undergo speciation.

For $2 \leq m \leq N-2$ changing to $1 \leq n \leq N-1$, we must consider that $\overline{n} > n$ individuals have been drawn from the focal species and then it changes to size n,

$$\mathbb{P}(m \to n|\text{speciation}) = \sum_{\overline{n}=n+1}^{N} \binom{N-2}{\overline{n}-2} \left(\frac{m}{N}\right)^{\overline{n}-2} \left(1-\frac{m}{N}\right)^{N-\overline{n}} \rho_n(n|m,\overline{n}), \tag{7.3.9}$$

where the -2 is due to the fact that if there is speciation, then at least two individuals of that specific species have already been drawn.

Normalization

Let us show now that the transition matrix defined this way is stochastic, i.e., $\sum_{n=0}^{N} \mathbb{P}(m \to n) = 1$ for every m.

1. For $m = 1$:

   According to equation (7.3.5),

   $$\sum_{n=0}^{N} \mathbb{P}(1 \to n) = \sum_{n=0}^{N} \delta_{n,0} = 1$$

   q.e.d.

2. For $2 \leq m \leq N-2$:

   In this case,

$$\begin{aligned}
\sum_{n=0}^{N} \mathbb{P}(m \to n) =& (1-p_s(m)) \sum_{n=0}^{N} \binom{N}{n} \left(\frac{m}{N}\right)^{n} \left(1-\frac{m}{N}\right)^{N-n} \\
&+p_s(m) \sum_{n=1}^{N-1} \sum_{\overline{n}=n+1}^{N} \binom{N-2}{\overline{n}-2} \left(\frac{m}{N}\right)^{\overline{n}-2} \left(1-\frac{m}{N}\right)^{N-\overline{n}} \rho_n(n|m,\overline{n}) \\
=& (1-p_s(m)) \\
&+p_s(m) \sum_{\overline{n}=2}^{N} \sum_{n=1}^{\overline{n}-1} \binom{N-2}{\overline{n}-2} \left(\frac{m}{N}\right)^{\overline{n}-2} \left(1-\frac{m}{N}\right)^{N-\overline{n}} \rho_n(n|m,\overline{n}) \\
=& (1-p_s(m)) \\
&+p_s(m) \sum_{\overline{n}=2}^{N} \binom{N-2}{\overline{n}-2} \left(\frac{m}{N}\right)^{\overline{n}-2} \left(1-\frac{m}{N}\right)^{N-\overline{n}} \sum_{n=1}^{\overline{n}-1} \rho_n(n|m,\overline{n}) \\
=& (1-p_s(m)) + p_s(m) = 1
\end{aligned}$$

q.e.d.

3. For $m = N - 1$ and $m = N$:

   Now,

$$\begin{aligned}\sum_{n=0}^{N} \mathbb{P}(m \to n) =& (1 - p_s(m)) \sum_{n=0}^{N} \delta_{n,N} + p_s(m) \sum_{n=1}^{N-1} \rho_n(n|N-1, \overline{n} = N) \\ =& (1 - p_s(m)) + p_s(m) = 1\end{aligned}$$

   q.e.d.

4. For $m = 0$:

   This was the only non discussed case so far, and because it is quite subtle, we discuss it in the next section.

The case m=0

It is important to emphasize what this Markov process is describing: a given species of size m_t at time t is changing its size due to a stochastic process which is a combination of random sampling followed by something we call *speciation* (which can also be understood as a random sampling from the first random sample). Thus, the sequence $\{m_t, m_{t+1}, \ldots\}$ may eventually reach size zero (extinction) and stop changing. It means $\mathbb{P}(0 \to n) = \delta_{n,0}$, which obviously normalizes the case $m = 0$. However, this introduces an absorbing state in the system: $\mathcal{P}_{t\to\infty}(n = 0) = 1$ and this does not provide the relative abundance distribution we are looking for.

Since our aim is to obtain the RAD, when the species we are following reaches size zero, we can substitute it for any other species from the system and then start to follow its sequence of sizes. But the species sizes in the system are distributed according to the RAD at that specific time, then

$$\mathbb{P}(m_t = 0 \to n) = \mathcal{P}_t(n) \tag{7.3.10}$$

and since the RAD is normalized, it completes the proof of normalization and the transition matrix is therefore stochastic. However, the problem now is more complex since the transition matrix depends on the current distribution $\mathcal{P}_t(n)$, characterizing a *non-linear Markov chain* [50]. Indeed, the absorbing state $\mathcal{P}(n = 0) = 1$ is a solution of the system, and then we are interested in the non-trivial solution of the equation

$$\begin{aligned}\mathcal{P}(n) &= \sum_{m=0}^{N} \mathbb{P}_\infty(m \to n)\mathcal{P}(m) \\ &= \mathcal{P}(n)\mathcal{P}(0) + \sum_{m=1}^{N} \mathbb{P}(m \to n)\mathcal{P}(m)\end{aligned} \tag{7.3.11}$$

which we can obtain numerically.

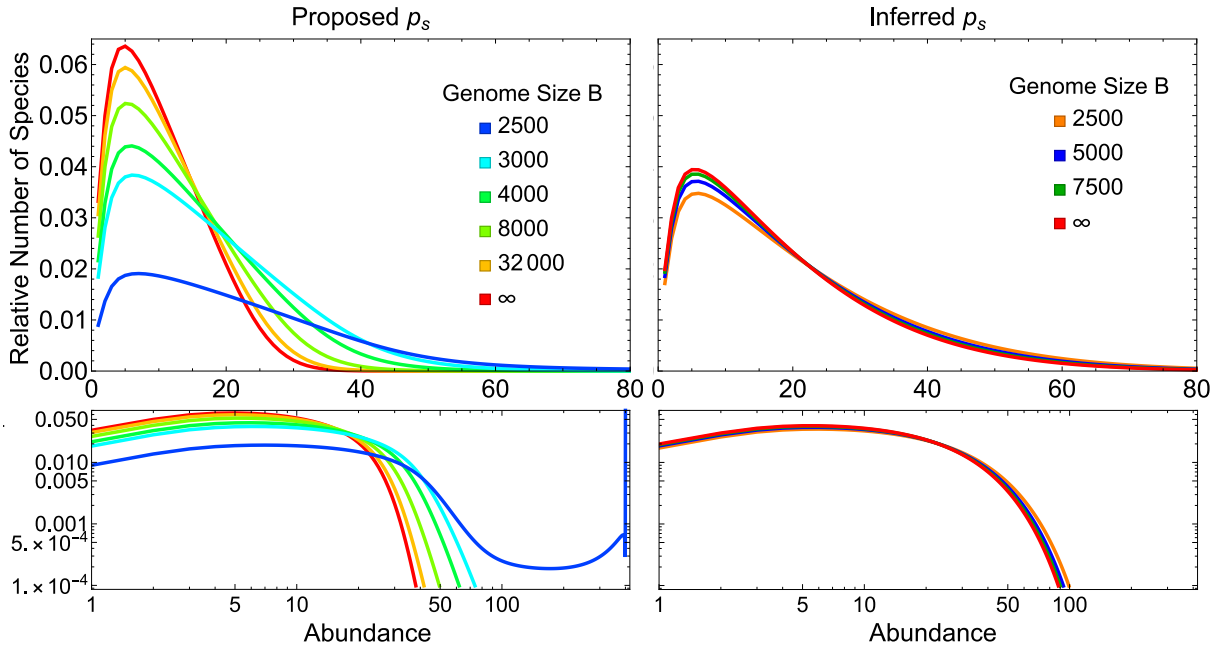


Figure 7.4: **Relative abundance distribution: Numerical results.** The figure shows the RADs resulting from equation (7.3.12) for different values of the genome size B. The left panels shows the results when considering $p_s(m)$ as Eq.(7.2.5) and the right panels the results for $p_s(m)$ as (7.2.6). The bottom panels show the upper ones in log scale. In the figure, the simulation parameters are $N = 400$, $\mu = 0.0025$ and $q_{min} = 0.8$.
Source: Figure produced by the author.

7.3.2 Results

Starting with only one species, $\mathcal{P}_0(N) = 1$, we recursively ran

$$\mathcal{P}_{t+1}(n) = \mathcal{P}_t(n)\mathcal{P}_t(0) + \sum_{m=1}^{N} \mathbb{P}(m \to n)\mathcal{P}_t(m) \tag{7.3.12}$$

for a number T of iterations in which $|\mathcal{P}_T(n) - \mathcal{P}_{T-1}(n)| < \epsilon\mathcal{P}_{T-1}(n)$, for every n where ϵ is a small value (we used $\epsilon = 10^{-8}$). Since we are starting the system with only one species, we considered parameter values such that $B > B_c$. In this way, we can use the proposed probability of speciation (Eq.(7.2.5)). Moreover, we have also inferred the probability of speciation (Eq.(7.2.1)) from some simulations and used it in our Markov chain framework. There are three different results we can analyze: RADs, Richness and Extinction probability ($\mathcal{P}(n = 0)$).

Relative species abundance distribution

Figure 7.4 shows the numerical solutions of the RADs for different values of B, considering the proposed probability of speciation (left panels) as also the inferred probability of speciation (right panels). We observe that when using our proposition, different genome sizes end up in very different RADs, although all the effective supports fall approximately on the same region of the domain for sufficiently large genome sizes, as also all curves

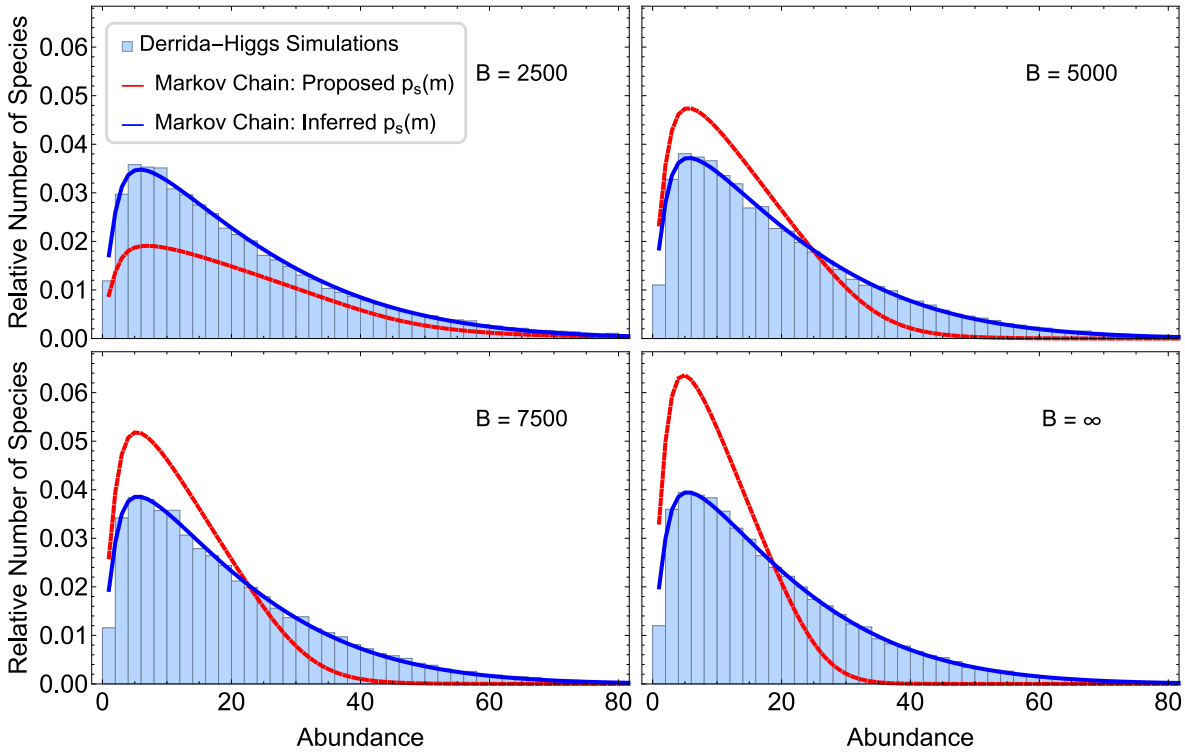


Figure 7.5: **Relative abundance distribution: Comparing the results.** The figure shows the RAD coming from the Derrida-Higgs evolution (background histogram) and the numerical solutions of equation (7.3.12) when considering the proposed probability of speciation (red line) and the inferred probability of speciation (blue line). Different panels show the results for different genome sizes. In the figure, the simulation parameters are $N = 400$, $\mu = 0.0025$ and $q_{min} = 0.8$. For the histogram, the set of simulations is the same as those from Fig.7.1 and more information can be found in Appendix A.4.
Source: Figure produced by the author.

being unimodal and displaying the same RAD well-known feature: many rare species and a few common species. When B is close to the transition value B_c (like $B = 2500$, as it can be seen from Fig. 7.2), the bahavior of the RAD can be a bit different, finding equilibrium with also a peak close to the population size N (as it can be seen from in the bottom left panel of Fig.7.4).

On the other hand, the RADs resulting from the inferred probability of speciation are much more close to each other, slightly increasing their peak value as the genome size increases. The overall behaviors of all these observed RADs directly reflect the variety of probability of speciation curves displayed in Fig.7.2.

Although consistent, our proposition for $p_s(m)$ (Eq.(7.2.5)) is motivated by the heuristic solution for the transition. It is not concerned with what happens to small species within a bigger community. Thus, the probability of a population of size N to undergo speciation does not need to be the same of a species of size $m < N$ to undergo speciation within a larger community.

Figure 7.5 compares the numerically calculated RADs with the ones found in the Derrida-Higgs dynamics for different genome sizes. The background histogram comes from the Derrida-Higgs evolution. The red curve is the numerically calculated RAD considering the proposed probability of speciation and the blue curve is the RAD when

considering the inferred probability of speciation. We observe a good agreement between the data and the RAD when using the inferred probability of speciation. It corroborates our Markov chain model for the high diversity regime, showing an important dependence on the probability of speciation.

On the other hand, in the absence of any data, our proposition for $p_s(m)$ is not a bad starting point, being able to give important insights about the real Derrida-Higgs RAD, as an estimative of its effective support and the distribution mode.

Species richness and extinction probability

The species richness $\mathcal{S}$ is calculated with equation (7.1.3) with the RAD given by the equilibrium of Eq. (7.3.12) renormalized to one after the probability $\mathcal{P}(n=0)$ is dropped out, since extinct species are non-observed. The right panel of Fig. 7.7 shows the results obtained for richness. The joined dark dots are the results of our proposition for $p_s(m)$, the opened diamonds are the results of the Markov chain with the inferred probability of speciation and the opened circles are the results obtained from the Derrida-Higgs dynamics. Again, the results with the inferred $p_s(m)$ are in remarkable agreement with the dynamics, while the richness can be overestimated when we consider our proposition for $p_s(m)$.

In order to infer the probability of extinction from the Derrida-Higgs dynamics, we consider the following procedure. At a given time t, there are $\mathcal{S}_t$ species. Then, the average number of species that will be extinct is given by $E_t = \mathcal{P}(0)\mathcal{S}_t$, with $\mathcal{P}(0)$ the probability of extinction, which is assumed to be a constant. Thus,

$$\sum_t \frac{E_t}{\mathcal{S}_t} \sim \mathcal{P}(0)t, \tag{7.3.13}$$

and the angular coefficient of the line $\sum_t E_t/\mathcal{S}_t$ is the probability we are looking for. This value can be obtained numerically with linear regression, considering only the generations after the equilibration time. We take the average value of many simulations. Fig.7.6 shows the obtained lines for a set of 50 simulations with the same parameters.

The results for $\mathcal{P}(0)$ are displayed in the left panel of Fig.7.7. The opened circles are the results from the Derrida-Higgs model. The joined dark dots are the results from the Markov chain considering our proposition for $p_s(m)$ and the opened diamonds are the results considering the inferred $p_s(m)$. In this case, we also find a much better agreement of the data with the Markov chain with the inferred probability of speciation. However, now the agreement is not as good as for the richness results. A possible reason can be due to the necessity of a better adjustment to the curve $p_s(m)$, a fitting able to get with a good precision a single element of the vector $\mathcal{P}(n)$. Notwithstanding, this is still a remarkable result.

7.3.3 Concluding remarks

Species Abundance Distributions are important descriptors of real communities [124] and in the Derrida-Higgs model they are an important outcome, reaching stationarity despite the constant genetic evolution and species turnover dynamics. A unified theory for the entire Derrida-Higgs model is still a challenge and finding equations from which

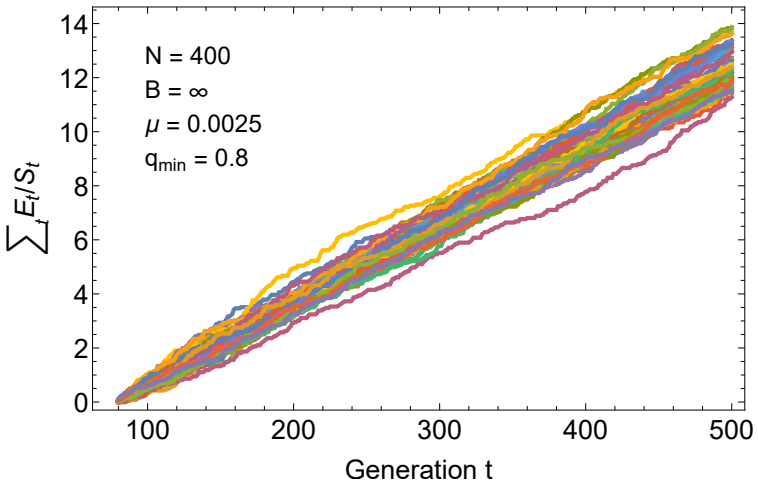


Figure 7.6: **Calculating the probability of extinction.** The figure shows the curve $\sum_t E_t/\mathcal{S}_t$ for 50 different simulations of the Derrida-Higgs model (each color represents a different run) for the same set of parameters. The angular coefficient of each line is an estimative of the probability of extinction per generation and number of species. We take the average of all angular coefficients to infer the real value of $\mathcal{P}(0)$.
Source: Figure produced by the author.

we are able to observe the Species Abundance Distribution as an emergent result would constitute a great achievement. On the other hand, the lack of a complete analytical description of this process did not prevent us from introducing a quite accurate and simple mathematical framework to deal with the high-diversity regime.

Using a Markov chain to model this phase is a simple but straightforward strategy and has shown to be very effective, despite its non-linearity. We did not work on its analytical properties, which can by itself turn out to be a challenge, hence we have focused on a simple mathematical investigation, following the same methodology as we have used for studying the computational Derrida-Higgs model, i.e., starting always with a single species with abundance equaling the carrying capacity (the population size N).

Constructing Markov chains introduces the task of calculating the transition matrix, which in this case has non-trivial boundary conditions due to sexual reproduction. The probabilities involved in the speciation process are also important inputs of this matrix and we investigated it carefully. The number of new species after a speciation event as also the distribution of their sizes have shown to be simple to model and their approximations did not seem to have greatly affected the results. However, the probability of speciation is still a gap in the theory. Our first attempt for $p_s(m)$ (Eq.(7.2.5)) ended up being very different from the values inferred from the data (Eq.(7.2.6)), although the resulting RAD also shows some reasonable properties, like the mode value and the effective support. Notwithstanding, when we used the inferred probability of speciation in the transition matrix, the results were remarkably good.

In this way, we finish our analysis of the Derrida-Higgs model. The use of probability theory combined with a careful computational investigation ended up in a new and useful theory for this very interesting population dynamics model. The following chapter works on an extension of the Derrida-Higgs model in order to incorporate the coevolution of a second genetic material. We also apply the techniques we developed in the first chapter of this part of the thesis to try to achieve analytical results.

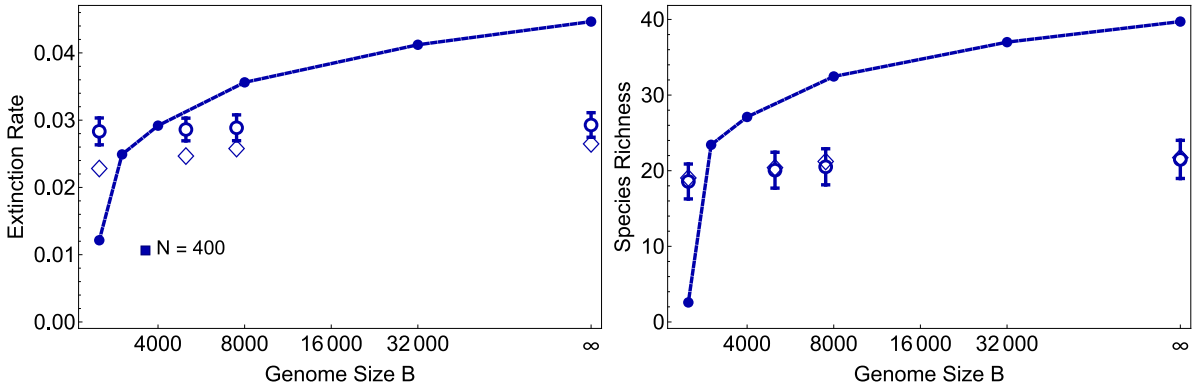


Figure 7.7: **Results on richness and extinction probability.** The left panel shows the results on the extinction rate and the right panel shows the results on species richness for different values of genome size. The opened circles are the results obtained from the Derrida-Higgs dynamics, the error bars are the standard deviation of all the data collected after the equilibration time for 50 different simulations. The joined dark points are the results of the Markov chain when the probability of speciation is given by our proposition (Eq.(7.2.5)), while the opened diamonds, the results when $p_s(m)$ is given by the inferred probability speciation (Eq.(7.2.6)). In the figure, the simulation parameters are $\mu = 0.0025$ and $q_{min} = 0.8$. For the Derrida-Higgs results, the set of simulations is the same as those from Fig.7.1 and more information can be found in in Appendix A.4. Source: Figure produced by the author.

Chapter 8

The mito-nuclear DNA interaction model

In eukaryotic individuals, cells pose a specific organelle responsible for cell respiration, the *mitochondrion* [131]. This organelle differs from the others while having its own genetic material, the mitochondrial DNA (mDNA), which contains many genes used during the cell respiration process [131, 119, 132]. The mDNA is much smaller than the nuclear DNA (nDNA) (which also has genes for cell respiration) and, while being non-recombinant (it is of maternal inheritance), it has a higher mutation rate than the nDNA [119].

The coevolution between the nDNA and the mDNA is essential to the cell, since the different evolutionary rates between these two genetic materials can lead to non compatibilities among its genes directed to cell respiration, making this process inefficient. Thus, mutations in the mDNA must be compensated by changes in the nDNA [132].

The mDNA is also used for species identification [119] and phylogeny reconstruction [133], since it is possible to recognize a species by a standard sequence fragment of the mDNA [119, 132]. This is known as the barcode property of the mitochondrial DNA. The origin of this property was investigated in a model proposed by Princepe and Aguiar in 2021 [120]. They used a spatial version of the Derrida-Higgs model (previously introduced by Aguiar [134]) as a framework and included a binary sequence posing the same properties as the mDNA. They concluded that the barcode property emerges as a consequence of the space structure and conjectured that it would disappear for sympatric communities.

As we have developed a theory for the (sympatric) Derrida-Higgs model, we now apply the same techniques to the sympatric version of the Princepe-Aguiar model of mito-nuclear coevolution. We try to find analytical properties of the system and also to recover numerical results on the barcode property, thus corroborating their conclusions. In what follows, we present the model and its details, followed by the analytical theory and numerical results.

8.1 The model

In order to model the coevolution between the mitochondrial and nuclear genetic materials, and study its effects, Princepe and Aguiar included a second binary string to represent the mitochondrial DNA [120]. Since the mDNA is of maternal inheritance,

individuals are divided into males and females, being non hermaphrodites now. The evolution now is non-neutral, and the individual fitness is calculated according to a measure of similarity between the genetic sequences.

8.1.1 The dynamics

Now, the population is described by N individuals, whose sexes are randomly assigned. An individual α has a nuclear genetic material $\mathbf{S}^\alpha = \{n_1^\alpha, \ldots, n_B^\alpha\}$ and also a *mitochondrial genetic material* $\mathbf{S}_M^\alpha = \{m_1^\alpha, \ldots, m_{B_M}^\alpha\}$, with $n_i^\alpha = \pm 1$ and $m_i^\alpha = \pm 1$ and $B_M \leq B$. We define the *nuclear genetic distance* between individuals α and β as

$$d_N^{\alpha\beta} = \sum_{k=1}^{B} \frac{1}{2} |n_k^\alpha - n_k^\beta|, \tag{8.1.1}$$

and the *mitochondrial genetic distance* as

$$d_M^{\alpha\beta} = \sum_{k=1}^{B_M} \frac{1}{2} |m_k^\alpha - m_k^\beta|, \tag{8.1.2}$$

notwithstanding, we also define the *mito-nuclear genetic distance* as the fraction of distinct alleles between the mitochondrial and nuclear DNA

$$d_{MN}^{\alpha\beta} = \frac{1}{B_M} \sum_{k=1}^{B_M} \frac{1}{2} |m_k^\alpha - n_k^\beta|, \tag{8.1.3}$$

in such a way that an individual α with a full correspondence between its genetic materials has $d_{MN}^{\alpha\alpha} \equiv d_{MN}^{\alpha} = 0$.

Since the mitochondrial and nuclear DNAs should *match* in order to maximize cellular respiration, we define the *fitness* of the individual α as

$$\omega^\alpha = \exp\left[-(d_{MN}^\alpha)^2 / 2\sigma_\omega^2\right], \tag{8.1.4}$$

being σ_ω the coupling strength parameter between both genetic materials. The smallest σ_ω is, the greater is the coupling strength.

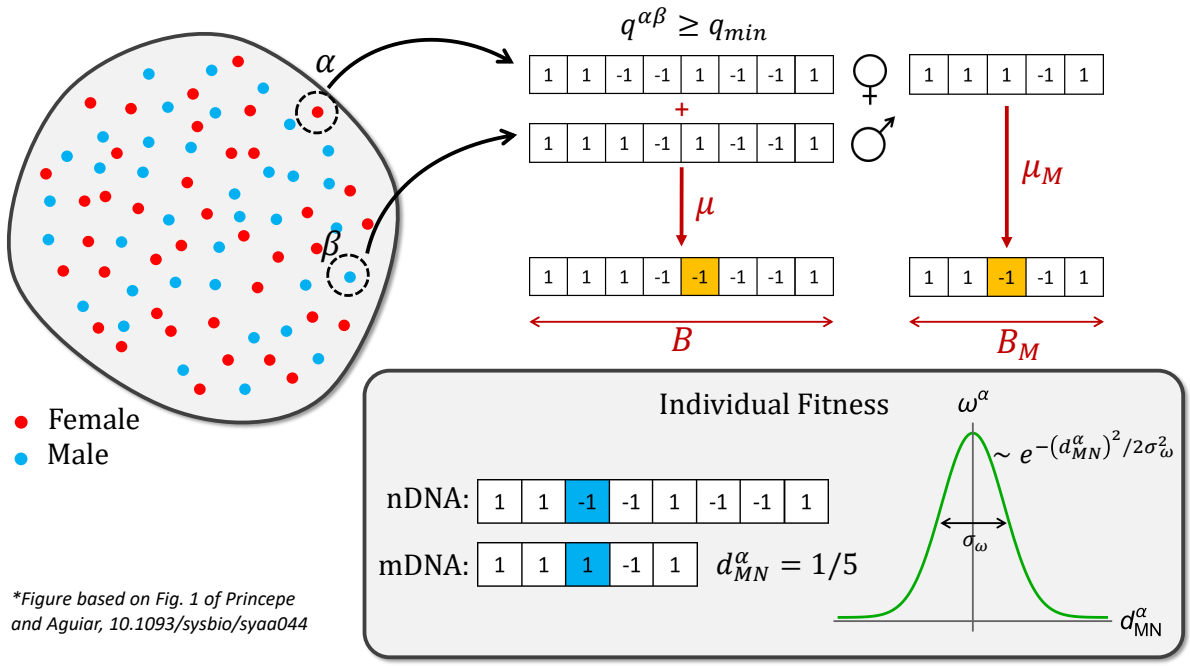


Figure 8.1: **The sympatric Princepe-Aguiar model.** The figure shows the model structure. At time t, a population of N individuals is divided in males (blue points) and females (red points). A focal individual is chosen according to its fitness and its mating pair is chosen such that their sexes are different and $q^{\alpha\beta} \geq q_{min}$. Their nuclear genomes (of size B) are combined to generate the nuclear genome of the offspring, but the mitochondrial genome (of size B_M) is passed directly from the female parent to the offspring. Different mutation rates act on each genome. The individual fitness ω^{α} is calculated according to the mito-nuclear distance d^{α}_{MN} of an individual α, calculated as the fraction of different alleles between their genomes (in the first B_M entries of the nuclear DNA). This figure is based on Fig. 1 of Princepe and Aguiar [120].
Source: Figure produced by the author.

Two individuals α and β can mate if and only if their genomes differ by at most G alleles, i.e., mating is possible only when $d^{\alpha\beta} \leq G$. This condition is equivalent to that in the Derrida-Higgs involving the genetic similarity, we only need to use the relation $q^{\alpha\beta} = 1 - 2d^{\alpha\beta}/B$ (with $q_{min} = 1 - 2G/B$). Also, compatibility now is also defined by sexual differences: males (females) are only compatible with females (males). The mating pairs are drawn according to their fitness. The focal individual is chosen with a probability proportional to its fitness, $\mathcal{P}(\alpha \text{ is focal}) \sim \omega^{\alpha}$, and then, as in the Derrida-Higgs model, its partner β is chosen from the subset $\mathbb{N}_{\alpha}$ of individuals that can mate with α also according to its fitness $\mathcal{P}(\beta \text{ is the partner of } \alpha) \sim A_{\alpha\beta}\omega^{\beta}$, with $A_{\alpha\beta}$ the adjacency matrix element of the underlying network.

As in the Derrida-Higgs model, the offspring γ has its nDNA $\mathbf{S}^{\gamma}$ as a combination of its parents nDNA, i.e., n_i^{γ} equals n_i^{α} or n_i^{β} with the same probability 1/2. Then, every allele can flip its value with rate μ_N. However, the mDNA is of maternal inheritance, and then it is entirely copied from the female parent to the offspring: $m_i^{\gamma} = m_i^{\text{Female Parent}}$. As for the nDNA, every allele of the mDNA can flip with a mutation rate μ_M. The sex of the offspring is randomly assigned with the same probability for male or female.

The process of choosing mating pairs and generating an offspring is then repeated N

times and an entirely new population at time $t+1$ is generated from the population at time t, without generational overlap. The dynamics is repeated this way. Fig.8.1 summarizes the model.

8.1.2 Identifying species and barcode

In the absence of the mDNA (and the fitness attributed to its coupling with the nDNA), the model follows exactly as the Derrida-Higgs model. Species (called now as *nuclear species*) are still identified as components of the underlying network: reproductively isolated groups of individuals. What we aim now is to look for an association between the species defined by the genetic nuclear distance and their mDNA. Thus, we define *mitochondrial species* by constructing a sequence of adjacency matrix with elements

$$B_{\alpha\beta}^{(d)} = U\left(d - d_M^{\alpha\beta}\right), \tag{8.1.5}$$

where $U(x) = 1$ if $x \geq 0$ and $U(x) = 0$ otherwise and $d = B_M, B_M - 1, B_M - 2, \ldots$ is a mitochondrial distance threshold, a parameter that can vary in order to maximize the agreement between *nuclear* and *mitochondrial* species.

Hence, for a given mitochondrial distance matrix (of elements $d_M^{\alpha\beta}$), as d decreases, the number of components of the corresponding network increases. When this number equals at least the number of nuclear species, we set the mitochondrial distance threshold $G_M \equiv d$ and then define these components as the mitochondrial species.

If all individuals belonging to a given nuclear species belong to the same mitochondrial species, and this mitochondrial species is not found in any other nuclear species, then there is a *species bijection*. We define the ratio of species bijection to the number of mitochondrial species as the *barcode success*. If this fraction is close to 1, then we may say that the barcode property of the mitochondrial DNA has emerged in the system. Fig.8.2 shows this numerical process of reducing the value d in order to find G_M ($q_{min}^M = 1 - 2G_M/B_M$) and to analyze the barcode success.

8.2 The analytical theory

As we have done for the Derrida-Higgs model, we are going to analyze the evolution of the similarity distributions in this model, also following the calculation path of Fig.5.5. Now, there is the dependence of the results on the fitness distribution, as we are going to explicitly see in the analytical results. Although the model was introduced considering the Hamming distances between the genomes, working with the similarity quantities is more suitable, since they are normalized, which provides a better comparison among different parameter sets. We must be careful only with the mito-nuclear distance $d_{MN}^{\alpha\beta}$ and the corresponding similarity, $c^{\alpha\beta}$. Because $d_{MN}^{\alpha\beta}$ is defined as a normalized quantity (Eq.(8.1.3)), the relation between them is given by

$$c^{\alpha\beta} = 1 - 2d_{MN}^{\alpha\beta}, \tag{8.2.1}$$

.

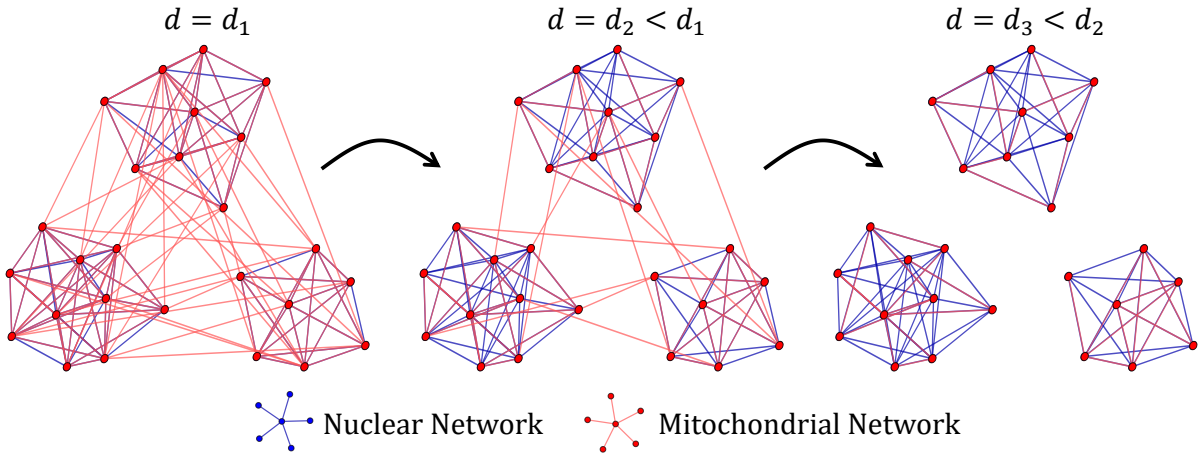


Figure 8.2: **Mitochondrial and nuclear species.** The figure shows the population network with three nuclear species (three components in the blue network). The red network shows the mitochondrial networks for different values of the threshold d. When the mitochondrial distance threshold d is reduced, the number of connections in the red network also reduces, until the moment we recognize the same number of mitochondrial species (components in the red network) as the number of nuclear species. If there is a bijection between the mitochondrial species and the nuclear species (as in the figure for $d = d_3$), then the barcode success is equal to 1.
Source: Figure produced by the author.

Fig.8.3 shows some results of simulations with different coupling coefficients σ_ω. When $\sigma_\omega \to \infty$, the nuclear similarity evolution recovers the sympatric Derrida-Higgs model.

8.2.1 The nuclear similarity distribution

Let us start with the nuclear similarity distribution. We keep the same notation we have used so far and notice that on what concerns the nDNA, the only difference regards how to choose the mating pairs. The probability of drawing individual p_1 to be focal now is given by

$$\mathbb{P}(p_1 \text{ as focal}) = \rho_{p_1} \equiv \frac{\omega^{p_1}}{Z} = \frac{1}{Z} e^{-(d_{MN}^{p_1})^2/2\sigma_\omega^2}, \tag{8.2.2}$$

where $Z = \sum_i \omega^i$ is the normalization. And after choosing p_1, p_2 is chosen with probability

$$\mathbb{P}(p_2|p_1 \text{ is focal}) = \frac{A_{p_1p_2}\omega^{p_2}}{Z_{p_1}} = \frac{A_{p_1p_2}}{Z_{p_1}} e^{-(d_{MN}^{p_1})^2/2\sigma_\omega^2}, \tag{8.2.3}$$

where $Z_{p_1} = \sum_{p_2} A_{p_1p_2}\omega^{p_2}$ is the normalization now. Thus, it is not hard to see that the probability distribution for a given nuclear similarity at time $t+1$, $\mathcal{P}(q_{t+1}^{\alpha\beta})$ is given by

$$\mathcal{P}(q_{t+1}^{\alpha\beta}) = \sum_{\mathbf{S}_{t+1}^\alpha, \mathbf{S}_{t+1}^\beta} \sum_{p_1,p_2} \sum_{p_1',p_2'} \delta\left(q_{t+1}^{\alpha\beta}, \mathbf{S}_{t+1}^\alpha \cdot \mathbf{S}_{t+1}^\beta / B\right) \frac{\omega^{p_1} A_{p_1p_2}\omega^{p_2}}{Z Z_{p_1}} \frac{\omega^{p_1'} A_{p_1'p_2'}\omega^{p_2'}}{Z Z_{p_1'}}$$

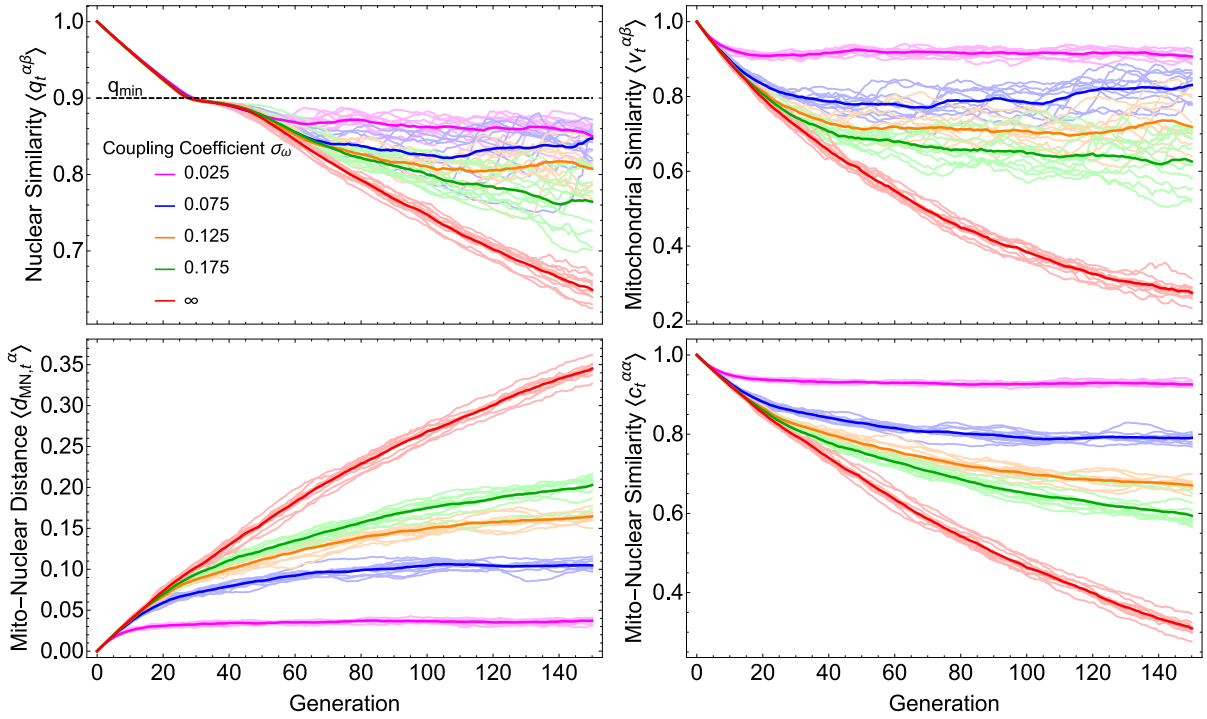


Figure 8.3: **Mitochondrial and nuclear similarities evolution.** In the figure, the average similarity (darker curves) over 10 simulations (lighter curves) with different coupling coefficients is presented. The top left panel shows the evolution of the nuclear similarity $q^{\alpha\beta}$; the top right, the mitochondrial similarity $\nu^{\alpha\beta}$; the bottom left, the mito-nuclear distance d_{MN}^{α} and its corresponding similarity at the bottom right c^{α}. In the figure, the simulation parameters are $N = 500$, $\mu = 0.001$, $q_{min} = 0.9$, $B = 7500$, $B_M = 250$, $\mu_M = 0.003$.
Source: Figure produced by the author.

$$\times \prod_{i=1}^{B} \left[\frac{1}{2} + \frac{n_{i,t+1}^{\alpha} e^{-2\mu}}{4} (n_{i,t}^{p_1} + n_{i,t}^{p_2}) \right] \left[\frac{1}{2} + \frac{n_{i,t+1}^{\alpha} e^{-2\mu}}{4} (n_{i,t}^{p_1'} + n_{i,t}^{p_2'}) \right], \tag{8.2.4}$$

where instead of $1/NN_{p_1}$ appears the term $\omega^{p_1}\omega^{p_2}/ZZ_{p_1}$ and instead of $1/NN_{p_1'}$, the term $\omega^{p_1'}\omega^{p_2'}/ZZ_{p_1'}$.

Thus, for the first moment, we find

$$\mathbb{E}(q_{t+1}^{\alpha\beta}) = \frac{e^{-4\mu}}{4Z^2} \sum_{p_1 p_2} \sum_{p_1' p_2'} \frac{\omega^{p_1} A_{p_1 p_2} \omega^{p_2}}{Z_{p_1}} \frac{\omega^{p_1'} A_{p_1' p_2'} \omega^{p_2'}}{Z_{p_1'}} \left(q_t^{p_1 p_1'} + q_t^{p_2 p_1'} + q_t^{p_1 p_2'} + q_t^{p_2 p_2'} \right). \tag{8.2.5}$$

In the absence of the nuclear similarity threshold q_{min}, we can rearrange the indexes and find

$$\mathbb{E}(q_{t+1}^{\alpha\beta}) = e^{-4\mu} \sum_{p_1 p_1'} \frac{\omega^{p_1} \omega^{p_1'}}{Z^2} q_t^{p_1 p_1'}$$

$$
\begin{aligned}
&= e^{-4\mu} \sum_{p_1 p_1'} \rho_{p_1} \rho_{p_1'} q_t^{p_1 p_1'} \\
&= e^{-4\mu} \sum_{p_1} \left[\rho_{p_1}^2 + \sum_{p_1' \neq p_1} \rho_{p_1} \rho_{p_1'} q_t^{p_1 p_1'} \right] \\
&= e^{-4\mu} N \left[\langle \rho_{p_1}^2 \rangle_P + (N-1) \langle \rho_{p_1} \rho_{p_1'} q_t^{p_1 p_1'} \rangle_P \right]. \qquad (8.2.6)
\end{aligned}
$$

8.2.2 The mitochondrial similarity distribution

For the mitochondrial similarity $\nu^{\alpha\beta}$,

$$
\nu^{\alpha\beta} = \frac{1}{B_M} \sum_{i=1}^{B_M} m_i^\alpha m_i^\beta, \qquad (8.2.7)
$$

we must remember that the mDNA has maternal inheritance, hence, let us consider that the parent p_1 of α and the parent p_1' of β are the mothers. Thus, given the mother p_1, the probability of the allele i of the mDNA of α of being equal to m_i^α is

$$
\mathbb{P}(m_i^\alpha = \pm m_i^{p_1}) = \frac{1}{2} \left[1 + e^{-2\mu_M} m_i^\alpha m_i^{p_1} \right], \qquad (8.2.8)
$$

and then

$$
\begin{aligned}
\mathcal{P}(\nu_{t+1}^{\alpha\beta}) = &\sum_{\mathbf{S}_{M,t+1}^\alpha, \mathbf{S}_{M,t+1}^\beta} \sum_{p_1,p_2} \sum_{p_1',p_2'} \delta \left(\nu_{t+1}^{\alpha\beta}, \mathbf{S}_{M,t+1}^\alpha \cdot \mathbf{S}_{M,t+1}^\beta / B_M \right) \\
&\times \frac{\omega^{p_1} A_{p_1 p_2} \omega^{p_2}}{Z} \left(\frac{1}{Z_{p_1}} + \frac{1}{Z_{p_2}} \right) \frac{\omega^{p_1'} A_{p_1' p_2'} \omega^{p_2'}}{Z} \left(\frac{1}{Z_{p_1'}} + \frac{1}{Z_{p_2'}} \right) \\
&\times \prod_{i=1}^{B_M} \frac{1}{2} \left[1 + e^{-2\mu_M} m_i^\alpha m_i^{p_1} \right] \frac{1}{2} \left[1 + e^{-2\mu_M} m_i^\beta m_i^{p_1'} \right], \qquad (8.2.9)
\end{aligned}
$$

and we emphasize that the indexes p_1 and p_1' run over the females only, while p_2 and p_2' run over the males. Because of this difference regarding the indexes, we cannot manipulate this expression as we did for equation (5.3.23).

We shall now calculate the expected value $\mathbb{E}(\nu_{t+1}^{\alpha\beta})$

$$
\begin{aligned}
\mathbb{E}(\nu_{t+1}^{\alpha\beta}) &= \sum_\nu \nu \mathcal{P}(\nu_{t+1}^{\alpha\beta} = \nu) \\
&= \sum_{p_1,p_2} \sum_{p_1',p_2'} \frac{\omega^{p_1} A_{p_1 p_2} \omega^{p_2}}{Z} \left(\frac{1}{Z_{p_1}} + \frac{1}{Z_{p_2}} \right) \frac{\omega^{p_1'} A_{p_1' p_2'} \omega^{p_2'}}{Z} \left(\frac{1}{Z_{p_1'}} + \frac{1}{Z_{p_2'}} \right) \\
&\times \frac{1}{B_M} \sum_{\mathbf{S}_M^\alpha, \mathbf{S}_M^\beta} \sum_{j=1}^{B_M} m_j^\alpha m_j^\beta \prod_{i=1}^{B_M} \frac{1}{2} \left[1 + e^{-2\mu_M} m_i^\alpha m_i^{p_1} \right] \frac{1}{2} \left[1 + e^{-2\mu_M} m_i^\beta m_i^{p_1'} \right] \\
&= \sum_{p_1,p_2} \sum_{p_1',p_2'} \frac{\omega^{p_1} A_{p_1 p_2} \omega^{p_2}}{Z} \left(\frac{1}{Z_{p_1}} + \frac{1}{Z_{p_2}} \right) \frac{\omega^{p_1'} A_{p_1' p_2'} \omega^{p_2'}}{Z} \left(\frac{1}{Z_{p_1'}} + \frac{1}{Z_{p_2'}} \right)
\end{aligned}
$$

$$\times \frac{1}{B_M}\sum_{j=1}^{B_M}\left(\sum_{\mathbf{S}_M^\alpha} m_j^\alpha \prod_{i=1}^{B_M}\frac{1}{2}\left[1+e^{-2\mu_M} m_i^\alpha m_i^{p_1}\right]\right)\left(\sum_{\mathbf{S}_M^\beta} m_j^\beta \prod_{i=1}^{B_M}\frac{1}{2}\left[1+e^{-2\mu_M} m_i^\beta m_i^{p_1'}\right]\right). \tag{8.2.10}$$

The last two terms in parenthesis can be calculated as follows. Defining

$$G_i \equiv \frac{1}{2}\left[1+e^{-2\mu_M} m_i^\alpha m_i^{p_1}\right], \tag{8.2.11}$$

we can write

$$\sum_{\mathbf{S}_M^\alpha} m_j^\alpha \prod_{i=1}^{B_M}\frac{1}{2}\left[1+e^{-2\mu_M} m_i^\alpha m_i^{p_1}\right] = \left(\sum_{m_1^\alpha=\pm 1} G_1\right)\cdots\left(\sum_{m_j^\alpha=\pm 1} m_j^\alpha G_j\right)\cdots\left(\sum_{m_{B_M}^\alpha=\pm 1} G_{B_M}\right), \tag{8.2.12}$$

and similar for β. Now, calculating each sum

$$\begin{aligned}\sum_{m_i^\alpha=\pm 1} G_i &= 1,\\ \sum_{m_i^\alpha=\pm 1} m_i^\alpha G_i &= e^{-2\mu_M} m_i^{p_1}.\end{aligned} \tag{8.2.13}$$

Hence,

$$\sum_{\mathbf{S}_M^\alpha} m_j^\alpha \prod_{i=1}^{B_M}\frac{1}{2}\left[1+e^{-2\mu_M} m_i^\alpha m_i^{p_1}\right] = e^{-2\mu_M} m_i^{p_1}. \tag{8.2.14}$$

Thus,

$$\begin{aligned}\mathbb{E}(\nu_{t+1}^{\alpha\beta}) &= \sum_{p_1,p_2}\sum_{p_1',p_2'}\frac{\omega^{p_1}A_{p_1p_2}\omega^{p_2}}{Z}\left(\frac{1}{Z_{p_1}}+\frac{1}{Z_{p_2}}\right)\frac{\omega^{p_1'}A_{p_1'p_2'}\omega^{p_2'}}{Z}\left(\frac{1}{Z_{p_1'}}+\frac{1}{Z_{p_2'}}\right)\\ &\times\frac{1}{B_M}\sum_{j=1}^{B_M} e^{-4\mu_M} m_j^{p_1} m_j^{p_1'}\\ &= e^{-4\mu_M}\sum_{p_1,p_2}\sum_{p_1',p_2'}\frac{\omega^{p_1}A_{p_1p_2}\omega^{p_2}}{Z}\left(\frac{1}{Z_{p_1}}+\frac{1}{Z_{p_2}}\right)\frac{\omega^{p_1'}A_{p_1'p_2'}\omega^{p_2'}}{Z}\left(\frac{1}{Z_{p_1'}}+\frac{1}{Z_{p_2'}}\right)\nu_t^{p_1p_1'}\end{aligned} \tag{8.2.15}$$

$$= e^{-4\mu_M}\sum_{p_1,p_1'}\rho_{p_1}\rho_{p_1'}\left(1+\sum_{p_2}\frac{A_{p_1p_2}\omega^{p_2}}{Z_{p_2}}\right)\left(1+\sum_{p_2'}\frac{A_{p_1'p_2'}\omega^{p_2'}}{Z_{p_2'}}\right)\nu_t^{p_1p_1'}. \tag{8.2.16}$$

$$\tag{8.2.17}$$

In the absence of assortative reproduction, $Z_{p_2} = Z_{p_2'} \equiv Z^{(F)}$ is the normalization considering all females (because all females are connected to all the males p_2). Also, $\sum_{p_2} A_{p_1p_2}\omega^{p_2} = \sum_{p_2'} A_{p_1'p_2'}\omega^{p_2'} \equiv Z^{(M)}$ is the normalization considering all males, since any

male p_2 or p_2' can mate with any female p_1 in the system. Then, in this case, the evolution of the first moment reduces to

$$\mathbb{E}(\nu_{t+1}^{\alpha\beta}) = e^{-4\mu_M} \sum_{p_1,p_1'} \rho_{p_1}\rho_{p_1'} \left(1 + \frac{Z^{(M)}}{Z^{(F)}}\right)^2 \nu_t^{p_1p_1'}, \tag{8.2.18}$$

and we can assume, for large populations, that $Z^{(M)} \approx Z^{(F)}$ and that half of the population is male and the other half, female. Then

$$\begin{aligned}\mathbb{E}(\nu_{t+1}^{\alpha\beta}) &= 4e^{-4\mu_M} \sum_{p_1,p_1'} \rho_{p_1}\rho_{p_1'}\nu_t^{p_1p_1'} \\ &= 2e^{-4\mu_M} N \left[\langle \rho_{p_1}^2 \rangle_P + \left(\frac{N}{2} - 1\right) \langle \rho_{p_1}\rho_{p_1'}\nu_t^{p_1p_1'} \rangle_P\right],\end{aligned} \tag{8.2.19}$$

in which the averages over the probabilities ρ can be considered sex-independent.

8.2.3 The mito-nuclear similarity distribution

Now, we introduce the mito-nuclear similarity, which concerns the coupling between the two genetic sequences,

$$c^{\alpha\beta} = \frac{1}{B_M} \sum_{i=1}^{B_M} n_i^\alpha m_i^\beta. \tag{8.2.20}$$

There are two different cases to consider. Let us start with $\alpha \neq \beta$. Combining all the probabilities we have calculated so far, it is not hard to write the probability distribution of a given value at time $t+1$, given the population at time t, as

$$\begin{aligned}\mathcal{P}(c_{t+1}^{\alpha\beta}) = &\sum_{\mathbf{S}_{t+1}^\alpha, \mathbf{S}_{M,t+1}^\beta} \sum_{p_1,p_2} \sum_{p_1',p_2'} \delta\left(c_{t+1}^{\alpha\beta}, \mathbf{S}_{t+1}^\alpha \cdot \mathbf{S}_{M,t+1}^\beta / B_M\right) \\ &\times \frac{\omega_{p_1} A_{p_1p_2} \omega_{p_2}}{Z} \left(\frac{1}{Z_{p_1}} + \frac{1}{Z_{p_2}}\right) \frac{\omega_{p_1'} A_{p_1'p_2'} \omega_{p_2'}}{Z} \left(\frac{1}{Z_{p_1'}} + \frac{1}{Z_{p_2'}}\right) \\ &\times \prod_{i=1}^{B_M} \left[\frac{1}{2} + \frac{n_i^\alpha e^{-2\mu}}{4}(n_i^{p_1} + n_i^{p_2})\right] \frac{1}{2}\left[1 + e^{-2\mu_M} m_i^\beta m_i^{p_1'}\right],\end{aligned} \tag{8.2.21}$$

where again we consider p_1 and p_1' as females and p_2 and p_2' as males. Following the previous calculations, the expected value can be obtained as

$$\begin{aligned}\mathbb{E}(c_{t+1}^{\alpha\beta}) &= \sum_c c\mathcal{P}(c_{t+1}^{\alpha\beta} = c) \\ &= \sum_{p_1,p_2} \sum_{p_1',p_2'} \frac{\omega_{p_1} A_{p_1p_2} \omega_{p_2}}{Z} \left(\frac{1}{Z_{p_1}} + \frac{1}{Z_{p_2}}\right) \frac{\omega_{p_1'} A_{p_1'p_2'} \omega_{p_2'}}{Z} \left(\frac{1}{Z_{p_1'}} + \frac{1}{Z_{p_2'}}\right) \\ &\times \frac{1}{B_M} \sum_{j=1}^{B_M} n_j^\alpha m_j^\beta \sum_{\mathbf{S}_{t+1}^\alpha, \mathbf{S}_{M,t+1}^\beta} \left[\frac{1}{2} + \frac{n_i^\alpha e^{-2\mu}}{4}(n_i^{p_1} + n_i^{p_2})\right] \frac{1}{2}\left[1 + e^{-2\mu_M} m_i^\beta m_i^{p_1'}\right]\end{aligned}$$

$$
\begin{aligned}
&= \sum_{p_1,p_2}\sum_{p_1',p_2'} \frac{\omega_{p_1}A_{p_1p_2}\omega_{p_2}}{Z}\left(\frac{1}{Z_{p_1}}+\frac{1}{Z_{p_2}}\right)\frac{\omega_{p_1'}A_{p_1'p_2'}\omega_{p_2'}}{Z}\left(\frac{1}{Z_{p_1'}}+\frac{1}{Z_{p_2'}}\right)\\
&\times \frac{1}{B_M}\sum_{j=1}^{B_M}\frac{e^{-2\mu}}{2}(n_j^{p_1}+n_j^{p_2})e^{-2\mu_M}m_j^{p_1'}\\
&= \frac{e^{-4\overline{\mu}}}{2}\sum_{p_1,p_2}\sum_{p_1',p_2'} \frac{\omega_{p_1}A_{p_1p_2}\omega_{p_2}}{Z}\left(\frac{1}{Z_{p_1}}+\frac{1}{Z_{p_2}}\right)\frac{\omega_{p_1'}A_{p_1'p_2'}\omega_{p_2'}}{Z}\left(\frac{1}{Z_{p_1'}}+\frac{1}{Z_{p_2'}}\right)\left(c_t^{p_1p_1'}+c_t^{p_2p_1'}\right),
\end{aligned}
\tag{8.2.22}
$$

where we have defined

$$
\overline{\mu} \equiv \frac{\mu+\mu_M}{2}. \tag{8.2.23}
$$

In the case with no assortative reproduction, let us consider again, for a large population, $Z_{p_1} = Z_{p_1'} = Z^{(M)}$ and $Z_{p_2} = Z_{p_2'} = Z^{(F)}$ with $Z^{(M)} \approx Z^{(F)} \approx Z/2$, ($Z = Z^{(M)} + Z^{(F)}$). Thus,

$$
\begin{aligned}
\mathbb{E}(c_{t+1}^{\alpha\beta}) &= \frac{8e^{-4\overline{\mu}}}{Z^4}\sum_{p_1,p_2}\sum_{p_1',p_2'}\omega_{p_1}A_{p_1p_2}\omega_{p_2}\omega_{p_1'}A_{p_1'p_2'}\omega_{p_2'}\left(c_t^{p_1p_1'}+c_t^{p_2p_1'}\right)\\
&= \frac{2e^{-4\overline{\mu}}}{Z^2}\left(\sum_{p_1,p_1'}\omega_{p_1}\omega_{p_1'}c_t^{p_1p_1'}+\sum_{p_2,p_1'}\omega_{p_2}\omega_{p_1'}c_t^{p_2p_1'}\right)\\
&= 2e^{-4\overline{\mu}}\sum_{\overline{p},p_1'}\rho_{\overline{p}}\rho_{p_1'}c_t^{\overline{p}p_1'},
\end{aligned}
\tag{8.2.24}
$$

where the index $\overline{p}$ runs over the entire population (males and females) and then

$$
\mathbb{E}(c_{t+1}^{\alpha\beta}) = e^{-4\overline{\mu}}N\left[(N-1)\langle\rho_{\overline{p}}\rho_{p_1'}c_t^{\overline{p}p_1'}\rangle_P + \langle\rho_{p_1'}^2c_t^{p_1'p_1'}\rangle_P\right]. \tag{8.2.25}
$$

However, for the diagonal terms $\alpha = \beta$, the result is different, with

$$
\begin{aligned}
\mathcal{P}(c_{t+1}^{\alpha\alpha}) = &\sum_{\mathbf{S}_{t+1}^{\alpha},\mathbf{S}_{M,t+1}^{\alpha}}\sum_{p_1,p_2}\delta\left(c_{t+1}^{\alpha\alpha},\mathbf{S}_{t+1}^{\alpha}\cdot\mathbf{S}_{M,t+1}^{\alpha}/B_M\right)\frac{\omega_{p_1}A_{p_1p_2}\omega_{p_2}}{Z}\left(\frac{1}{Z_{p_1}}+\frac{1}{Z_{p_2}}\right)\\
&\times\prod_{i=1}^{B_M}\left[\frac{1}{2}+\frac{n_i^{\alpha}e^{-2\mu}}{4}(n_i^{p_1}+n_i^{p_2})\right]\frac{1}{2}\left[1+e^{-2\mu_M}m_i^{\alpha}m_i^{p_1}\right],
\end{aligned}
\tag{8.2.26}
$$

with the expected value given by

$$
\mathbb{E}(c_{t+1}^{\alpha\alpha}) = \frac{e^{-4\overline{\mu}}}{2}\sum_{p_1,p_2}\frac{\omega_{p_1}A_{p_1p_2}\omega_{p_2}}{Z}\left(\frac{1}{Z_{p_1}}+\frac{1}{Z_{p_2}}\right)(c^{p_1p_1}+c^{p_2p_1}). \tag{8.2.27}
$$

Again, in the absence of assortative reproduction, and for a large population, we find

$$\mathbb{E}(c_{t+1}^{\alpha\alpha}) = \frac{e^{-4\overline{\mu}}}{2} N \left[N \langle \rho_{p_1} \rho_{p_2} c_t^{p_1 p_2} \rangle_P + \langle \rho_{p_1} c_t^{p_1 p_1} \rangle_P \right]. \tag{8.2.28}$$

8.2.4 The average evolutions: Summing up

Let us sum up the previous results considering ensemble averages. To simplify the indexes, let us consider F as female, M as male and α and β as any sex. Thus, we find

$$\langle q_{t+1}^{\alpha\beta} \rangle = e^{-4\mu} N \left[(N-1) \langle \rho_\alpha \rho_\beta q_t^{\alpha\beta} \rangle + \langle \rho_\alpha^2 \rangle \right], \tag{8.2.29}$$

$$\langle \nu_{t+1}^{\alpha\beta} \rangle = 2e^{-4\mu_M} N \left[\left(\frac{N}{2} - 1 \right) \langle \rho_{F_1} \rho_{F_2} \nu_t^{F_1 F_2} \rangle + \langle \rho_\alpha^2 \rangle \right], \tag{8.2.30}$$

$$\langle c_{t+1}^{\alpha\beta} \rangle = e^{-4\overline{\mu}} N \left[(N-1) \langle \rho_\alpha \rho_F c_t^{\alpha F} \rangle + \langle \rho_F^2 c_t^{FF} \rangle \right], \tag{8.2.31}$$

$$\langle c_{t+1}^{\alpha\alpha} \rangle = e^{-4\overline{\mu}} \frac{N}{2} \left[N \langle \rho_M \rho_F c_t^{MF} \rangle + \langle \rho_F c_t^{FF} \rangle \right], \tag{8.2.32}$$

and propose the following simplifications

$$\langle q_{t+1}^{\alpha\beta} \rangle = e^{-4\mu} N \left[(N-1) \langle \rho_\alpha \rangle^2 \langle q_t^{\alpha\beta} \rangle + \langle \rho_\alpha^2 \rangle \right], \tag{8.2.33}$$

$$\langle \nu_{t+1}^{\alpha\beta} \rangle = 2e^{-4\mu_M} N \left[\left(\frac{N}{2} - 1 \right) \langle \rho_\alpha \rangle^2 \langle \nu_t^{\alpha\beta} \rangle + \langle \rho_\alpha^2 \rangle \right], \tag{8.2.34}$$

$$\langle c_{t+1}^{\alpha\beta} \rangle = e^{-4\overline{\mu}} N \left[(N-1) \langle \rho_\alpha \rangle^2 \langle c_t^{\alpha\beta} \rangle + \langle \rho_\alpha^2 c_t^{\alpha\alpha} \rangle \right], \tag{8.2.35}$$

$$\langle c_{t+1}^{\alpha\alpha} \rangle = e^{-4\overline{\mu}} \frac{N}{2} \left[N \langle \rho_\alpha \rangle^2 \langle c_t^{\alpha\beta} \rangle + \langle \rho_\alpha c_t^{\alpha\alpha} \rangle \right], \tag{8.2.36}$$

in which we also overcame the sex differences due to the fact that, once starting with a genetically identical population, the evolution equations are the same regardless of the sex of the individuals. Now, let us present two different results by considering the existence or not of coupling between the genetic sequences.

8.2.5 Without mito-nuclear coupling ($\sigma_\omega \to \infty$)

If we consider the system without any coupling between the genetic sequences, i.e., $\sigma_\omega \to \infty$, then the individuals have all the same probability of being chosen as focal, $\rho_\alpha \to 1/N$. In this case, the evolution of the nuclear similarity becomes exactly as in the Derrida-Higgs model, as expected. Also, for the mito-nuclear similarities we find

$$\langle c_{t+1}^{\alpha\beta} \rangle = e^{-4\overline{\mu}} \left[\left(1 - \frac{1}{N} \right) \langle c_t^{\alpha\beta} \rangle + \frac{1}{N} \langle c_t^{\alpha\alpha} \rangle \right], \tag{8.2.37}$$

$$\langle c_{t+1}^{\alpha\alpha} \rangle = \frac{e^{-4\overline{\mu}}}{2} \left[\langle c_t^{\alpha\beta} \rangle + \langle c_t^{\alpha\alpha} \rangle \right], \tag{8.2.38}$$

whose equilibrium solution is

$$\langle c^{\alpha\beta}_{t\to\infty}\rangle = \langle c^{\alpha\alpha}_{t\to\infty}\rangle = 0, \tag{8.2.39}$$

thus completely decoupling the genetic sequences, as also expected, once the fitness is the source of coupling. Fig.8.4 shows the evolution of the system without any coupling and we can see the system decoupling in the long time limit. A very interesting result we can see from the simulations is that even when there is species formation (Fig.8.4b), the average mito-nuclear similarity still follows the prediction for the case without mating restrictions (the dashed curves in the figure).

8.2.6 With mito-nuclear coupling ($\sigma_\omega < \infty$)

Of course, the case of interest is in the presence of coupling, which appears by means of fitness, resulting in a non-uniform probability distribution ρ_α. In this case, we cannot easily treat the terms $\langle \rho_\alpha^2 c_t^{\alpha\alpha}\rangle$ and $\langle \rho_\alpha c_t^{\alpha\alpha}\rangle$ which appear in the evolution equations (8.2.35) and (8.2.36), but a quite simple study can be pursued.

Let us suppose the probabilities ρ_α assume the form

$$\rho_\alpha = \rho_\alpha(c_t^{\alpha\alpha}) \sim \exp\left[-\left(c_t^{\alpha\alpha}-1\right)^2/2\sigma_\omega^2\right], \tag{8.2.40}$$

since the fitness is chosen to be a normal function of the mito-nuclear distance. Now, let us also assume that the distribution of $c_t^{\alpha\alpha}$ is also normal, with variance σ_c^2,

$$\mathcal{P}(c_t^{\alpha\alpha}) \sim \exp\left[-\left(c_t^{\alpha\alpha}-\langle c_t^{\alpha\alpha}\rangle\right)^2/2\sigma_c^2\right]. \tag{8.2.41}$$

Hence,

$$\langle \rho_\alpha c_t^{\alpha\alpha}\rangle \propto \int c_t^{\alpha\alpha} \exp\left[-\frac{\left(c_t^{\alpha\alpha}-1\right)^2}{2\sigma_\omega^2} - \frac{\left(c_t^{\alpha\alpha}-\langle c_t^{\alpha\alpha}\rangle\right)^2}{2\sigma_c^2}\right] dc_t^{\alpha\alpha}, \tag{8.2.42}$$

in which the exponent in brackets can be written as

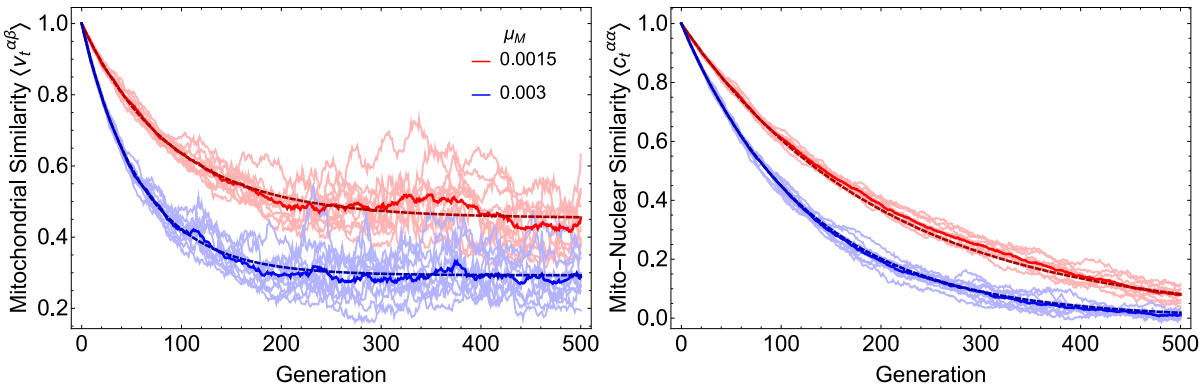
$$-\frac{\left(c_t^{\alpha\alpha}-\overline{c}\right)^2}{2\overline{\sigma}^2} - \frac{\left(\langle c_t^{\alpha\alpha}\rangle - 1\right)^2}{2\left(\sigma_\omega^2+\sigma_c^2\right)},$$

with

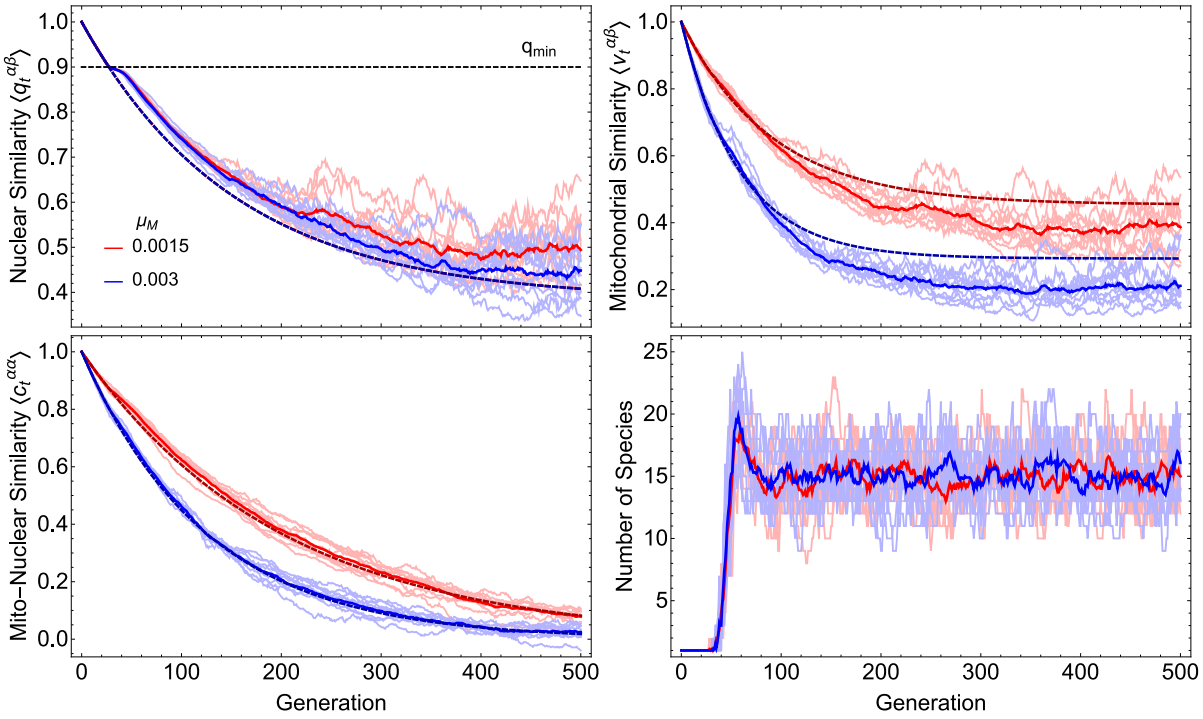
$$\overline{c} = \frac{\sigma_\omega^2\langle c_t^{\alpha\alpha}\rangle + \sigma_c^2}{\sigma_\omega^2+\sigma_c^2}, \tag{8.2.43}$$

and,

$$\overline{\sigma}^2 = \left[\frac{1}{\sigma_\omega^2}+\frac{1}{\sigma_c^2}\right]^{-1}. \tag{8.2.44}$$



(a) Absence of q_{min}.



(b) $q_{min} = 0.9$

Figure 8.4: **Evolution of the system in the absence of coupling.** In the figure, the average similarity (darker continuous curves) over 10 simulations (lighter curves) for two different values of μ_M, $\mu_M = 0.0015$ (red curves) and $\mu_M = 0.003$ (blue curves). The dashed lines are the numerical results of equations (8.2.33) to (8.2.36). In (a), there are no restrictions to mating. In (b), $q_{min} = 0.9$. In the figure, the simulation parameters are $N = 400$, $\mu = 0.001$, $B = 7500$ and $B_M = 250$.
Source: Figure produced by the author.

Thus,

$$\langle \rho_\alpha c_t^{\alpha\alpha} \rangle \propto \exp\left[-\frac{\left(\langle c_t^{\alpha\alpha}\rangle - 1\right)^2}{2\left(\sigma_\omega^2 + \sigma_c^2\right)}\right] \int c_t^{\alpha\alpha} \exp\left[-\frac{\left(c_t^{\alpha\alpha} - \overline{c}\right)^2}{2\overline{\sigma}^2}\right] dc_t^{\alpha\alpha}$$
$$= \exp\left[-\frac{\left(\langle c_t^{\alpha\alpha}\rangle - 1\right)^2}{2\left(\sigma_\omega^2 + \sigma_c^2\right)}\right] \left(\frac{\sigma_\omega^2 \langle c_t^{\alpha\alpha}\rangle + \sigma_c^2}{\sigma_\omega^2 + \sigma_c^2}\right). \tag{8.2.45}$$

The same calculation can be performed for $\langle \rho_\alpha^2 c_t^{\alpha\alpha} \rangle$, and the final result can be achieved by simply changing $\sigma_\omega^2 \to \sigma_\omega^2/2$,

$$\langle \rho_\alpha^2 c_t^{\alpha\alpha} \rangle \propto \exp\left[-\frac{\left(\langle c_t^{\alpha\alpha}\rangle - 1\right)^2}{2\left(\sigma_\omega^2/2 + \sigma_c^2\right)}\right] \left(\frac{\sigma_\omega^2/2 \langle c_t^{\alpha\alpha}\rangle + \sigma_c^2}{\sigma_\omega^2/2 + \sigma_c^2}\right). \tag{8.2.46}$$

The interesting results concerning these equations appear when we consider different limits:

- *(i)* $\sigma_\omega \gg \sigma_c$ (weak coupling limit): In this case,

$$\langle \rho_\alpha c_t^{\alpha\alpha} \rangle = C_0 \langle c_t^{\alpha\alpha} \rangle, \tag{8.2.47}$$
$$\langle \rho_\alpha^2 c_t^{\alpha\alpha} \rangle = C_0' \langle c_t^{\alpha\alpha} \rangle, \tag{8.2.48}$$

  where C_0 and C_0' are actually functions of $\langle c_t^{\alpha\alpha}\rangle$, which we are going to assume as constants, neglecting its complexity in order to be able to understand some general features of the model. Now, considering the evolution equations (8.2.35) and (8.2.36), the equilibrium solution is still trivial $\langle c_\infty^{\alpha\alpha}\rangle = 0$, being the system still uncoupled.

- *(ii)* σ_ω comparable to σ_c (strong coupling limit): Now,

$$\langle \rho_\alpha c_t^{\alpha\alpha} \rangle = C_1 + C_2 \langle c_t^{\alpha\alpha} \rangle, \tag{8.2.49}$$
$$\langle \rho_\alpha^2 c_t^{\alpha\alpha} \rangle = C_1' + C_2' \langle c_t^{\alpha\alpha} \rangle, \tag{8.2.50}$$

  in which again C_1, C_1', C_2 and C_2' are also functions of $\langle c_t^{\alpha\alpha}\rangle$ that we again treat as constants. However, now equations (8.2.35) and (8.2.36) lead to a non-trivial equilibrium solution

$$\langle c_\infty^{\alpha\alpha} \rangle = \frac{C_1 + N\langle \rho_\alpha \rangle_{Eq}^2 C_1' \left[e^{4\overline{\mu}}/N - (N-1)\langle \rho_\alpha \rangle_{Eq}^2\right]^{-1}}{2e^{4\overline{\mu}}/N - C_2 - N\langle \rho_\alpha \rangle_{Eq}^2 C_2' \left[e^{4\overline{\mu}}/N - (N-1)\langle \rho_\alpha \rangle_{Eq}^2\right]^{-1}}, \tag{8.2.51}$$

  showing that in this case, the coupling leads to a dependence between the genetic sequences.

These were quite strong approximations, but they were able to show us an important feature of the system: the genetic sequences lose coupling when $\sigma_\omega \gg \sigma_c$, i.e., when the coupling strength is much weaker (larger) than the broadness of the mito-nuclear similarity. Of course, the broadness σ_c is also a function of the coupling and this interplay is very

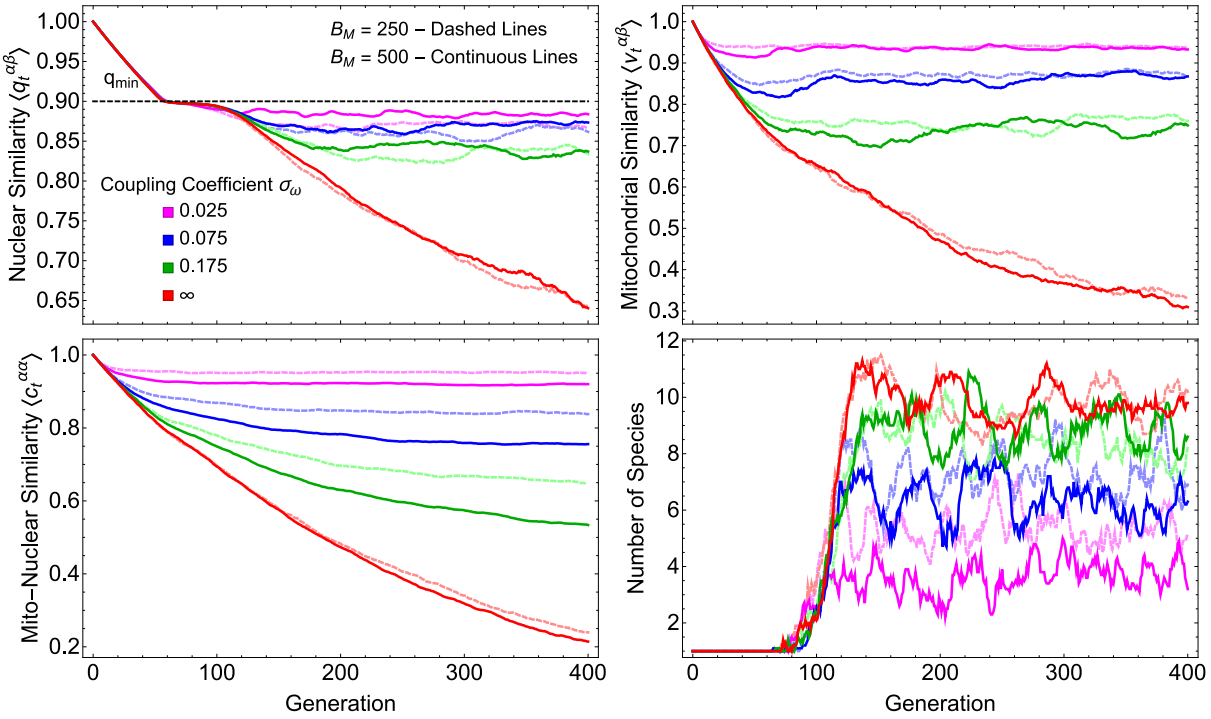


Figure 8.5: **Genome size B_M influence on the mito-nuclear coupling.** In the figure, the average similarity over 10 simulations of each parameter set with different coupling coefficients is presented. The lighter dashed curves are for $B_M = 250$ and the continuous darker curves for $B_M = 500$. The mito-nuclear similarity is always greater (i.e., higher coupling) for smaller genome sizes. In the figure, the simulation parameters are $N = 500$, $\mu = 0.0005$, $q_{min} = 0.9$, $B = 17000$ and $\mu_M = 0.0015$.
Source: Figure produced by the author.

complex, being this result a qualitative one, with the lack of a more formal quantitative expression.

In order to numerically test this result, as we learned from the Derrida-Higgs theory, we know that the σ_c decreases as the genome size increases. Thus, for the same coupling coefficient σ_ω, a greater value of B_M should result in a less coupled system. Fig.8.5 shows this result. In the following section we work on the variance of the genetic similarities and construct the algorithm for infinite genome sizes, showing the limiting case of the previous result, in which no matter the coupling, the system does not remain coupled in the long-time limit.

8.3 Infinite genome size

An algorithm for an infinite genome size lies on the assumption that in this case, given the parents of a pair of individuals, the genetic similarities are defined with probability one, once in this limit the variance should be zero. Let us first calculate the expressions involved in the algorithm and then show that the variance goes to zero, validating the algorithm.

8.3.1 The algorithm for infinite genomes

Every probability distribution of any similarity we have calculated so far has two parts: (a) the first concerns the network, which itself encodes the sampling process. The sum over the possible pairs of parents, the adjacency matrix, the fitness factors and the normalization factors, they are all telling how we should count the possible outcomes that come from the (b) sexual reproduction process. This part considers that, given the parents of both individuals (α and β, as we have been used), there is a correct way of mixing their genomes, which is written in the products, and then summed over many different genome outcomes ($\mathbf{S}^{\alpha}_{t+1}$ and $\mathbf{S}^{\beta}_{t+1}$) and filtered over the ones that give the desired similarity (by means of the δ function).

With this observation, we do not need to rewrite the probability distributions in order to calculate what we want, let us only focus on the reproduction part, once the algorithm draws the parents of an individual a priori.

Nuclear similarity

Let us start with the Nuclear similarity (which recovers the results from Derrida and Higgs). Given the parents of α and β, the probability of a given nuclear similarity is given by (from Eq.(8.2.4)),

$$\begin{aligned}&\mathcal{P}(q^{\alpha\beta}_{t+1}|\text{given the parents})\\ &=\sum_{\mathbf{S}^{\alpha}_{t+1},\mathbf{S}^{\beta}_{t+1}}\delta\left(q^{\alpha\beta}_{t+1},\mathbf{S}^{\alpha}_{t+1}\cdot\mathbf{S}^{\beta}_{t+1}/B\right)\prod_{i=1}^{B}\left[\frac{1}{2}+\frac{n^{\alpha}_{i,t+1}e^{-2\mu}}{4}(n^{p_1}_{i,t}+n^{p_2}_{i,t})\right]\left[\frac{1}{2}+\frac{n^{\alpha}_{i,t+1}e^{-2\mu}}{4}(n^{p'_1}_{i,t}+n^{p'_2}_{i,t})\right],\end{aligned} \tag{8.3.1}$$

in which p_1 and p_2 are parents of α and p'_1 and p'_2 are parents of β. In order to calculate the expected value, we follow the same procedures used before and find

$$\mathbb{E}(q^{\alpha\beta}_{t+1}|\text{given the parents})=\frac{e^{-4\mu}}{4}(q^{p_1p'_1}_t+q^{p_1p'_2}_t+q^{p_2p'_1}_t+q^{p_2p'_2}_t). \tag{8.3.2}$$

Mitochondrial similarity

Now, for the mitochondrial similarity, from (8.2.9)

$$\begin{aligned}&\mathcal{P}(\nu^{\alpha\beta}_{t+1}|\text{given the parents})\\ &=\sum_{\mathbf{S}^{\alpha}_{M,t+1},\mathbf{S}^{\beta}_{M,t+1}}\delta\left(\nu^{\alpha\beta}_{t+1},\mathbf{S}^{\alpha}_{M,t+1}\cdot\mathbf{S}^{\beta}_{M,t+1}/B_M\right)\prod_{i=1}^{B_M}\frac{1}{2}\left[1+e^{-2\mu_M}m^{\alpha}_i m^{p_1}_i\right]\frac{1}{2}\left[1+e^{-2\mu_M}m^{\beta}_i m^{p'_1}_i\right],\end{aligned} \tag{8.3.3}$$

also keeping the notation that p_1 and p'_1 are females. For the expected value, one can easily show

$$\mathbb{E}(\nu^{\alpha\beta}_{t+1}|\text{given the parents})=e^{-4\mu_M}\nu^{p_1p'_1}_t. \tag{8.3.4}$$

Mito-nuclear similarity

Finally, from Eq.(8.2.21)

$$\mathcal{P}(c_{t+1}^{\alpha\beta}|\text{given the parents})$$
$$= \sum_{\mathbf{S}_{t+1}^{\alpha},\mathbf{S}_{M,t+1}^{\beta}} \delta\left(c_{t+1}^{\alpha\beta}, \mathbf{S}_{t+1}^{\alpha}\cdot\mathbf{S}_{M,t+1}^{\beta}/B_M\right)\prod_{i=1}^{B_M}\left[\frac{1}{2}+\frac{n_i^{\alpha}e^{-2\mu}}{4}(n_i^{p_1}+n_i^{p_2})\right]\frac{1}{2}\left[1+e^{-2\mu_M}m_i^{\beta}m_i^{p_1'}\right], \tag{8.3.5}$$

and the expected value

$$\mathbb{E}(c_{t+1}^{\alpha\beta}|\text{given the parents}) = \frac{e^{-4\overline{\mu}}}{2}(c_t^{p_1p_1'}+c_t^{p_2p_1'}). \tag{8.3.6}$$

and this equation holds even for $\alpha=\beta$.

The algorithm for infinite genomes

As we are going to show, once the parents are given, when the genome sizes are both infinite, we should be able to write $\mathbb{E}(quantity|\text{given the parents}) = quantity$, once the variance $\text{Var}(quantity)\to 0$. Hence, for infinite genome sizes, we start with a clonal population with an equal number of males and females; the matrices $q_0^{\alpha\beta}=1$, $\nu_0^{\alpha\beta}=1$ and $c_0^{\alpha\beta}=1$. An individual p_{focal} is chosen according to its fitness $\omega_{p_{focal}}$ and then its partner $p_{partner}$ is chosen, according to its fitness, among the ones that can mate with p_{focal}, i.e., they should have different sexes and $q^{p_{focal}p_{partner}}\geq q_{min}$. N pairs like this are drawn. The matrices are updated according to

$$q_{t+1}^{\alpha\beta} = \frac{e^{-4\mu}}{4}(q_t^{p_1p_1'}+q_t^{p_1p_2'}+q_t^{p_2p_1'}+q_t^{p_2p_2'}), \tag{8.3.7}$$
$$\nu_{t+1}^{\alpha\beta} = e^{-4\mu_M}\nu_t^{p_1p_1'}, \tag{8.3.8}$$
$$c_{t+1}^{\alpha\beta} = \frac{e^{-4\overline{\mu}}}{2}(c_t^{p_1p_1'}+c_t^{p_2p_1'}), \tag{8.3.9}$$

in which p_1 and p_2 are respectively the female and male parents of α and p_1' and p_2' are respectively the female and male parent of β.

8.3.2 The similarity variance (given the parents)

Following our outline, let us now calculate the variance of the genetic similarities given the parents.

Nuclear similarity

Starting with the nuclear similarity, we recover the results from the Derrida-Higgs,

$$\mathbb{E}((q_{t+1}^{\alpha\beta})^2|\text{given the parents})$$

$$
\begin{aligned}
&= \sum_q q^2 \mathcal{P}(q_{t+1}^{\alpha\beta} = q | \text{given the parents}) \\
&= \sum_{\mathbf{s}_{t+1}^{\alpha}, \mathbf{s}_{t+1}^{\beta}} (q_{t+1}^{\alpha\beta})^2 \prod_{i=1}^{B} \left[\frac{1}{2} + \frac{n_{i,t+1}^{\alpha} e^{-2\mu}}{4} (n_{i,t}^{p_1} + n_{i,t}^{p_2}) \right] \left[\frac{1}{2} + \frac{n_{i,t+1}^{\alpha} e^{-2\mu}}{4} (n_{i,t}^{p_1'} + n_{i,t}^{p_2'}) \right] \\
&= \frac{1}{B} + \frac{e^{-8\mu}}{16} \left[(q_t^{p_1 p_1'} + q_t^{p_1 p_2'} + q_t^{p_2 p_1'} + q_t^{p_2 p_2'}) \right]^2 - \frac{e^{-8\mu}}{4B} \left(1 + q_t^{p_1 p_2} + q_t^{p_1' p_2'} + q_t^{p_1 p_2 p_1' p_2'} \right),
\end{aligned}
\tag{8.3.10}
$$

and then the variance

$$
\mathrm{Var}(q_{t+1}^{\alpha\beta} | \text{given the parents}) = \frac{1}{B} - \frac{e^{-8\mu}}{4B} \left(1 + q_t^{p_1 p_2} + q_t^{p_1' p_2'} + q_t^{p_1 p_2 p_1' p_2'} \right), \tag{8.3.11}
$$

which is zero for $B \to \infty$, q.e.d.

Mitochondrial similarity

For the mitochondrial similarity we have

$$
\begin{aligned}
&\mathbb{E}((\nu_{t+1}^{\alpha\beta})^2 | \text{given the parents}) \\
&= \sum_{\nu} \nu^2 \mathcal{P}(\nu_{t+1}^{\alpha\beta} = \nu | \text{given the parents}) \\
&= \sum_{\mathbf{s}_{t+1}^{\alpha}, \mathbf{s}_{t+1}^{\beta}} (\nu_{t+1}^{\alpha\beta})^2 \prod_{i=1}^{B_M} \frac{1}{2} \left[1 + e^{-2\mu_M} m_i^{\alpha} m_i^{p_1} \right] \frac{1}{2} \left[1 + e^{-2\mu_M} m_i^{\beta} m_i^{p_1'} \right] \\
&= \frac{1}{B_M^2} \sum_{\mathbf{s}_{t+1}^{\alpha}, \mathbf{s}_{t+1}^{\beta}} \sum_{j,k} m_j^{\alpha} m_j^{\beta} m_k^{\alpha} m_k^{\beta} \prod_{i=1}^{B_M} \frac{1}{2} \left[1 + e^{-2\mu_M} m_i^{\alpha} m_i^{p_1} \right] \frac{1}{2} \left[1 + e^{-2\mu_M} m_i^{\beta} m_i^{p_1'} \right] \\
&= \frac{1}{B_M^2} \sum_j \left[\left(\sum_{\mathbf{s}_{M,t+1}^{\alpha}} \prod_{i=1}^{B_M} \frac{1}{2} \left[1 + e^{-2\mu_M} m_i^{\alpha} m_i^{p_1} \right] \right) \left(\sum_{\mathbf{s}_{M,t+1}^{\beta}} \prod_{i=1}^{B_M} \frac{1}{2} \left[1 + e^{-2\mu_M} m_i^{\beta} m_i^{p_1'} \right] \right) \right] \\
&+ \frac{1}{B_M^2} \sum_{k \neq j} \left[\left(\sum_{\mathbf{s}_{M,t+1}^{\alpha}} m_j^{\alpha} m_k^{\alpha} \prod_{i=1}^{B_M} \frac{1}{2} \left[1 + e^{-2\mu_M} m_i^{\alpha} m_i^{p_1} \right] \right) \left(\sum_{\mathbf{s}_{M,t+1}^{\beta}} m_j^{\beta} m_k^{\beta} \prod_{i=1}^{B_M} \frac{1}{2} \left[1 + e^{-2\mu_M} m_i^{\beta} m_i^{p_1'} \right] \right) \right] \\
&= \frac{1}{B_M} + \frac{1}{B_M^2} \sum_{k \neq j} e^{-8\mu_M} m_j^{p_1} m_k^{p_1} m_j^{p_1'} m_k^{p_1'} \\
&= \frac{1}{B_M} + \frac{e^{-8\mu_M}}{B_M^2} \sum_{j=1}^{B_M} m_j^{p_1} m_j^{p_1'} \left(-m_j^{p_1} m_j^{p_1'} + \sum_{k=1}^{B_M} m_k^{p_1} m_k^{p_1'} \right) \\
&= \frac{(1 - e^{-8\mu_M})}{B_M} + \left(e^{-4\mu_M} \nu_t^{p_1 p_1'} \right)^2,
\end{aligned}
\tag{8.3.12}
$$

and then the variance

$$
\mathrm{Var}(\nu_{t+1}^{\alpha\beta} | \text{given the parents}) = \frac{(1 - e^{-8\mu_M})}{B_M}, \tag{8.3.13}
$$

which is again zero for $B_M \to \infty$, q.e.d.

Mito-nuclear similarity

Now, for the mito-nuclear similarity,

$$
\begin{aligned}
&\mathbb{E}((c_{t+1}^{\alpha\beta})^2|\text{given the parents}) \\
&= \sum_c c^2 \mathcal{P}(c_{t+1}^{\alpha\beta} = c|\text{given the parents}) \\
&= \sum_{\mathbf{s}_{t+1}^{\alpha},\mathbf{s}_{t+1}^{\beta}} (c_{t+1}^{\alpha\beta})^2 \prod_{i=1}^{B_M} \frac{1}{2}\left[1 + e^{-2\mu_M} m_i^{\alpha} m_i^{p_1}\right] \frac{1}{2}\left[1 + e^{-2\mu_M} m_i^{\beta} m_i^{p_1'}\right] \\
&= \frac{1}{B_M^2} \sum_{\mathbf{s}_{t+1}^{\alpha},\mathbf{s}_{t+1}^{\beta}} \sum_{j,k} n_j^{\alpha} m_j^{\beta} n_k^{\alpha} m_k^{\beta} \prod_{i=1}^{B_M} \left[\frac{1}{2} + \frac{n_i^{\alpha} e^{-2\mu}}{4}(n_i^{p_1} + n_i^{p_2})\right] \frac{1}{2}\left[1 + e^{-2\mu_M} m_i^{\beta} m_i^{p_1'}\right] \\
&= \frac{1}{B_M^2} \sum_j \left[\left(\sum_{\mathbf{s}_{M,t+1}^{\alpha}} \prod_{i=1}^{B_M} \left[\frac{1}{2} + \frac{n_i^{\alpha} e^{-2\mu}}{4}(n_i^{p_1} + n_i^{p_2})\right]\right)\left(\sum_{\mathbf{s}_{M,t+1}^{\beta}} \prod_{i=1}^{B_M} \frac{1}{2}\left[1 + e^{-2\mu_M} m_i^{\beta} m_i^{p_1'}\right]\right)\right] \\
&+ \frac{1}{B_M^2} \sum_{k \neq j} \left[\left(\sum_{\mathbf{s}_{M,t+1}^{\alpha}} n_j^{\alpha} n_k^{\alpha} \prod_{i=1}^{B_M} \left[\frac{1}{2} + \frac{n_i^{\alpha} e^{-2\mu}}{4}(n_i^{p_1} + n_i^{p_2})\right]\right)\left(\sum_{\mathbf{s}_{M,t+1}^{\beta}} m_j^{\beta} m_k^{\beta} \prod_{i=1}^{B_M} \frac{1}{2}\left[1 + e^{-2\mu_M} m_i^{\beta} m_i^{p_1'}\right]\right)\right] \\
&= \frac{1}{B_M} + \frac{e^{-8\overline{\mu}}}{4B_M^2} \sum_{k \neq j} (n_j^{p_1} + n_j^{p_2})(n_k^{p_1} + n_k^{p_2}) m_j^{p_1'} m_k^{p_1'} \\
&= \frac{1}{B_M} + \frac{e^{-8\overline{\mu}}}{4B_M^2} \sum_{j=1}^{B_M} (n_j^{p_1} m_j^{p_1'} + n_j^{p_2} m_j^{p_1'}) \left(-(n_j^{p_1} m_j^{p_1'} + n_j^{p_2} m_j^{p_1'}) + \sum_{k=1}^{B_M} (n_k^{p_1} m_k^{p_1'} + n_k^{p_2} m_k^{p_1'})\right) \\
&= \frac{1}{B_M} + \left(\frac{e^{-4\overline{\mu}}}{2}(c_t^{p_1 p_1'} + c_t^{p_2 p_1'})\right)^2 - \frac{e^{-8\overline{\mu}}}{2B_M^2} \sum_{j=1}^{B_M} (1 + n_j^{p_1} n_j^{p_2}) \\
&= \frac{1}{B_M} - \frac{e^{-8\overline{\mu}}}{2B_M}(1 + q_t^{p_1 p_2}) + \left(\frac{e^{-4\overline{\mu}}}{2}(c_t^{p_1 p_1'} + c_t^{p_2 p_1'})\right)^2,
\end{aligned}
\tag{8.3.14}
$$

and the variance reads

$$
\text{Var}(c_{t+1}^{\alpha\beta}|\text{given the parents}) = \frac{1}{B_M} - \frac{e^{-8\overline{\mu}}}{2B_M}(1 + q_t^{p_1 p_2}), \tag{8.3.15}
$$

equalling zero for $B_M \to \infty$, q.e.d.

Hence, we have validated the proposed algorithm in the infinite genome case.

8.4 Barcode computational results

Following Princepe and Aguiar [120], the parameters for our simulations were chosen as $\mu_M = 3\mu$, as also $\mu_M < 1/B_M$, thus avoiding the *error threshold* limit[1], in which the genetic space would be widely occupied, and thus the mitochondrial population fragmented into many rare species as a simple consequence of a high mutation rate. Also, the nuclear genome size must be large enough in order to allow species formation, and following biological constraints, $B_M < B$.

We remember that the mitochondrial similarity threshold q^M_{min} is calculated as the smallest value such that the number of mitochondrial species $\mathcal{S}_M$ is at least the number of nuclear species $\mathcal{S}$. The barcode success is then calculated as the ratio of the species' bijection (biunivocal correspondence between mitochondrial and nuclear species) to the number of mitochondrial species.

Fig.8.6 shows the results for a given set of parameters. We calculate the barcode success for every generation for 10 different simulations of each parameter set. We first notice the dependence on the number of nuclear species to the coupling coefficient. The greater the coupling (smaller σ_ω), the smaller the number of species, which is an effect of the selection induced by the fitness distribution. The smaller the σ_ω, the narrower the fitness distribution, intensifying the selection.

In the neutral case ($\sigma_\omega \to \infty$), there is no emergency of barcode, reaching success around 0.5. When we increase the coupling, we observe worse results. We check its consistency by analyzing a single generation (the last one) and we present the results as a boxplot of all 10 simulations. We also notice that for greater coupling, the mitochondrial similarity threshold needs to be more restrictive in order to reach at least $\mathcal{S}$ mitochondrial species.

Figure 8.7 shows the boxplots for barcode success and corresponding q^M_{min} over a larger set of points. They include all the generations from $t = 200$ to $t = 400$ of 10 simulations of each parameter set. This result regards the same simulations of Fig.8.5, and we observe the same results of the previous simulations (Fig.8.6): the greater the coupling, the worse the barcode success.

A small difference can be observed for different genome sizes: the smaller B_M, the worse the barcode, which is consistent with the previous results, which says that the greater the genome size, the smaller the coupling.

This result seems to be very counterintuitive, since we would expect a better barcode success for a greater coupling, or at least no differences from the neutral case, which is not what we observe. This is an effect of how we define the barcode success, being a direct consequence of the number of mitochondrial species $\mathcal{S}_M$ compared to the number of nuclear species $\mathcal{S}$. When we analyze the relative difference of these numbers $(\mathcal{S}_M - \mathcal{S})/\mathcal{S}$, shown in Fig.8.8, we see that this deviation is larger for larger couplings and, as a consequence, a smaller number of species bijections for larger couplings, resulting in a worse barcode success.

[1]For self-replicating systems, high mutation rates can make the genetic sequences explore the whole configuration space, leading to loss of adaptation. The limiting mutation rate is known as the *error threshold* [135].

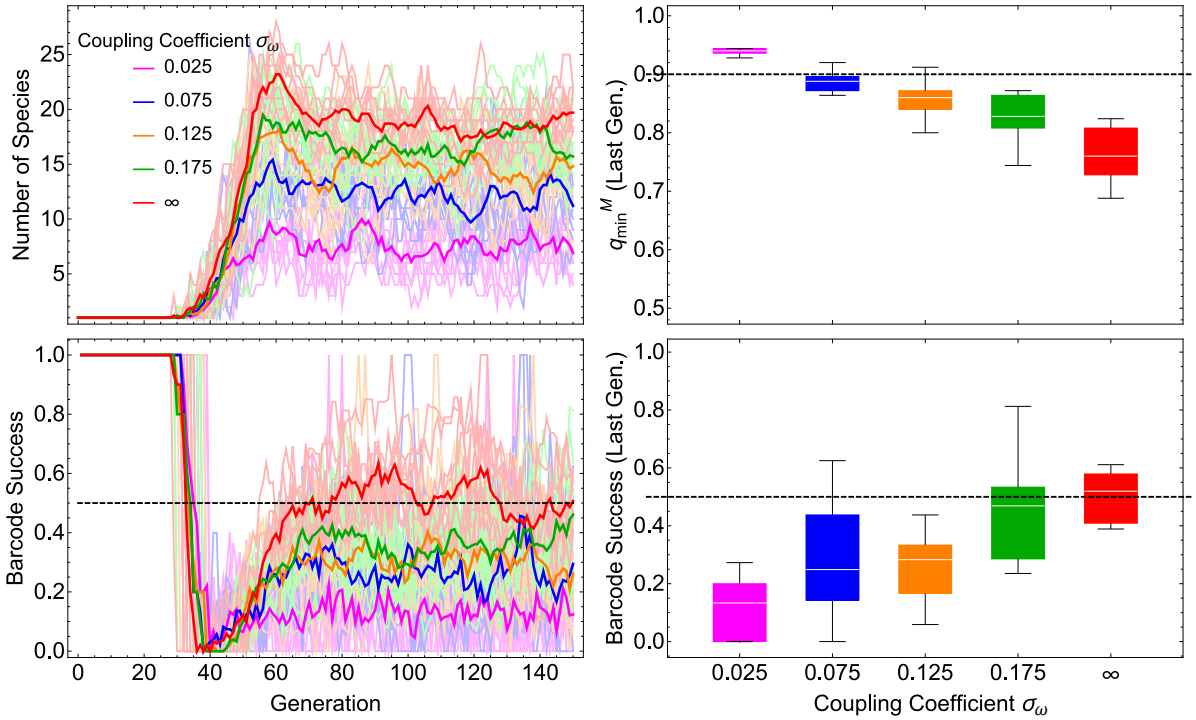


Figure 8.6: **Barcode success evolution.** (These results are from the same simulations of Fig.8.3). In the figure, the average result over 10 simulations of each parameter set, with different coupling coefficients, are presented. The bottom left panel shows the barcode success calculated over every generation, while the right bottom curve shows a boxplot representation of the 10 simulations only at the last generation ($t = 150$). The dashed black curve shows the value 0.5, which coincides with the non-interacting case ($\sigma_\omega \to \infty$). The top right panel shows a box plot of the results for the similarity threshold calculated also at the last generation, with the black dashed line showing the nuclear similarity threshold q_{min}. In the figure, the simulation parameters are $N = 500$, $\mu = 0.001$, $q_{min} = 0.9$, $B = 7500$, $B_M = 250$ and $\mu_M = 0.003$.
Source: Figure produced by the author.

8.5 Concluding remarks

The coevolution between the nuclear and mitochondrial genetic materials are fundamental for the cell respiration process, maximizing the energy uptake of eukaryotic biota by means of evolution [132]. When the coupling between both genetic materials is not good, the expression of genes related to the respiration process is not optimized, thus reducing the fitness of the individual [136]. As a consequence, the mitochondrial DNA becomes specific to its species, posing the so-called barcode property, i.e., to analyze the mitochondrial genome sequence is enough to recognize its species [119, 132].

The Derrida-Higgs model provides a simple framework for modeling this evolutive system and such an extension was proposed by Princepe and Aguiar in 2021 for allopatric communities [120]. Their main result was that the barcode property appears as a result of the population structure in space. The greater the overlap of species in space, the worse the barcode success, although they did not study the full sympatric case, their result conjectures what we have finally shown in our result: barcode does not emerge in

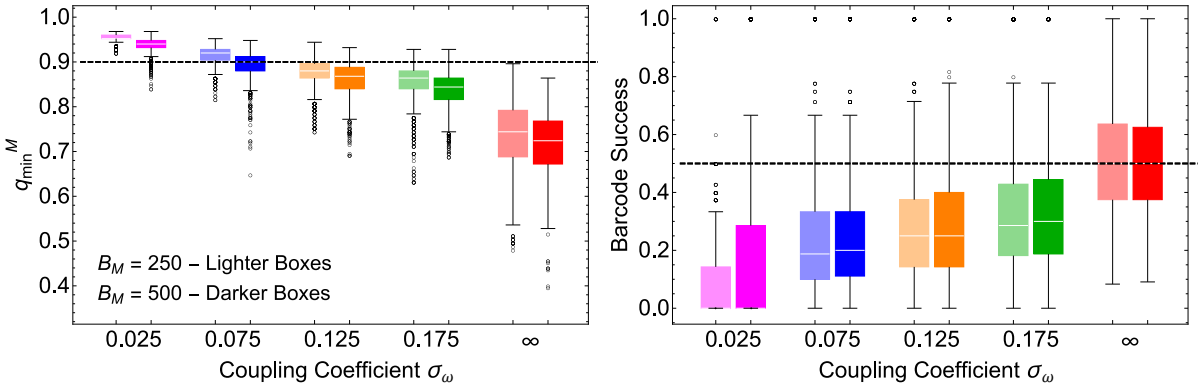


Figure 8.7: **Barcode success and mitochondrial threshold q^M_{min}.** (These results are from the same simulations of Fig.8.5). In the figure, the boxplots show the results for q^M_{min} and barcode success over the last 200 generations (after richness equilibration). On the left, the dashed black line shows the nuclear similarity threshold q_{min}, while on the right it shows the value 0.5, expected for the neutral case. The lighter boxes are for $B_M = 250$ and the darker boxes for $B_M = 500$. In the figure, the simulation parameters are $N = 500$, $\mu = 0.0005$, $q_{min} = 0.9$, $B = 17000$ and $\mu_M = 0.0015$.
Source: Figure produced by the author.

a sympatric population (and evolving according to the rules of the model).

The results point out an even worse than 50% of success when the coupling increases from the neutral case. This result is explained by the number of mitochondrial species that emerge in the system in the best case, which, on average, is greater than the number of nuclear species, reducing the number of bijections between them.

Despite no barcode in the system, the coupling, which depends not only on the coupling coefficient σ_ω but also on the mitochondrial genome size B_M, still afects the evolution, resulting in a smaller richness the greater the coupling, which can be understood as the equilibrium value of the mito-nuclear similarity: the closer to the unit, the stronger it is (i.e., this system shows stabilizing selection effect).

Our findings corroborate the results of Princepe and Aguiar [120] as we have also introduced the analytical theory for the sympatric case. Much of this model can still be studied, like the species abundance distribution as a function of the coupling coefficient, species turnover rates [136], closed forms for the fitness distributions or different definitions for the barcode success, but we believe to have finished this chapter with a substantial contribution to the general theory of finite genome sympatric Derrida-Higgs models.

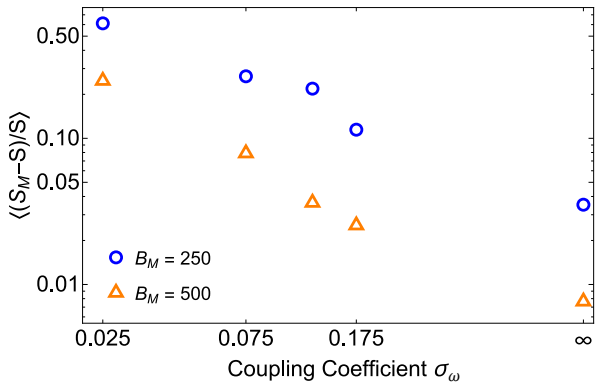


Figure 8.8: **Relative difference between Mitochondrial and Nuclear species.** (These results are from the same simulations of Fig.8.5). The figure shows the average relative distance between the number of Mitochondrial $\mathcal{S}_M$ and Nuclear species $\mathcal{S}$, $\langle(\mathcal{S}_M - \mathcal{S})/\mathcal{S}\rangle$, for all the generations from $t = 200$ to $t = 400$ (after richness equilibration). The blue circles are for the smaller mDNA ($B_M = 250$) and the orange circles are for the larger mDNA ($B_M = 500$). The plot is in log-log scale. In the figure, the simulation parameters are $N = 500$, $\mu = 0.0005$, $q_{min} = 0.9$, $B = 17000$ and $\mu_M = 0.0015$.
Source: Figure produced by the author.

Chapter 9

This part in a nutshell

The nature of the genetic code, its discrete and binary characterization, makes it very suitable for the design of evolutionary dynamics in computational and mathematical ways [16]. In 1991, B. Derrida and P. Higgs introduced a population dynamics with infinite loci that is able to cluster individuals in a way we can recognize different species [94]. The absence of a geographical structure, i.e., it is a sympatric community, makes the process of speciation a very striking result of the model. On the other hand, it can be understood as the interplay of two different evolutionary forces: mutations and genetic drift (the model has no differential fitness and is hence neutral), with one winning over the other in different regimes: low and high diversity regimes.

The transition parameters for infinite genome size were conjectured in the original 1991 paper, but it was shown by Aguiar in 2017 [118] that when the number of loci is finite, the Derrida and Higgs condition for speciation is not enough. Furthermore, the system showed a minimum genome size B_c to show species formation, but an analytical expression for B_c were not obtained. In this part, we have developed a theory for the genetic similarity distribution between the individuals in the population, which is a measure of how diverse the population is.

A heuristic theory for the transition and an approximated transition curve is obtained, i.e., we calculate B_c as a function of the other parameters of the model (the mutation rate, population size and assortative parameter). In addition, the high diversity phase is described as a Markov chain, in which the greatest challenge is to find the speciation probability of a species of a given size.

The Derrida-Higgs model is a very plastic dynamics, and can be adapted to different contexts and also extended in different directions [134]. In 2021, Princepe and Aguiar introduced a model of mito-nuclear coevolution grounded on the Derrida-Higgs model [120]. The mitochondrial genetic material is responsible for the expression of genes involved in the cell respiration process [122]. Its coevolution with the nuclear DNA is therefore an important process in eukaryotic cells [132]. It is also used for species identification due to its barcode property [119, 132], and this characteristic was studied in the Princepe-Aguiar model in an allopatric context [120].

We finish the present part by considering the Princepe-Aguiar model in a sympatric community, applying the formalism we have developed for the Derrida-Higgs dynamics in the coevolution of mitochondrial and nuclear genetic material. We find different coupling regimes for the genetic sequences, with numerical investigation that the barcode (as it is defined in the model) does not emerge in sympatric communities, corroborating the

conclusions of Princepe and Aguiar.

Part III

A Model of Viral Evolution

Chapter 10

Modeling epidemics

10.1 The COVID-19 pandemic in a (tiny) nutshell

At the end of 2019, the world started to face a massive spread of a new virus [137, 138], the SARS-CoV-2, responsible for the respiratory disease COVID-19 [139, 140]. This new virus was first detected in Wuhan, China, and in only a few months, it could be found in almost the entire world [139]. On March 11 of 2020, the World Health Organization (WHO) declared the COVID-19 pandemic [141]. In a glimpse, a worldwide *tour de force* was born in order to fight against what became one of the greatest global health problems of the modern era. Scientists from diverse areas engaged in data analysis forecast algorithms, epidemiological modeling [142, 143, 144], methods of non-pharmacological intervention [145, 146, 147], vaccine [148, 149, 150] and treatment [140, 151] research, as they were also trying to work close to governments, advising public health secretaries and healthcare centers [152].

The consequences of the SARS-CoV-2 pandemic were devastating [153, 154, 155]. Its transmission through the air is difficult to control [140, 156, 157], mainly in highly populated areas; its "flu-like" symptoms [140] were easily despised and, combined with a high number of asymptomatic cases [158, 159, 160], a natural reservoir for the virus have emerged throughout the population. Overcrowded hospitals and the sudden depletion of medical supplies, as well as the lack of efficient medical treatments [161, 162, 163], urged the necessity of quarantine and lockdown policies [145, 164].

On the other hand, its effects on countries' economies turned the COVID-19 pandemic into a political issue [165]. Right-aligned governments advocated for the reopening of commercial places and for the end social distancing policies [165, 166]. Misinformation and wrong information were very disappointing features of this era [167, 168, 169]. The "efficacy" of (useless) medicines was daily spread through social media, as also the use of non-pharmacological measures, such as the use of face masks and hand sanitizers, were ridicularized by those who did not *believe* in the pandemic. When vaccines finally appeared by the end of 2020, supported by intense scientific studies [170], anti-vaccine demonstrations and protests came out [171, 172, 173], embedded in a very political and denialist speech.

The first vaccine to be approved in Western countries was the Pfizer-BioNTech vaccine, on December 2 of 2020, in the United Kingdom [174, 175]. A different mathematical research line was then in vogue: optimal vaccination schemes. The age-dependent effects

of COVID-19 posed the elderly (and other groups of risk) in the first positions for getting vaccinated [176, 177]. Many vaccines, of different countries and companies, as also of different technologies were sold worldwide [170], bringing the first clear path to the end of the pandemic, albeit highlighting once more the discrepancies between rich and poor countries [178, 179].

In spite of the international efforts to end the pandemics, biological evolution showed off its role and many new strains of the SARS-CoV-2 were identified since the end of 2020 (e.g. Beta strain in South Africa [180, 181]). Mutants with different transmission rates were spreading out and the previously acquired immunity (artificial and natural) was not enough to block the emergence of new infection waves [180, 182, 183]. Vaccines needed to be adapted to take into account the new viral strains, and new researches were being conducted, but at this point, more efficient treatments had already been found and daily life started to go back to normal.

The World Health Organization decreed the end of the pandemic on May 5 of 2023 [184].

Until October 25, 2023, more than 771 million cases were confirmed and 6.9 million deaths were registered worldwide [155]. In Brazil, more than 704 thousand deaths were counted until the same date [155], after an incredibly disorganized and unscientific response to the pandemic by the national government [185, 166]. Despite of the hard task, scientists and science institutions fulfilled their duty to society. The COVID-19 pandemic, among many consequences, brought to light the importance of science. But not only its importance *per se*, but also how it is essential to maintain a comprehensible dialogue with society, being a trustworthy and perhaps the utmost source of information.

This part of the thesis presents our contributions to the comprehension of epidemic spreads. During the COVID-19 pandemic, we developed a framework to study the possible effects of genetic variability on the spread of a virus, which is a characteristic not included in the most common epidemiological models, i.e., mean-field *compartmental models* [22, 186]. These models are focused on the number of infections, which is indeed very important for epidemiological monitoring, still, multi-strain models [183, 187], such as ours, can be a source for many insights on how the spread of a disease behaves when it is genetically diverse. Results on this matter can be beneficial to, for instance, vaccination strategy research, since the efficacy of a given vaccine can decrease if the pathogen has many strains [182, 188].

The framework we are about to introduce is a compartmental model on a network, in which the pathogen (a RNA virus, as the SARS-CoV-2 [137]) is, inspired by the Derrida-Higgs model [94], described as a binary sequence. The dynamics and details of the model are going to be discussed in the next chapters, as also the results we have achieved. This research has already been published in two different scientific papers, Marquioni and de Aguiar 2020 [189] and Marquioni and de Aguiar 2021 [190].

10.2 On modeling epidemics

The classical way of modeling an epidemic is by means of compartmental models [186] at the population level, i.e., there is no individual tracking of the disease state and the epidemic is measured as a number describing the size of each compartment, e.g., the number of infected individuals or the number of recovered individuals. This type of dynamics has been written as sets of ordinary differential equations, subdividing the

population in many different compartments as also structuring it according to age intervals [191]. Equations (10.2.1) show a mean-field SEIR model [186], describing the dynamics of an epidemic among Susceptible (S), Exposed (E), Infected (I) and Recovered (R).

$$\begin{aligned}\dot{S} &= -\beta SI/N,\\ \dot{E} &= \beta SI/N - \sigma E,\\ \dot{I} &= \sigma E - \gamma I,\\ \dot{R} &= \gamma I.\end{aligned} \tag{10.2.1}$$

The parameters β, σ and γ are respectively the rates of transmission, transition from non-infectious to infectious and recovery.

This type of model is not restricted to epidemiology, and its importance ranges from chemical reactions [192] to ecology [193]. For instance, the Lotka-Volterra [22] model of predation divides a population into two compartments, one for predators and one for prey, and two differential equations describe the evolution of frequencies of each population.

The mathematical formalism is grounded on reaction networks theory [25, 26], in which the so-called *mass-action law* [194] sets the rates of change of each compartment. Everything that goes beyond mass-action law can, a priori, be included as a theory for the rates coefficients, making this type of model a very general and powerful formalism. Stochasticity can be added with noise terms [195], or as random couplings between equations [196], where stochastic analysis [43] is a very valuable tool. On the other hand, population level models do not include individuals' evolution, being unable to track the state of a specific agent in the population, nor it is possible to an measure individual's response to external inputs. This is where *agent-based models* (ABMs, or *individual-based models*, IBMs) come to the scene [197, 29].

In the following model, the individuals are still divided into compartments, albeit we now know in which compartment each individual is. Thus, one can ask if individual x is infected or not. In order to do that, we consider individuals as vertices of a network, with the edges representing individuals who are in contact to each other. Hence such a network is called the *contact network*. We are still going to be interested in population quantities, but this type of model is also suitable for studying specific behaviors, such as the role of very connected vertices (local *hubs* [198]) or the spread in complex heterogeneous populations [199].

Our interest in an agent-based approach to model an epidemic is reminiscent of the Derrida-Higgs model, allowing us to describe the viruses infecting the individuals, their replication and transmission processes. We are not including important features such as immunity waning [200] and transmission heterogeneities across different viral strains [181], but as it is going to be clear, it is possible to inferr its effects. Notwithstanding, this is not a limitation of the framework, it is only our set of assumptions, which can be easily changed in different studies. We therefore argue that this model can be as general as any compartmental model can be, relying on computational power and numerical analysis for obtaining strong results.

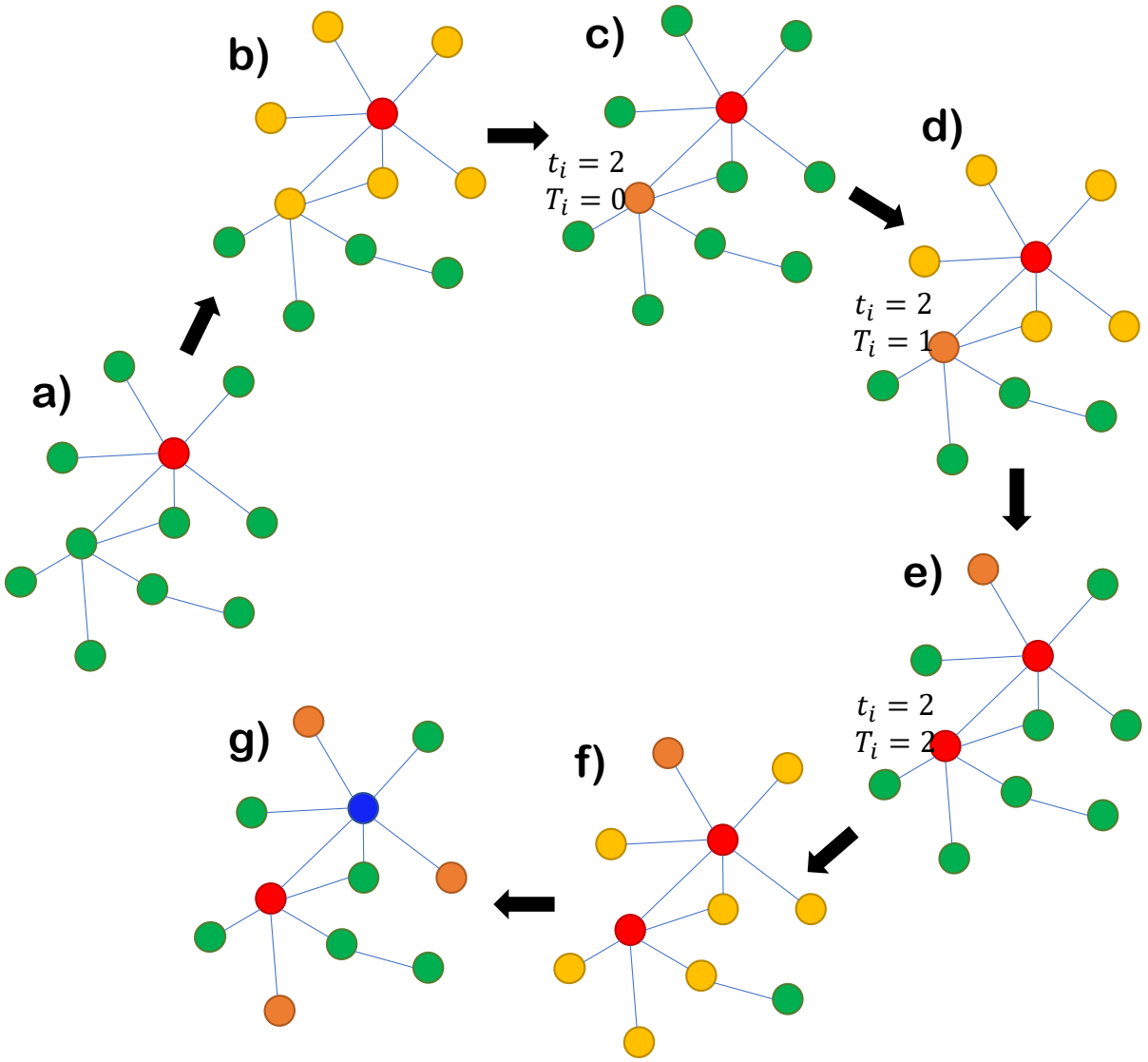


Figure 10.1: **Illustration of virus spread on the network.** (a) initially only one individual is infected (red) and all the others susceptible (green); (b) neighbors of infected might get the virus (yellow); (c) one neighbor did get the virus and become exposed (orange); (d) neighbors of first infected individual might still get the virus, but the orange node is still in incubation time; (e) another node gets the virus from the first infected becoming exposed and the old exposed becomes infected; (f) more nodes might get the virus (yellow) and; (g) some do become exposed while the first infected becomes recovered. Source: Figure from Marquioni and de Aguiar, 2020 [190].

10.3 The model

Our model can be split into two parts: the spread and the evolution, which are described below.

10.3.1 The spread

We consider a compartmental model of SEIR type, i.e., individuals are divided in

- *Susceptible (S)*: those that were not infected and non-resistant to infection;
- *Exposed (E)*: individuals who are infected but still not infectious;
- *Infected (I)*: individuals who are infected and infectious;
- *Recovered (R)*: those that were infected but have already recovered from the infection, being non-susceptible anymore.

Individuals are described as the nodes of a network, in which the edges describe the contact between the individuals. This *contact network* is fixed in time for the present model, and the final state of a spread is dependent on its structure.

An epidemic in the network runs as follows:

1. An infected (I) node can infect any susceptible (S) node to which it is connected with *probability of transmission* p_I in every time step.
2. Whenever an individual gets infected, it remains exposed (E) (i.e., not infectious) for an *incubation time* t_i, drawn from a distribution $\mathcal{P}(t_i)$.
3. After the incubation time, an exposed individual becomes infectious (I), being also able to spread the disease to other susceptible nodes.
4. Every infected (and infectious) individual can recover with probability r at the end of the iteration step.

Fig. 10.1 summarizes the spread part of the model.

10.3.2 The evolution

The novelty we address in this work is a *microscopic* description of the pathogen evolution in terms of binary sequences. We introduce a suitable agent-based framework to understand how the contact structure among individuals can affect the evolution of a RNA-virus that is spreading throughout a community.

A virus is described as a $2B$ binary string representing its genetic material, where B is the genome size and every pair of bits (b_{2i-1}, b_{2i}) defines a nucleotide, for instance, $(0,0) = A$, $(0,1) = U$, $(1,0) = C$ and $(1,1) = G$. We consider a single sequence as a proxy for the infection in an individual. This sequence can mutate as long as the individual remains infected (or exposed), with mutation rate μ per nucleotide [201]. The evolution follows as

1. When a transmission event occurs, the binary sequence is copied from the infected to the infectee;
2. When an individual recovers, its virus cannot mutate anymore and the *final virus* is saved;
3. All viruses (within infected and exposed) can mutate in every time step;

Fig. 10.2 sums up this dynamics. In the present study, all viruses are considered equivalent, showing no fitness differences or distinct transmission and death rates. Individuals also acquire perfect cross-immunity once infected by any strain.

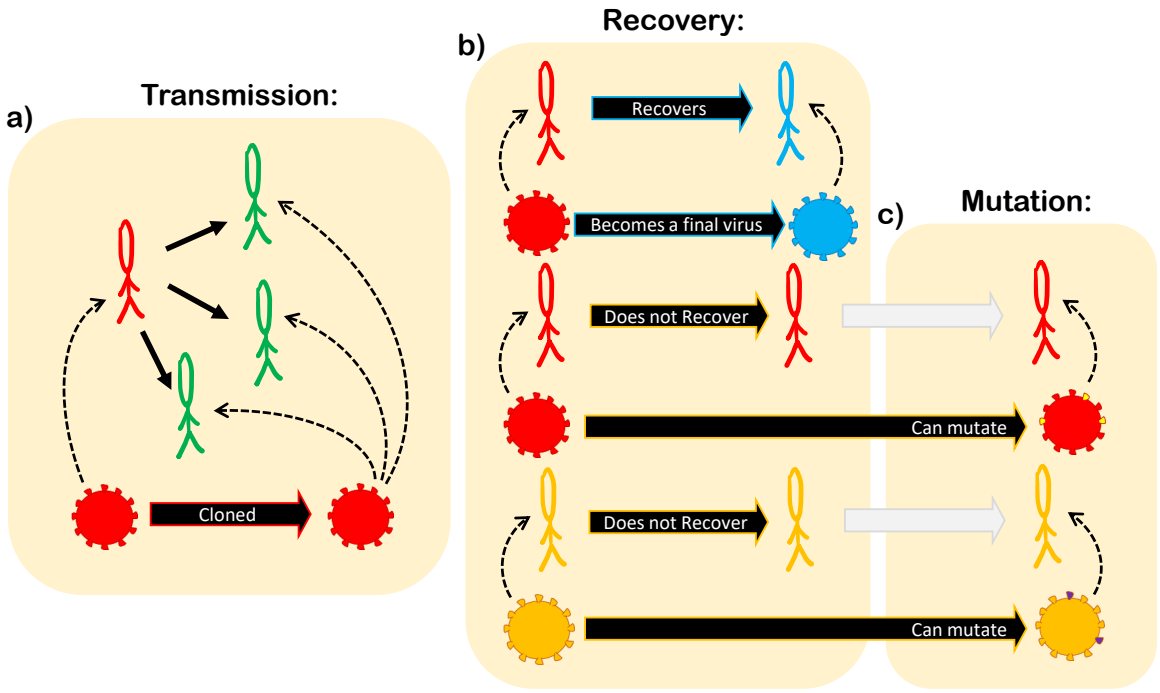


Figure 10.2: **Model dynamics.** *(a)* infected individuals (red) can transmit the virus to their susceptible first neighbors (green). When transmission is successful the virus is cloned to the new host, which is now an exposed individual (yellow) and will be able to mutate only in the next iteration. *(b)* infected individuals can recover with probability r. When an individual recovers (blue), its virus stops mutating and becomes a "final virus." *(c)* viruses on infected (red) or exposed (yellow) individuals can mutate.
Source: Figure from Marquioni and de Aguiar, 2021 [190].

10.3.3 The parameters

The *Basic Reproduction Number* R_0 measures the number of secondary infections emerging from a single infection in a completely susceptible population [22, 186]. It can be estimated by statistical methods and it is an important parameter of an epidemic, and it has been, for the COVID-19 pandemic, reported in many different studies [202]. A theoretical way of describing R_0 is

$$R_0 = p_I D \tau_{symp}, \tag{10.3.1}$$

where p_I is the transmission prabability, D is the average number of contacts of an individual in the population and τ_{symp} is the average duration of symptoms, (considered to be the period when individuals remain infectious). From this equation, given R_0, the transmission probability can be calculated as

$$p_I = \frac{R_0}{D\tau_{symp}}. \tag{10.3.2}$$

We may also consider a probability of recovery r per time unit

$$r \approx \lambda = 1/\tau_{symp}, \tag{10.3.3}$$

with the approximation valid for $\lambda \ll 1$. (The detailed calculation is shown in App. B.1). D is the average degree of the contact network. In the quarantine studies (which are going to be presented) we considered a scale-free network (Barabasi-Albert) in order to include heterogeneity.

We have conducted a simple study of the effect of network size and average degree on the final epidemic size (percentage of the population that was infected at the end of the epidemic) and on the average time to the infection curve peak. The results are shown in the App. B.2. From this investigation, to conduct further studies, we set networks with $N = 2000$ nodes and $D \approx 98$ (nominal $D = 100$).

For the incubation times, we considered the gamma distribution $\Gamma(\alpha, \beta)$ from Wu [203], with mean on 6.5 days and standard deviation 2.6 days.

The full list of parameters is shown in Table 10.1.

Table 10.1: **Simulation parameters.** The number of nodes in each simulation is described properly [189, 190].

Parameter	Value
R_0	2.4 [204]
Average Symptoms Duration τ_0	14 days [143, 146]
Networks Average Degree D *	100
Incubation Time Distribution $\mathcal{P}(\tau)$	$\Gamma(6.25, 25/26)$ [205, 203]
Mutation Rate μ	0.001 substitution per base, per year [206, 201]
Genome Size B	29900 bases [204]
For the Quarantine Results:	
Network size N	2000
Recovery probability r	1/14 (day^{-1})

*This is the input average degree for the network construction, but the actual value for each realization fluctuates. For the communities simulations, this is the parameter for constructing each isolated network, as also for the control case $p = 0$ [190].

10.4 The analysis

We have investigated epidemiological and evolutionary effects of the spread. During the COVID-19 pandemic, the lack of efficient treatments and vaccines forced governments to adopt quarantine and lockdown policies as measures of epidemic control [145, 164], slowing down the spread and preventing the overload of health systems [162]. We then studied the effect of quarantines in the course of an epidemic.

We have modeled a quarantine regime as a decreased transmission probability, reducing it by a factor $(1 - Q)$ in which Q is the quarantine intensity, varying from 0 to 1. The quarantine starts in day t_s and lasts for t_d days, and we present its results in the next chapter (Ref. [189]).

The pathogen evolution was also a very important feature during the COVID-19 pandemic, since the emergence of new viral strains was responsible for reinfection cases and

new infection waves [181]. Our goal was to understand how the contact network structure could shape viral diversity regardless of the effects of viral fitness differences acquired due to mutations and subsequent selection pressure [190].

We define the genetic distance between two viruses as the Hamming distance between the binary strings considering the nucleotides, i.e., the number of different nucleotides between them,

$$d^{\alpha\beta} = B - \sum_{i=1}^{B}(|b^{\alpha}_{2i-1} - b^{\beta}_{2i-1}|-1)(|b^{\alpha}_{2i} - b^{\beta}_{2i}|-1), \tag{10.4.1}$$

where α and β are different viruses and $b^{\gamma}_{j} \in \{0, 1\}$. The average genetic distance at a given time between all pairs of viruses (those that are still mutating (which we are going to call *active viruses*) and those that are not mutating anymore (*inactive viruses*)) is our measure of diversity.

We first study the viral diversity in networks without an explicit modular structure[1]. We show how one can analytically describe the evolution of the average genetic distance and apply our results to data collected in China at the beginning of the epidemic. Then, we consider modular networks as a simple model of a viral spread throughout different communities and show how the connectivity between close communities can shape the genetic variability. These results are also presented in the next chapter.

[1]A module of a network is a group of nodes with dense connectivity among each other but sparser connectivity with the other nodes of the network [207]. By networks "without an explicit modular structure" we mean Erdos-Renyi and Barabasi-Albert networks.

Chapter 11

Results and discussion: On quarantine regimes

The results we present in this chapter were published in two different papers [189, 190], whose "Results and Discussion" sections are being fully reproduced here.

11.1 Results and discussion

From Marquioni and de Aguiar, *Chaos, Solitons and Fractals*, 2020 [189].

Unlike the mean field SEIR model, Eqs.(10.2.1), the present IBM version on networks is probabilistic and different outcomes are obtained every time the model is ran with the same set of parameters. To obtain statistically significant data (while keeping simulation time reasonable) we have ran the model 25 times for different quarantine duration and intensities, beginning $t_s = 20$, 30, and 40 days after the first infected node appears (at the beginning of the simulation). The results were divided in two different scenarios, the *best* and the *worst* cases. For each set of parameters, the best scenario consists of simulations where the infection peak is lower than the average peak of the full set of simulations, whereas the worst scenario contains the set with higher than average peaks. This approach is important because in many cases the epidemic response to the quarantine is not satisfactory, and this might be solely due to stochastic effects, a common feature of real systems. As an example, Fig. 11.1 shows the evolution curves of infected plus exposed individuals for all 25 replicas for $Q = 0.9$ and $t_d = 10$ weeks. Since independent populations, represented by different Barabási-Albert networks generated with the same specifications, under the same quarantine parameters might respond drastically different to quarantine, we also need to know the probability of each outcome.

Figures 11.2, 11.3 and 11.4 show results for average peak height, time of infection peak and fraction of recovered individuals at the end of the epidemic (i.e., all individuals that had contact with the virus, as we do not take mortality into account). The results in each case are separated into best and worst scenarios and we compute the probability that a best scenario will happen. For example, a specific set of parameters might result in ending the epidemic, but its probability of occurrence can be too low, excluding it as

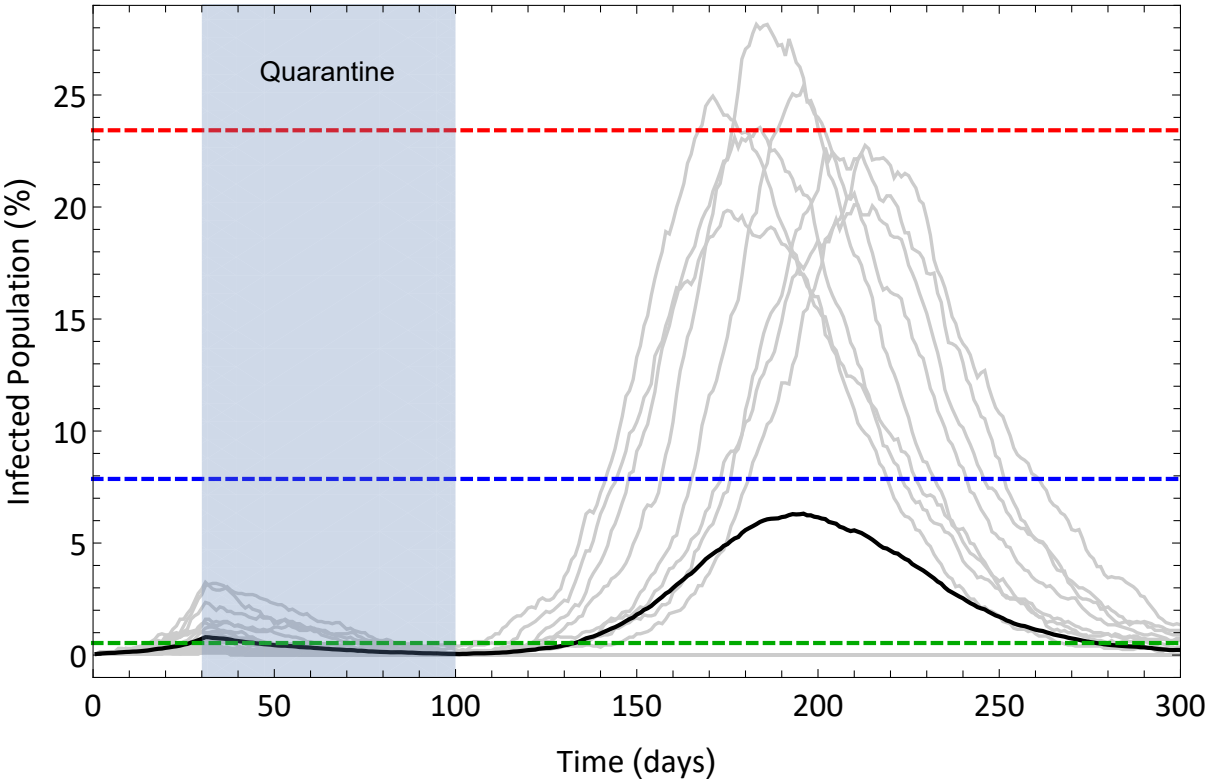


Figure 11.1: **Evolution of number of infected plus exposed individuals.** Evolution of number of infected plus exposed individuals for $Q = 0.9$, $t_s = 30$ days and $t_d = 10$ weeks for 25 replicas of the simulation. The blue dashed line shows the average height of the highest peak of each curve. Red and green dashed lines show the average peak of worst (8 replicas) and best (17 replicas) scenarios respectively, i. e., the average peak of the curves in which there is a second peak, after the quarantine, and the average peak of those in which there is not a second peak. The average of all curves (black thick line) is not representative of any actual curve. The gray shaded area indicates the quarantine period.
Source: Figure from Marquioni and de Aguiar, 2020 [189].

a recommended policy. All results are displayed as heat-maps.

Fig. 11.2 shows how peak height varies with quarantine duration, intensity and start date. This information is complemented by Fig. 11.3, that shows how peak center changes with quarantine parameters, and Fig. 11.4, displaying the proportion of recovered individuals at the end of the epidemic. The purple ellipse in Fig. 11.2 marks the parameter region where quarantine is very intense and lasts for more than 8 weeks, an ideal situation that works around 90% of the times but is very hard to enforce in practice. In this case the epidemic stops quickly (blue areas in Fig. 11.3) and less than 10% of the population is infected (green areas in Fig. 11.4).

The red ellipse in Fig. 11.2 shows a transition zone where the best scenario corresponds to substantial curve flattening. The center of the red ellipse is at $Q \approx 0.5$ for $t_s = 20$ but shifts to $Q \approx 0.9$ for $t_s = 40$, showing the importance of starting quarantine early. For all values of t_s the red ellipse is centered at $t_d \approx 6$ weeks, which is a relatively short duration. Peak center, however, is not delayed in the best case scenarios. Importantly, best case scenarios are very unlikely in this region, occurring with probability around 20%.

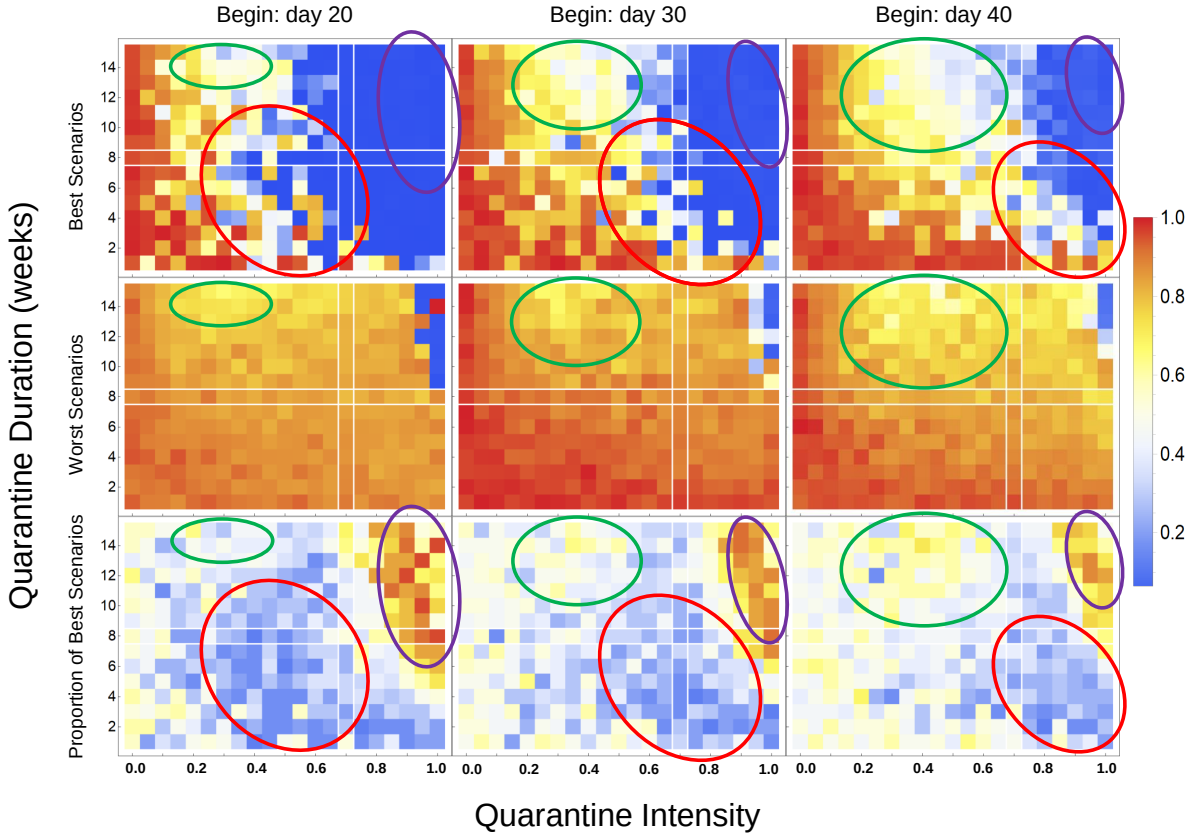


Figure 11.2: **Infection peak height heatmap.** Peak height with respect to the average 'no quarantine' result, starting 20, 30 or 40 days after the first infection (left, middle and right columns respectively). Plots in the first and second rows show the best and worst scenarios. The third row shows the probability that a simulation results in a best scenario. Quarantine duration is measured in weeks (from 1 to 15) and quarantine intensity goes from 0 (no quarantine) to 1 (full individual lock-down, $p = 0$). Green, red and purple ellipses highlight parameter regions of interest. White vertical and horizontal reference lines mark $Q = 70\%$ and $t_d = 8$ weeks.
Source: Figure from Marquioni and de Aguiar, 2020 [189].

Finally, the region surrounded by the green ellipse in Fig. 11.2 corresponds to long but moderate intensity quarantines. For the three values of t_s considered peak height was reduced by about 50% in the best case scenarios, which happens about 50% of the times. Peak center was not significantly delayed in the best scenarios, but was pushed forward in the worst scenarios, where peak height was reduced to about 70% with respect to non-quarantine height. Interestingly, in both scenarios about 70% of the population was infected at the end of the simulation, showing that herd immunity was achieved (corresponding to the pink areas in Fig 11.4).

Quarantine can also be implemented in the mean field model, Eqs. 10.2.1.[146] This is accomplished by integrating the dynamical equations with the infection rate β_0 for $t \in [0, t_s]$, with the reduced value $\beta_Q = (1-Q)\beta_0$ during quarantine period $t_s < t < ts+t_d$ and again with β_0 for $t > t_s + t_d$. Fig. 11.5 shows how results of mean field model differ from the IBM simulations. Panel (a) shows the dynamics without quarantine according to the mean field (thick lines) and 25 simulations with the IBM. Panel (b) shows the

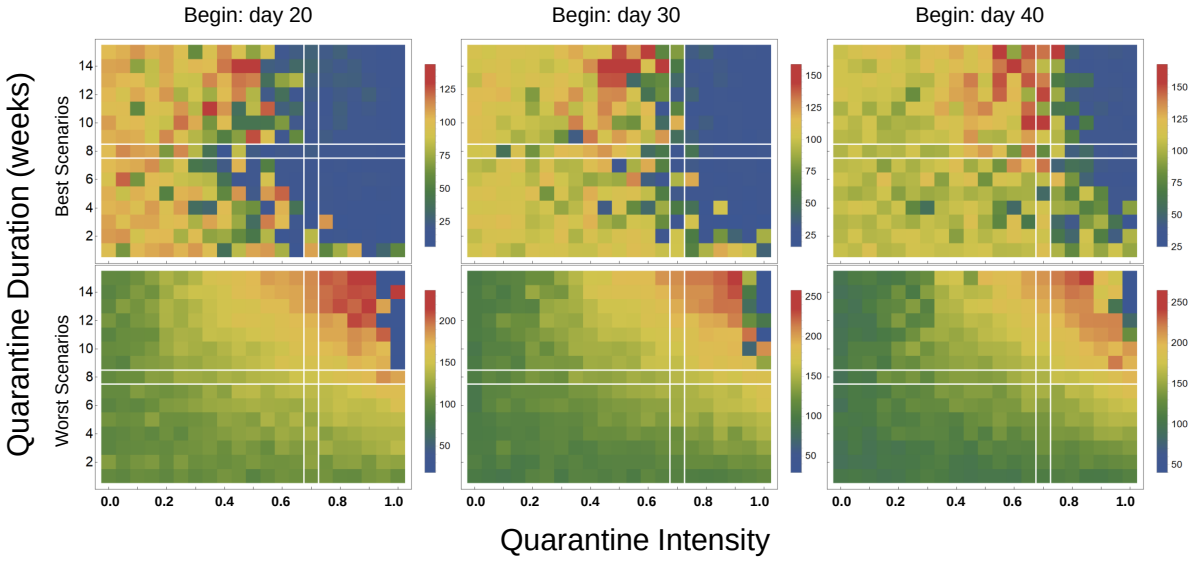


Figure 11.3: **Infection peak time heatmap.**Peak center (in days after the first infection) starting 20, 30 or 40 days after the first infection (left, middle and right columns respectively) for different quarantine intensities for best and worst scenarios.
Source: Figure from Marquioni and de Aguiar, 2020 [189].

effects of quarantine on the mean field model for $Q = 0.35$, $t_d = 10$ weeks and several starting times t_s. According to the mean field model quarantine is effective only if started later, otherwise the infection curve peaks at high values when the quarantine is over. The right panels compare IBM simulations (c) and mean field results (d) for $t_s = 30$ days and $t_d = 15$ weeks for several quarantine intensities Q. The mean field infection curves always grow to high values when quarantine is over, whereas the IBM simulations show many examples of low peak values and total epidemic control, with $I + E$ going to zero after the quarantine period. This highlights the importance of heterogeneous social interactions represented by the Barabási-Albert network and stochastic dynamics in epidemiological modeling.

11.2 Conclusions

In this paper we considered the effects of quarantine duration, starting date and intensity in the outcome of epidemic spreading in a population presenting heterogeneous degrees of connections. The model is stochastic and curves representing numbers of infected individuals vary considerably from one simulation to the other even when all model parameters are fixed. In order to distinguish between different outcomes we have divided them into two groups with the best and worst results based on the height of the infection peak (below or above the average height, respectively).

We have further divided the results into four qualitative classes delimited by the three ellipses in Fig. 11.2 plus the rest of the diagram. Besides the obvious region indicated by the purple ellipse where quarantine is very intense and long, we found that short but not so intense quarantine (red ellipse) does not work, since the probability of an outcome in the best scenario is very low. Instead, long but average intensity quarantine is both

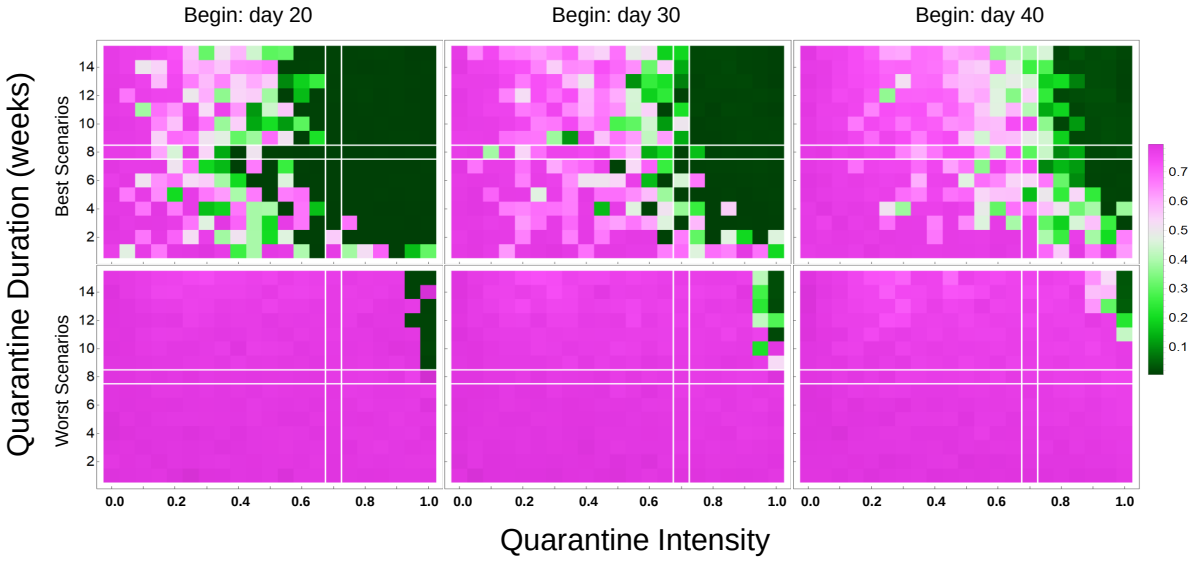


Figure 11.4: **Epidemic size heatmap.** Proportion of recovered individuals at the end of the epidemic for quarantine starting 20, 30 or 40 days after the first infection (left, middle and right columns respectively) for different quarantine intensities for best and worst scenarios.
Source: Figure from Marquioni and de Aguiar, 2020 [189].

likely to work and flattens the infection curve by around 50%, being the best alternative given the current assumptions. Indeed, the infection peak is considerably delayed in the region of the green ellipse when it falls into the worst scenario, confirming it as the best bet for preventing the health system breakdown (Fig. 11.3). The proportion of the population that had contact with the virus at the end of the epidemic (number of recovered individuals, Fig. 11.4) leads to more than 60% of the population, very close to achieving herd immunity. Comparing to the other regions, this seems to be the best option to control the epidemics under the model assumptions. We note, however, that the model does not account for deaths. If achieving herd immunity implies high mortality, the best option would be long and intense quarantine (purple ellipses in Fig. 11.2), the only way to avoid large number of infections and, therefore. high mortality.

We found that differences between mean field and stochastic models are very significant with respect to the effects of quarantine. In many cases the former cannot control the epidemic, as the infection peak grows again once the quarantine period is over, whereas the latter can end the epidemic in the best case scenarios. Morris *et. al* have investigated the optimal quarantine parameters for the mean field SIR model,[146]. For the strategy we designed in the present work (Q constant during the quarantine period $[t_s, t_s + t_d]$), they show the existence of an optimal value for t_s and Q for a given t_d, leading to the minimization of the infection peak. In cases where the infection curve shows a second peak, it reaches the height of the first peak, as illustrated by "Day 90" curve in Fig. 11.5 (b). Fig. 11.6 compares this curve with our model under the same quarantine parameters. The heterogeneity of network structure moves the peaks to earlier times, decreasing considerably the effect of the quarantine. In this case the mean field description is not a good approximation for the dynamics and the optimal solution[146] is not applicable without further adjustments.

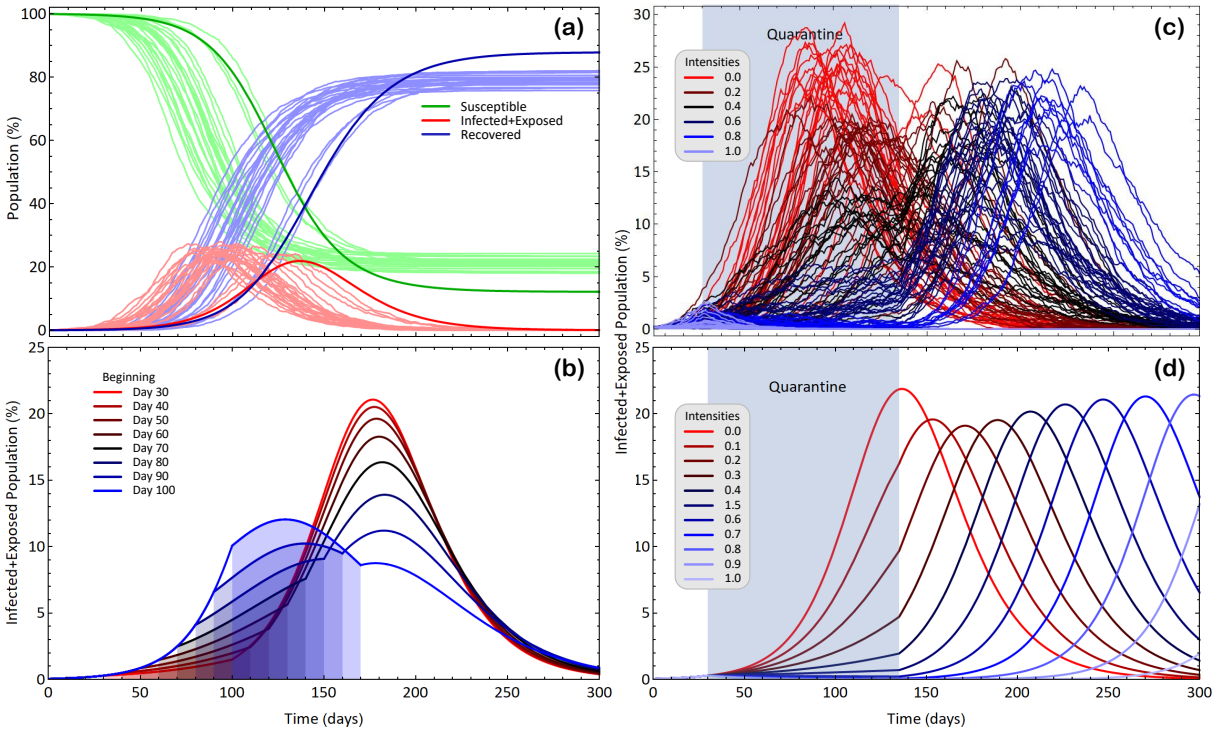


Figure 11.5: **Comparison between mean-field and network dynamics.** (a) dynamics without quarantine computed with mean field equations (thick lines) and IBM simulations (thin lines); (b) mean field results with $Q = 0.35$, $t_d = 10$ weeks and several starting dates t_s; (c) 25 IBM simulations and (d) mean field dynamics for $t_s = 30$ days, $t_d = 15$ weeks and several intensities Q. For the mean field equations, we set $N = 2000$, $\beta = R_0\gamma$, $\gamma = 1/14$, $\sigma = 1/\langle t_i \rangle$, and starting with one infected individual.
Source: Figure from Marquioni and de Aguiar, 2020 [189].

We recall that we used uniform decrease in infection rate as a proxy for quarantine. This is a simplified approach and other methods could be implemented to verify the robustness of the results. Also, different network topologies might affect the spread of the epidemics. Random uniform (Erdos-Renyi) [95] networks should produce results similar to mean field simulations, but small-world [113, 95] or other topologies could speed up or slow down the spread dynamics.

Our model is particularly suited to study spread between connected cities, that can be represented by modules of a larger network. We have also kept information about the virus DNA and its mutations, allowing us to reconstruct the phylogeny and classify its strains as it propagates. These results will be published in a forthcoming article.

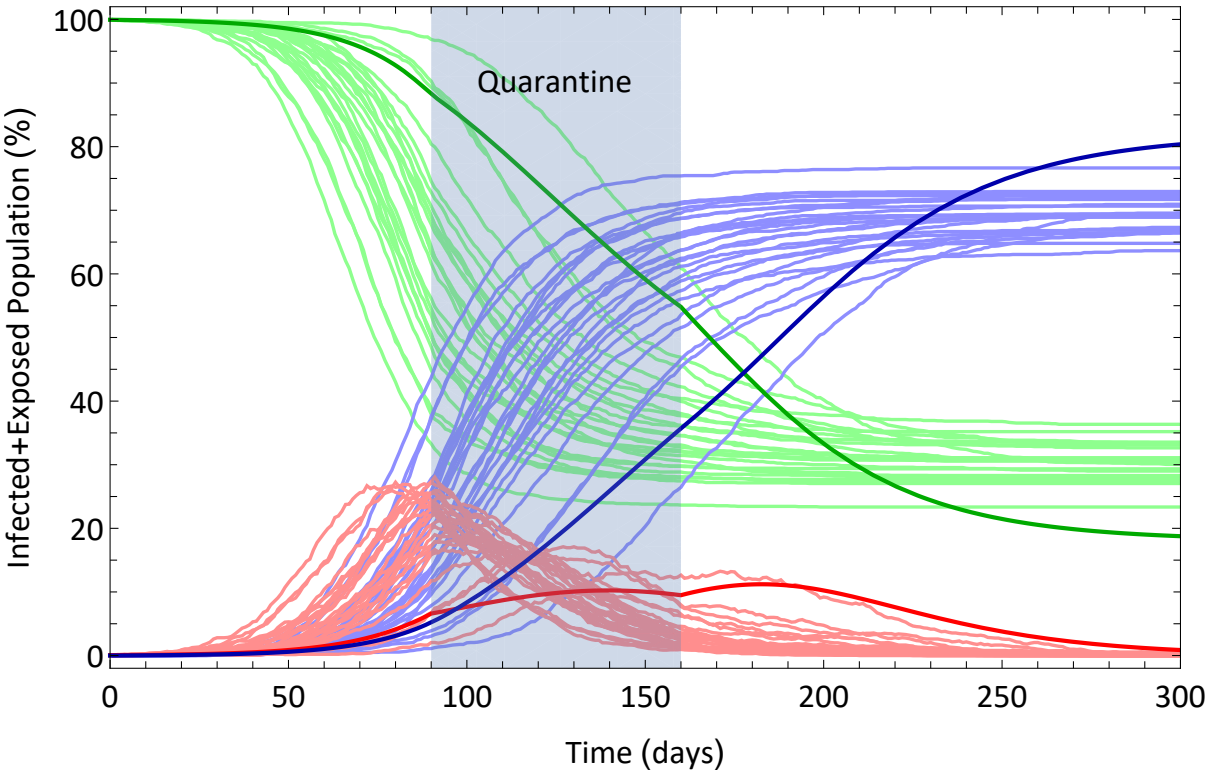


Figure 11.6: **Comparison of quarantine in mean-field and network dynamics.** Comparison between the mean field solution (thick lines) and our model (25 replicas, thin lines) under the same quarantine parameters ($Q = 0.35$, $t_s = 90$ days and $t_d = 10$ weeks). Colors represent: susceptible (green), infected plus exposed (red) and recovered (blue). Because most infection curves peak before the 90^{th} day, quarantine has little effect. In the mean field model, on the other hand, the infection peak is substantially reduced. Source: Figure from Marquioni and de Aguiar, 2020 [189].

Chapter 12

Results and discussion: On viral diversity

From Marquioni and de Aguiar, *PLoS ONE*, 2021 [190].

12.1 Analytical description

The analysis presented here to calculate the average genetic distance between all viruses, living and final, is suitable for compartmental models in general [22]. Although we develop it to the SEIR model, it can be applied to other models of this type. From now on we shall abbreviate *average genetic distance* by *average distance* for simplicity.

12.1.1 Single initial infection

Here we assume that the epidemic starts with a single infected individual. Our goal is to compute the average distance d_{t+1} at time $t+1$ given the average distance d_t at time t. Notice that at the *beginning of iteration* $t+1$, there are different *kinds* of viruses: those that are *already final* and have ceased to evolve (whose number is R_t); *viruses hosted* in exposed individuals (E_t), thus still evolving; and also those *hosted in infected* individuals (I_t). During the iteration, *new infections* appear (x_t) and some *infected individuals recover* (r_t), and thus do not evolve at this time step. Then, given d_t, we calculate the new average distance *between each kind of virus* which exists at the *end of iteration* $t+1$, as well as the new average distance *within each kind of virus*.

Given that $\mu \ll 1$, we consider that the probability that two mutations happen in the same nucleotide in the course of the epidemic is negligible. This is a good approximation if the epidemic duration T remains sufficiently small, $\mu T \ll 1$. We also consider that each new infection in the same iteration comes from different hosts, which is valid for $R_0/\tau_0 < 1$, with τ_0 the average duration of symptoms. This means that we do not expect more than one new infection per infected individual in a single iteration. Highly connected nodes, however, can break this assumption, giving rise to super-spreaders. Network heterogeneity, therefore, can show deviations from our estimation. Under these assumptions, the new average distance (at the end of iteration $t+1$) among the E_t is

$d_t + 2B\mu$, once they distanced d_t at the begging of iteration $t+1$ and evolved along the iteration, each virus getting $B\mu$ mutations. The new average distance between the E_t and the R_t is $d_t + B\mu$, since only the E_t evolved. We emphasize that the approximations used in this section are only for simplification of the analytical equations; the simulations in Section 12.3 run as previously described.

Once all average pairwise distances have been calculated, d_{t+1} is given by a weighted average, where the weigths are the number of pairs sharing that distance. For instance, the number of pairs between exposed and recovered individuals is $E_t R_t$, while the number of pairs within exposed individuals is $E_t(E_t-1)/2$.

All distances are calculated in appendix C.1, and we find the recurrence equation

$$\begin{aligned} d_{t+1} = \frac{1}{Z_t} \big(& d_t(R_t+E_t+I_t)(R_t+E_t+I_t-1) \\ & + x_t d_t \left(1 + 2B\mu \frac{R_t}{I_t+E_t+R_t}\right)(x_t - 3 + 2R_t + 2I_t + 2E_t) \\ & + 2B\mu(E_t+I_t-r_t)(E_t+I_t+R_t+x_t-1)\big) \end{aligned} \tag{12.1.1}$$

where $Z_t = (R_t+E_t+I_t+x_t)(R_t+E_t+I_t+x_t-1)$, $r_t = R_{t+1} - R_t$ and $x_t = (E_{t+1} - E_t) + (I_{t+1} - I_t) + (R_{t+1} - R_t)$.

Therefore, given the epidemic curves S_t, E_t, I_t and R_t, respectively the Susceptible, Exposed, Infected and Recovered at time t, we can infer the evolution of average genetic distances. Taking the limit of continuous time between events we find the approximation,

$$\dot{d} = \frac{2\dot{S}d\left(1 - B\mu R\left(2 - \frac{3}{N-S}\right)\right)}{(N-S)(N-1-S)} + 2B\mu\left(1 - \frac{R}{N-S}\right) \tag{12.1.2}$$

where $N - S = I + R + E$ and $\dot{S} = -(\dot{E} + \dot{I} + \dot{R})$. The derivation of this limit is described in App. C.1. Since this equation depends only on the continuous curves $S(t)$ and $R(t)$, the initial and final compartment, it can be added to the classic SEIR model to infer the genetic evolution, or to the SIR model, if the exposed compartment is kept empty, meaning that all hosts are infectious. This result holds if viral evolution occurs in the same way in every intermediate compartment and if every virus passes through all compartments. Adding more compartments with different dynamical behavior or changing the mutation mechanism through different compartments would change the equations (12.1.1) and (12.1.2) but the procedure described in the begging of this section to find d_{t+1} should remain the same.

12.1.2 Multiple initial infections

Eq.(12.1.1) considers the epidemic starting with a single infected individual. To consider $m > 1$ initial infections, we must include the distance among the m different lineages. Let $\mathfrak{D}_t$ be the average distance among all viruses at time t, $d_t^{(i)}$ the average distance among the viruses of lineage i at time t, $d_0^{(ij)}$ the distance between the initial viruses i and j, and

$d^{(i)}_{root,t}$ the average distance at time t of lineage i to the root of lineage i. Thus,

$$\begin{aligned}\mathfrak{D}_t = &\left[\sum_{i=1}^{m} d_t^{(i)} \left(R_t^{(i)} + E_t^{(i)} + I_t^{(i)}\right)\left(R_t^{(i)} + E_t^{(i)} + I_t^{(i)} - 1\right)/2\right. \\ &\left. + \sum_{i=1}^{m-1}\sum_{j=i+1}^{m} \left(d_0^{(ij)} + d^{(i)}_{root,t} + d^{(j)}_{root,t}\right)\left(R_t^{(i)} + E_t^{(i)} + I_t^{(i)}\right)\left(R_t^{(j)} + E_t^{(j)} + I_t^{(j)}\right)\right] \\ &\div \left[\left(\sum_{i=1}^{m}\left(R_t^{(i)} + E_t^{(i)} + I_t^{(i)}\right)\right)\left(\sum_{i=1}^{m}\left(R_t^{(i)} + E_t^{(i)} + I_t^{(i)}\right) - 1\right)/2\right] \end{aligned} \tag{12.1.3}$$

where $R_t^{(i)}$, $E_t^{(i)}$ and $I_t^{(i)}$ are, respectively, the number of recovered, exposed and infected individuals of lineage i at time t. The first sum represents the distances within each lineage i, while the double sum is due to the distance between each pair of lineages i and j. In this equation, we assume the $\mu \ll 1$ (for coronaviruses, μ lies in the range $\sim [10^{-5}, 10^{-2}]$ per site per year[201]) so that mutations for each virus are unlikely to occur twice at the same nucleotide.

For each lineage i, $d_t^{(i)}$ can be calculated from Eq.(12.1.1) or Eq.(12.1.2) and $d_0^{(ij)}$ must be a given matrix. The distance $d^{(i)}_{root,t}$ can be calculated similarly as Eq.(12.1.1),

$$d^{(i)}_{root,t+1} = d^{(i)}_{root,t} + \frac{B\mu}{E_t^{(i)} + I_t^{(i)} + R_t^{(i)} + x_t^{(i)}}\left(E_t^{(i)} + I_t^{(i)} - r_t^{(i)} + \frac{4x_t^{(i)} R_t^{(i)} d^{(i)}_{root,t}}{E_t^{(i)} + I_t^{(i)} + R_t^{(i)}}\right) \tag{12.1.4}$$

with the continuum limit

$$\dot{d}_{root} = B\mu\left[1 - \frac{R^{(i)}}{R^{(i)} + I^{(i)} + E^{(i)}}\left(1 - \frac{4d_{root}(\dot{E}^{(i)} + \dot{I}^{(i)} + \dot{R}^{(i)})}{R^{(i)} + I^{(i)} + E^{(i)}}\right)\right] \tag{12.1.5}$$

where $R^{(i)}$, $I^{(i)}$ and $E^{(i)}$ are SEIR variables for lineage (i). The details behind these results are described in App. Analytical calculations.

12.2 Viral spread throughout communities

As an application of our model and computational framework, we studied the genetic evolution of a viral spread throughout four weakly and linearly connected communities, i.e., a network with four modules, representing different cities. The goal is to understand how the average genetic distance between viruses in distant communities change if the connectivity between the intermediary communities changes.

We start by generating four independent Barabasi-Albert networks, named 1, 2, 3 and 4. Then, we connect individuals from networks i and $i+1$ with a *connection probability* p in a way they form a line of communities. The Barabasi-Albert network is chosen in order to include heterogeneity in the contact network [189]. Finally, we analyse the average genetic distance between viruses from cities 1 and 4 for different values of p. The epidemic starts with a single infected individual in city 1 and spreads through the entire network.

Although in our model we always consider that individuals acquire perfect cross-immunity against all strains after being infected the cross-immunity could in principle be lost if a new infecting virus were too different from the original infection. Thus, if the distance between viruses from cities 1 and 4 is large, an infected individual from city 4 that travels to city 1 might reinfect an already recovered individual. Although our simulations do not include this possibility, this is an interesting way to investigate how the risk of reinfection changes due to changes in the network topology.

12.3 Results and discussion

12.3.1 Single initial infection

We ran our model for random (Erdos-Renyi) and scalefree (Barabasi-Albert) networks and calculated the average genetic distance. We used networks of 200, 500, 1000 and 4000 nodes, and average degree D of 100 nodes, which was the same for all simulations. In the range of parameters we have used, changing the average degree has two main consequences. First, for large values ($D \gg R_0$), the deviations around the mean of many simulations decreases; and secondly, once the probability of infection is proportional to $1/D$, increasing D delays the peak of infection. We note that the greater the number of connections, the greater the number of attempts to infect neighbors within a single iteration. Thus, we have chosen a value of D that produces reasonably small deviations around the mean and, at the same time, enables fast computation. Changing D in the interval 50 to 200 resulted in no qualitative changes. The infection starts with a single infected individual chosen at random and evolves according to the description in section 10.3. Fig.12.1 shows comparisons between the simulated distance and the average distance calculated from Eq.(12.1.1) and Eq.(12.1.2). Each subfigure contains two different simulations and the mean-field solution for that respective set of parameters. We see that Eq.(12.1.2) approaches Eq.(12.1.1) only for Erdos-Renyi networks, since only this topology mimics the well-mixed hypothesis considered in mean-field models. Because each genetic evolution curve is calculated from the corresponding epidemic curves, we cannot average over many simulations, thus the error bars are simply the standard deviation of the distribution of distances among all viruses that appeared at that specific simulation time step. Another important feature of this analytical formulation is that, once it is an average description, it does not capture the random appearance or extinction of viral lineages, which can introduce important deviations from our analytical description.

12.3.2 Multiple initial infections

Fig.12.2 shows the evolution of epidemic in two different cities (non-connected networks of random and scalefree types), each one starting its infection with a single infected individual chosen at random. The evolution in each city is calculated with Eq.(12.1.1) (pink curves), while the distance between cities 1 and 2 is $d_t^{(1,2)} = d_0^{(1,2)} + d_{root,t}^{(1)} + d_{root,t}^{(2)}$, where $d_{root,t}^{(i)}$ is calculated with Eq.(12.1.4) (red curve) and the total average distance $\mathfrak{D}_t$ (green curve) is given by Eq.(12.1.3). The initial distance between the viruses that infected each city is $d_0^{(1,2)} = 0$ in panels (a) and (b), and $d_0^{(1,2)} = 5$ in panels (c) and (d).

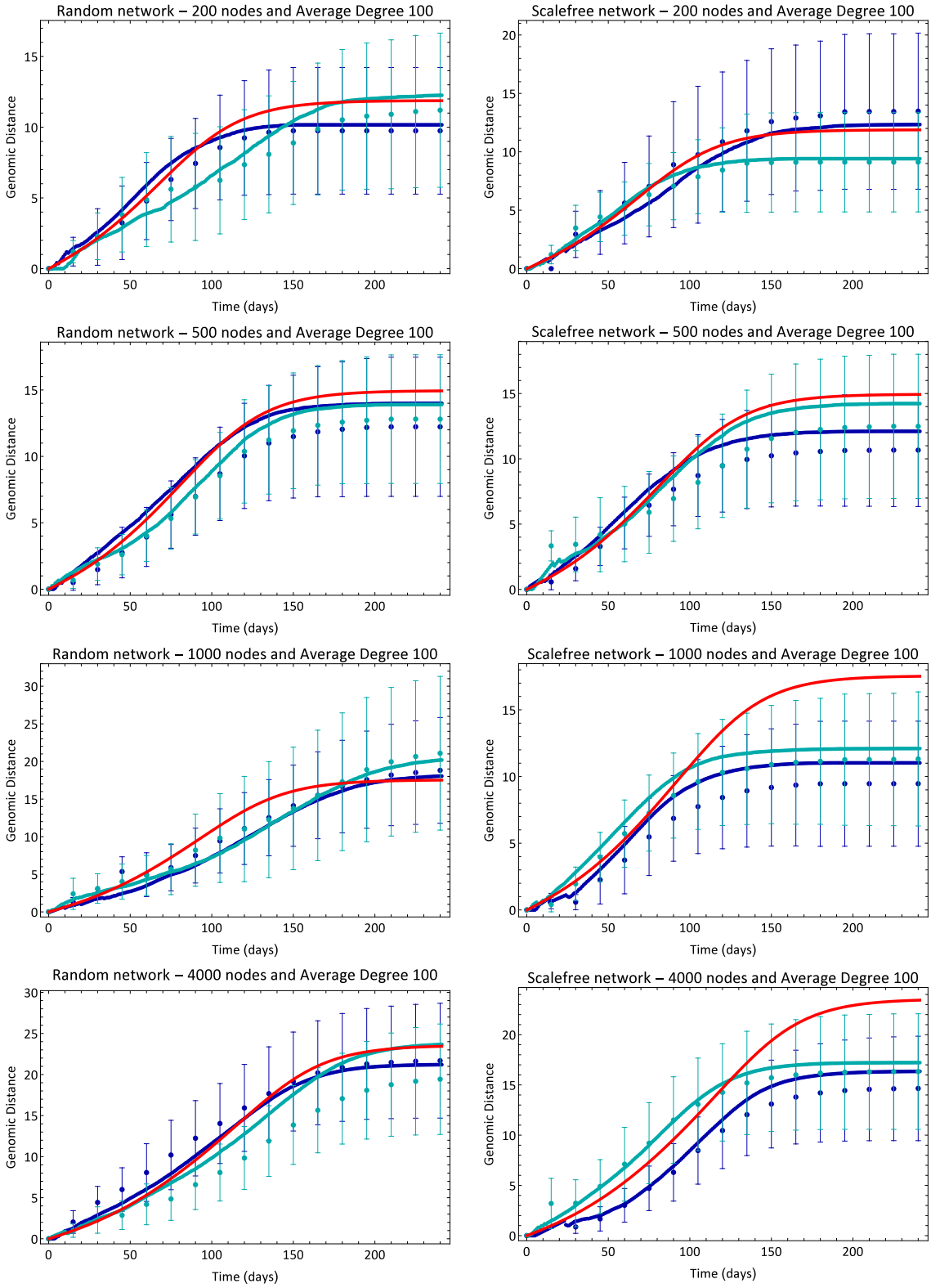


Figure 12.1: **Evolution of average genetic distance.** Blue lines and dots are, respectively, analytical (Eq.(12.1.1)) and simulation results for different simulations. Different shades of blue correspond to different simulations for the same set of parameters. The red line shows the result of mean-field Eq.(12.1.2). Error bars are standard deviation of the distance distribution in each simulation at each time.
Source: Figure from Marquioni and de Aguiar, 2021 [190].

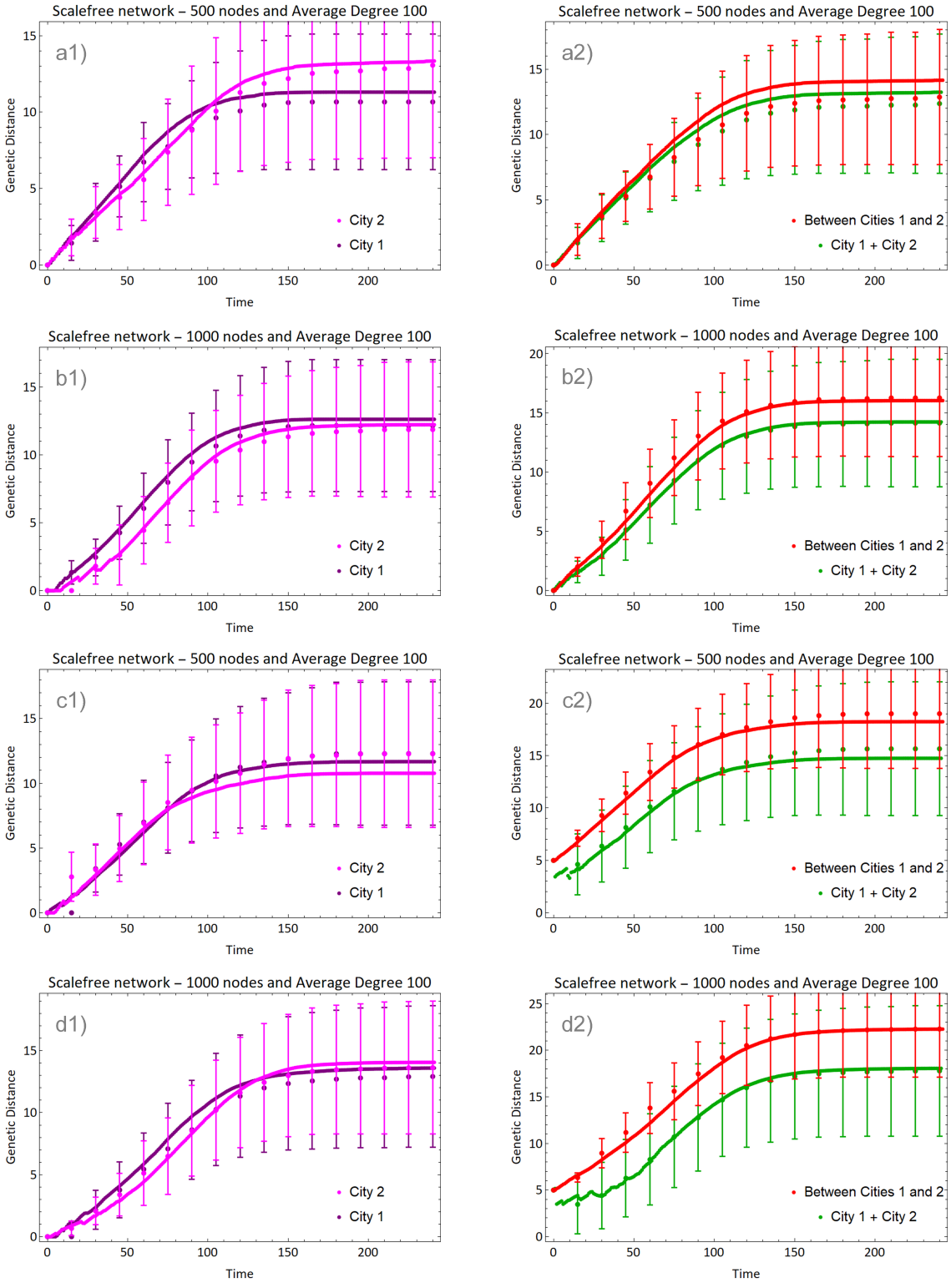


Figure 12.2: **Evolution of average genetic distance in two isolated cities (sizes indicated in the panels).** In (a) and (b) the initial viruses were identical and in (c) and (d) they differed by 5 nucleotides. Lines show the average distance within each city (pink), between cities (red) and total average distance (green).
Source: Figure from Marquioni and de Aguiar, 2021 [190].

12.3.3 The COVID-19 epidemic in China

Eq.(12.1.1) describes the evolution of average genetic distance between viruses in a single community and depends only on the epidemic curves. It might, therefore, be used to estimate the genetic evolution in real cases. The beginning of COVID-19 epidemic in China is a suitable example, considering the existence of a single patient zero. In any other country, the epidemic may have started with more than one individual, which would require the difficult task of tracking the lineages. The same applies to secondary waves of infection in China.

We obtained Chinese data from the Wolfram Data Repository[208], and corrected it as in reference [144]. Because of the existence of undetected cases, we estimated the real number of cases considering references [144, 209]. Because the number of exposed individuals is not directly available we choose to consider the simpler SIR model in this case. Notwithstanding, because the cases notification started only in January while the epidemic started in December, we extrapolated the data to previous dates, in order to calculate the genetic evolution since patient zero, as we have made in Fig.12.1. All these data corrections and considerations are described in the supporting information.

To compare the result of Eq.(12.1.1) with the real genetic evolution, we used carefully selected 55 real genomes sequenced and collected in China, also available in the Wolfram Data Repository[210]. The Hamming distance between each pair of genomes was obtained by first aligning every two genomes with the Needleman-Wunsch algorithm with score matrix $+1$ for match and -1 for mismatch[211]. Then, we considered the Hamming distance between a given pair of genomes as the number of mismatches that are not *indels*, i.e., we considered only nucleotide substitutions. The algorithm to estimate the distance evolution is explained in App. C.1, as we also detail the informations of the used genetic data.

Fig.12.3 shows the result obtained from Eq.(12.1.1) (brown line) and the estimated genetic evolution (blue dots). The interval around the brown line is an error of $\pm 10\%$ on the product μB, which is the only parameter in the equation (12.1.1). Despite all corrections to the epidemic data and the small number of real genomes we used to infer the real genetic evolution, except for a few points, all the inferred average genetic distances between RNA sequences lie in the predicted interval given by our theoretical model. Because the epidemic in China was readily contained, the average distance d_t saturated.

12.3.4 Communities and reinfection

In this section, we consider the spread of the epidemic through four communities, representing cities, connected linearly as in Fig.12.4. The connections within each network are of Barabasi-Albert type, with 1000 nodes and average degree 100 (following the same considerations on average degree already mentioned). Every node from network i can be connected to a node in netowrk $i+1$ with connection probability p. Once p is small (ranging from 0.0005 to 0.0035) the degree distribution is not considerably distorted from a scale-free one. Fig.12.4 shows an example of the contact network. From left to right, we number the communities, or cities, from 1 to 4. The epidemic starts with a single infection in city 1 and spread through the entire network. Fig.12.4 also shows the Infection curves obtained from a simulation. The infection peak delay from one city to other is responsible for the plateau-type curve of total infections.

To analyse the genetic evolution in this system we simulated the dynamic until the epi-

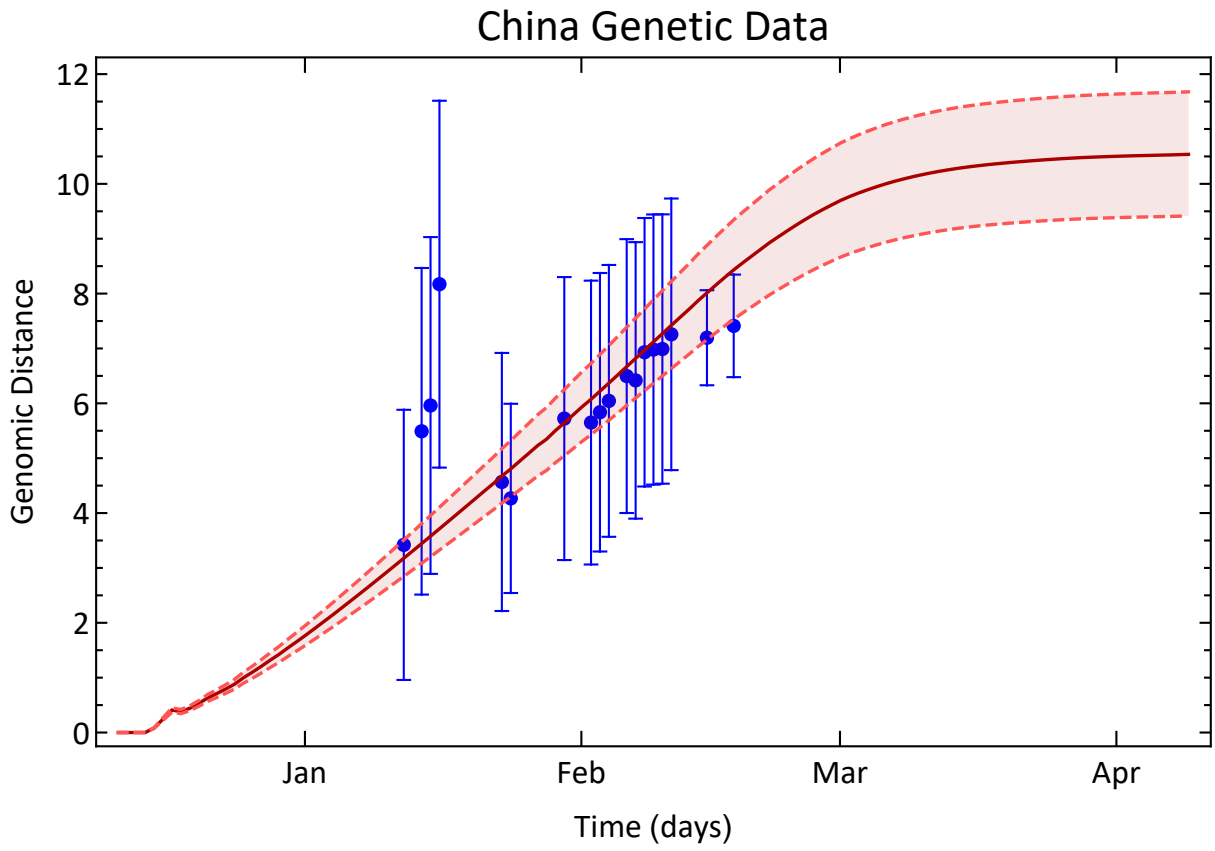


Figure 12.3: **The genetic evolution of SARS-CoV-2 in China.** Blue dots are the genetic distance among SARS-CoV-2 inferred from data collected in China between 12/23/2019 and 03/24/2020. The error bars are standard deviation of pairwise distance propagated through the equations. The brown line shows the genetic distance estimated with Eq.(12.1.1) and the Chinese epidemic data. The interval around the brown curve is a $\pm 10\%$ error interval on the value $B\mu$, which we considered to be $B\mu = 29900 \times 0.001/365$. Source: Figure from Marquioni and de Aguiar, 2021 [190].

demic was over and calculated the Hamming distance between every pair of final genomes α and β, constructing the distance matrix $d^{\alpha\beta}$ (Fig.12.5). Viruses are ordered according to their position in the line, i.e., first the genomes from city 1, then those from the city 2, and so on. We calculated the average distances D_{i-j} between the final genomes from cities i and j and compared with D_{i-i}, the average distance within city i.

As a null model, we run the epidemic over a single Barabasi-Albert network wih the total size of the 4 cities. City i, in this case, means the i-th quarter of the infected nodes. We plot the results of the null model as $p = 0$ in Fig.12.6 and Fig.12.7 for comparison. The single network behaves very differently from the four module network, not showing the same interesting results we find for the communities.

Fig.12.6 shows the ratio D_{4-4}/D_{4-1} as a function of the connection probability p. The results are averages over 20 different simulations for 7 different values of p. When p is small, $D_{4-4}/D_{4-1} < 1$, meaning that the viruses from city 4 are, in average, closer to each other than they are to the viruses from city 1. When p increases, the ratio D_{4-4}/D_{4-1} approaches 1, indicating that the viruses from city 4 are so close to each other as they are to viruses from city 1.

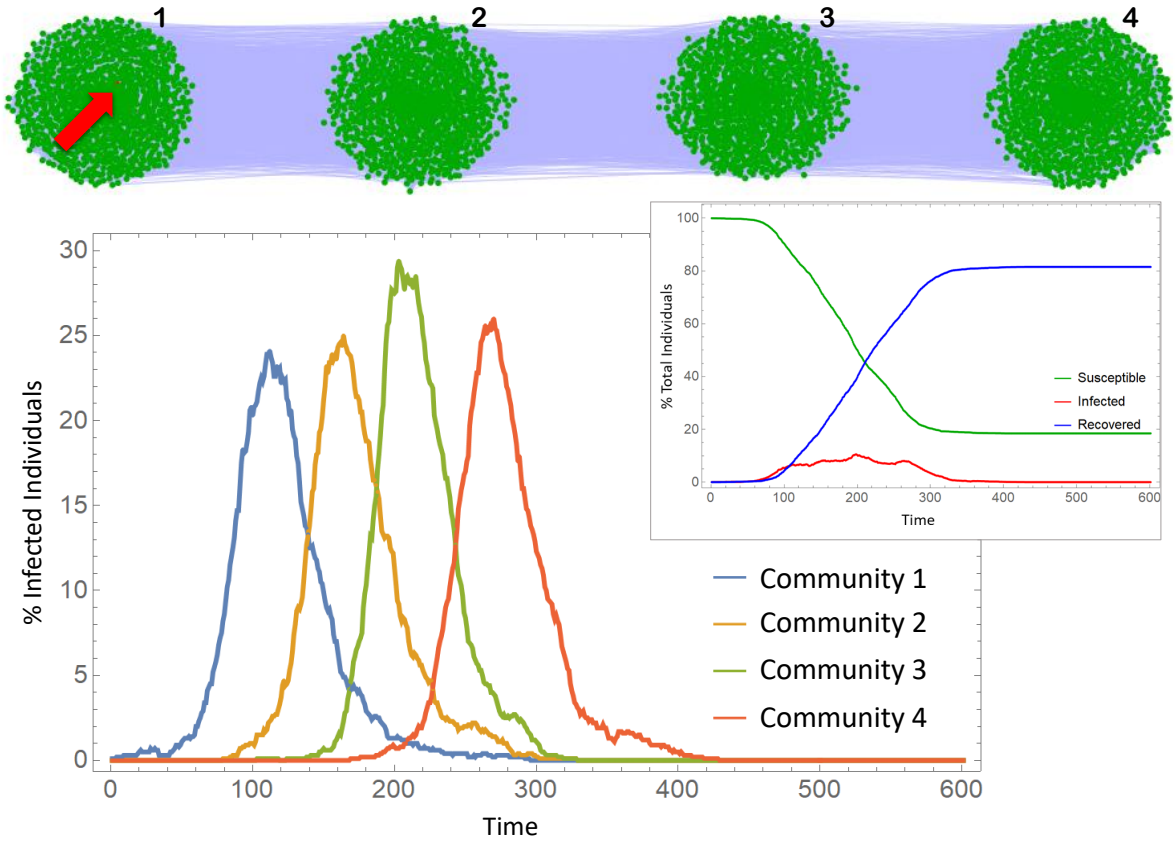


Figure 12.4: **Contact network of four communities on a line and infection curves.** Communities are Barabasi-Albert networks with 1000 nodes. We have kept the average degree constant and equal to 100 in all simulations. The infection starts with a single infected individual in the first community (red node indicated with the red arrow). The epidemic parameters are in Table 10.1.
Source: Figure from Marquioni and de Aguiar, 2021 [190].

In order to understand the origin of this effect we analyse the infection trees in each case (Fig.12.6, left). Each node in the trees represents a recovered individual and is connected upwards with whoever infected it. Colors represent cities and it is possible to count how many initial infections each city had along the epidemic, i.e., how many lineages has infected each city. When p is small, very few lineages were responsible for infecting city 4 but for higher values of p, this number increases. This is expected, since more connected communities should have more infection gates. This result is a consequence of the founder effect, i.e., only a few individuals, "the founders", give rise to a new population in the new location[212, 213]. However, the system passes through a non-trivial bistable point. When $p = 0.0015$, the values of D_{4-4}/D_{4-1} accumulate around two different values, one above 1 and another below 1. In this case the average is not a good descriptor of the actual system behaviour and there is a competition between different lineages infecting city 4. In simulations where $D_{4-4}/D_{4-1} > 1$, many lineages were successful in infecting the city 4, whereas when $D_{4-4}/D_{4-1} < 1$, only a few did so successfully.

Fig.12.7 shows the values D_{4-4} and D_{4-1} obtained in each simulation. The average over simulations of the average distance within the forth city D_{4-4} (highlighted blue

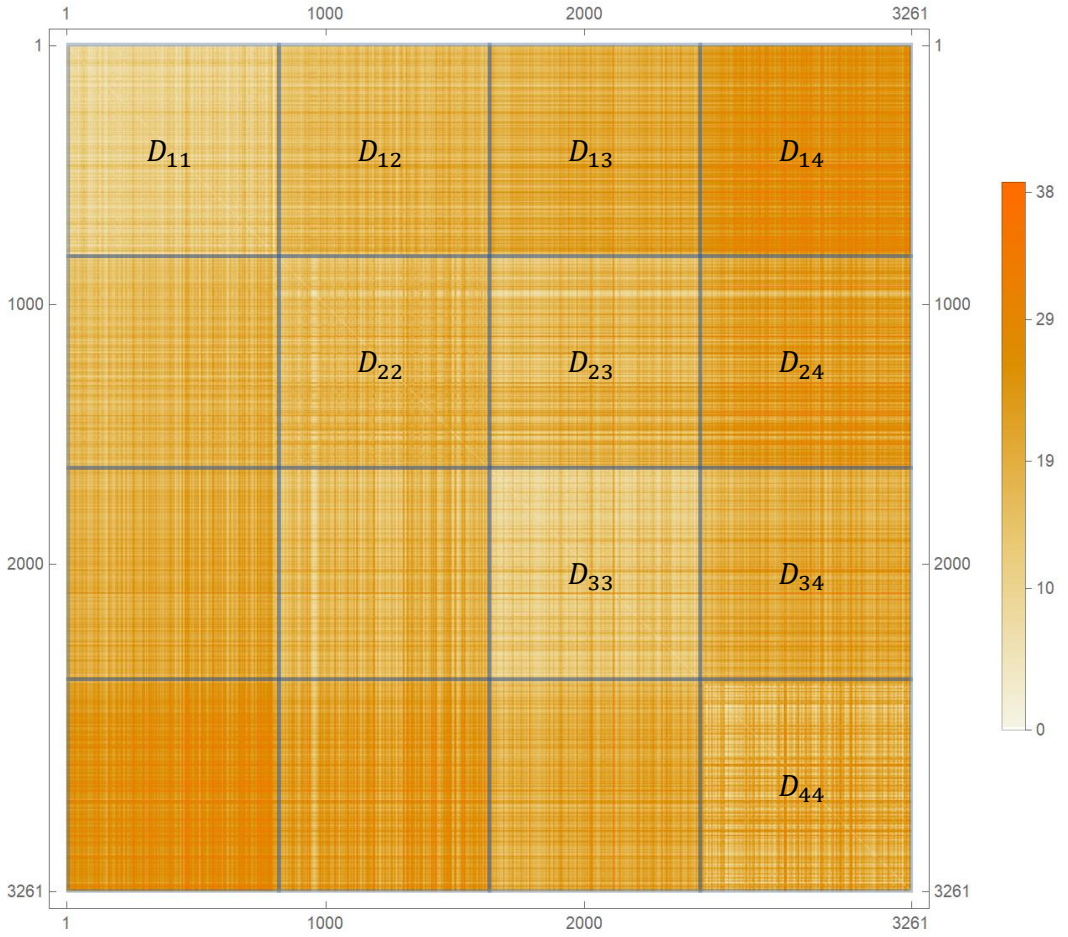


Figure 12.5: **Hamming distance between pairs of viruses.** The distance matrix is sorted by the city. Diagonal blocks show the distance between the viruses from a single city, while the non-diagonal blocks are the distances between the viruses from different cities.
Source: Figure from Marquioni and de Aguiar, 2021 [190].

circles) does not change considerably with p (around $D \approx 21$ nucleotides). Under a neutral evolutionary perspective, viruses will belong to different strains if they differ by more than G nucleotides, where G is a parameter whose value depends on the virus [214, 118]. If $D > G$, viruses in city 4 would belong, on average, to different strains when compared to city 1. As an example, if $G = 26$ new strains would arise, on average, in city 4 for $0 < p \leq 0.0010$, allowing a recovered individual from city 1 to be reinfected by an infected individual from city 4 if they are put in contact with each other (by travelling, for instance). Therefore, there is an increased risk of reinfection due to low connectivity among communities. In this sense, pandemics are more likely to originate new strains than epidemics, as they affect far more distant (therefore less connected) communities. One confirmed case of reinfection by COVID-19 in Hong-Kong had the virus differing by 24 nucleotides from the first infecting virus[215]. This distance matches a value for G for which the network connectivity would strongly influence the rise of reinfections.

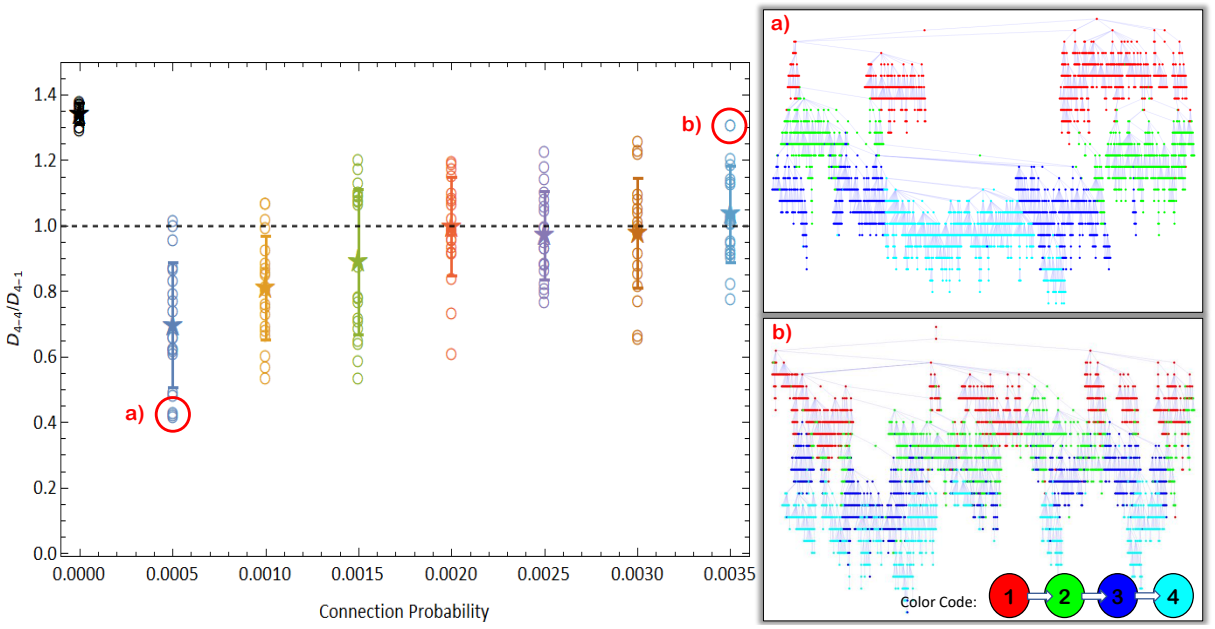


Figure 12.6: **Ratio between the average distance in city 4 and the average distance between cities 1 and 4.** Right panels show infection trees for the simulations highlighted with red circles. Open circles show results for individual simulations, the star is the average over 20 simulations and error bars are standard deviations. $p = 0$ represents a single Barabasi-Albert network with 4000 nodes (see text). Nodes in infection trees represent infected individuals, colored according to its city. City 4 (cyan) in panel (a), where $D_{4-4}/D_{4-1} < 1$, was almost entirely infected by a single viral lineage, while in panel (b) where $D_{4-4}/D_{4-1} > 1$, it was infected by many different viral lineages.
Source: Figure from Marquioni and de Aguiar, 2021 [190].

12.4 Conclusions

We have introduced an individual based model to describe the genetic evolution of a RNA-virus epidemic spreading . We used the SEIR model with four compartments on networks, but the evolutionary dynamics can be implemented in more compartmentalized epidemic models. We provided an analytical description that can be generalized for models with more compartments. An important result of this study is the mean-field approximation, Eq.(12.1.2), for the evolution of the average genetic distance, which can be added directly to the mean-field SIR or SEIR models.

Our analytical description of the average genetic distance between viruses is neutral and depends only on the epidemic curves. This allows us to project the evolutionary scenario without using the actual genome sequences. Deviations from these predictions in genetic data could reveal the strength of selection or network effects. We compared our prediction using only fifty complete genomes sequenced and collected in China and found good agreement.

We have also analysed the genetic evolution of the epidemic when it spreads over different communities. By changing the connection probability p between 4 linearly arranged communities we investigated how different the viruses infecting city 4 would be from their ancestors in city 1. Our simulations showed that when p is sufficiently small, the genetic difference between these viruses can be quite large, spanning 30 loci. This could allow an

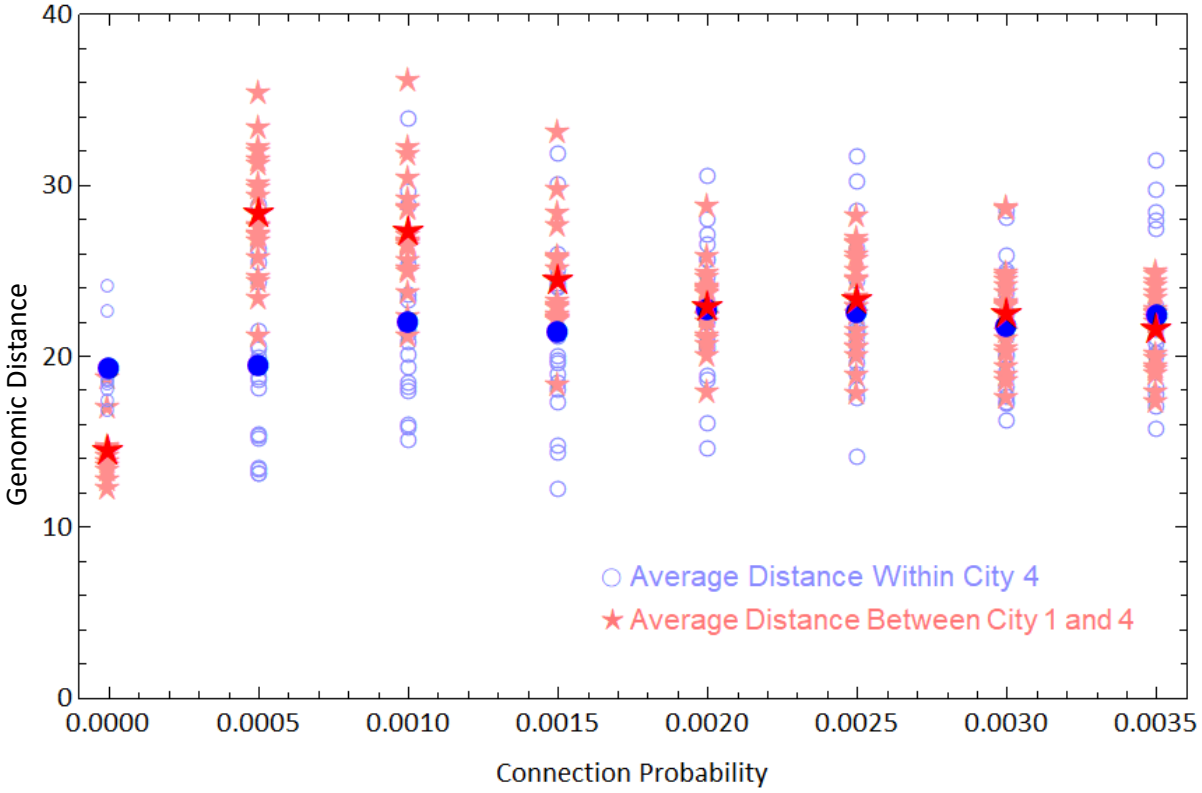


Figure 12.7: **Average genetic distances within cities 1 and 4.** Open blue circles are average distance between the viruses of city 4 from a single simulation, and the filled blue circle is average of these values. Light red stars are average distances between viruses from cities 1 and 4 and the dark red star is the average of these values. We ran 20 simulations for each value of connection probability.
Source: Figure from Marquioni and de Aguiar, 2021 [190].

infected individual from city 4 to reinfect a recovered individual from city 1. This is a consequence of the founder's effect, which is stronger if p is small as it decreases the number of infection gates of a community. Therefore, we expect increased risk of reinfection from contacts between travelling individuals living in distant territories.

Although the computational framework we described for the viral evolution is neutral, it can be adapted to including other evolutionary aspects, such as differential fitness for mutations in certain genome regions or loss of cross-immunity. These and other features are important topics to be added and studied in future works.

Chapter 13

This part in a nutshell

The COVID-19 pandemic highlighted on a global scale how the spread of a new pathogen is a real danger to human life and not only a science fiction plot of a book, movie, or TV show. In 2014, Katherine F. Smith et. al. analyzed infectious disease outbreaks from 1980 to 2013, counting more than 12000 outbreaks of more than 200 different human diseases, having affected millions of people [216]. But it was the COVID-19 that brought this dramatic reality to the mainstream. Global warming, rapid human growth and the invasion of natural environments share the responsibility for spillover events and the consequent spread of new zoonoses [217, 218]. The increase of epidemiological surveillance seems to be needed more than ever.

Biological evolution cannot be stopped and new epidemics caused by unseen pathogens are going to be faced. The understanding of how societal dynamics, people's behaviors and responses, the structure of human communities such as villages, cities and countries, can shape pathogen diversity and change the spread routes of infections is vital to implement any future control strategy. The optimization of mitigation strategies and non-pharmacological measures (which are necessary in the absence of vaccines and treatments) highly depends on how individuals' lives are connected to each other. Denser populations can face a higher transmission rate of airborne diseases, for instance, than not-so-dense communities [219].

In the model we developed, a population is described by the nodes of a network, which represents their mutual contact. Individuals are divided in compartments regarding their epidemiological state: susceptible (S), exposed (E), infected (I) and recovered (R). This is an individual-based version of the mean-field SEIR model. Contact heterogeneity is easily modeled with different network topologies. Although harder to treat it mathematically than its deterministic version, this stochastic epidemic model is suitable for evolutionary studies. Since we can track each individual's state, we can model the pathogen itself, to which we considered a binary string as a proxy for its genetic material.

As quarantine regimes were required to slow down the spread of COVID-19, we studied its effects in a scale-free network as a function of its *intensity, duration* and *starting day*. Due to the stochastic nature of the system, it is possible to observe different outcomes with different probabilities for each realization, which can be successful or not. Success in this analysis regards the peak of the infected curve, which policies aim to decrease and delay. Using COVID-19 parameters, we have identified three regions in the quarantine parameter space, which highlights the importance of strong and long quarantine regimes if the extinction of the epidemic is desired. On the other hand, the feasibility of imple-

menting such intense quarantine regime is very low, given its possible economic effects. Reducing the infection peak height, without erasing the viral spread, is therefore more viable, relying on not so intense quarantine states, but still with long duration periods [189].

But the viral evolution also had a significant impact in the COVID-19 pandemic, as reinfection cases were rising, evading acquired immunity and starting new infection waves. When the herd immunity threshold seemed to be achieved and the pandemic seemed to be under control (due to quarantine, use of masks, and vaccines), new viral strains were responsible for new cases and for the necessity of maintaining mitigation strategies. Viruses are under strong selection pressures, evading the immune system, resisting drugs, and being transmitted through different environmental conditions, and each one of these pressures affects the viral diversity. We then asked how much the contact network could shape the variability of a pathogen. Our case of study was the SARS-CoV-2 (the virus responsible for the COVID-19 disease), but our model is also suitable for other RNA-viruses.

In a neutral framework, including neutral mutations at random as long as a virus remains active in a host, we mainly found that poorly connected communities can increase the viral diversity as a result of the founder effect, which consequently increases the risk of the emergence of new strains and thus of reinfection cases. Our model also suggests a theory for connecting viral diversity to epidemiological quantities, like the number of infected and recovered, and despite its simplicity, we were able to describe the viral evolution in early cases in China (by the end of 2019 and the beginning of 2020) [190].

Albeit we did not investigate it, our model allows the study of reinfection dynamics, as also the introduction of non-equivalent strains (with different transmission rates, lethality, etc.). Also, the contact network could be well adjusted to real communities and to find reliable forecast results would be a matter of a reliable choice of parameters - which is itself a big challenge.

The model we have introduced resembles the Derrida-Higgs model, studied in the previous part of this thesis. It does not share the feature of sexual reproduction but uses a chain of bits to describe the genetic material (as in the one-parent model) and we advocate here for its power to depict evolutionary processes. In spite of its hard mathematical tractability, its possible extensions are limitless, relying on computational power and programming techniques. Epidemiology is indisputably a necessary science and we hope to have positively contributed to its understanding.

Final words

There is not so much more to be said after this intense mathematical journey. Of course, the work is not finished yet: different routes can be taken now, new extensions of the models discussed here, other generalizations, and analytical and numerical results can still be investigated. The Derrida-Higgs model still lacks of complete analytical theory. Albeit this work has formalized and characterized in detail the genetic similarity distribution, it has also evidenced how much the underlying network plays a very important and still not well understood role. If or when this "reproduction network" could share common features with real populations is for instance a very interesting question one could raise on the model.

We are not posing as a question the reality of the model: it is indeed not *real*, and I am always trying to be careful enough to say that the model "mimics" species formation, instead of saying it is a model of species formation. The mechanisms of genetics can be mapped onto binary chains, but the converse is not true. Not every binary chain dynamics displays all the intricate rules that give origin the complexity of life – nor even I believe all these rules are already known. Moreover, the Derrida-Higgs model does not even distinguish between genes and nucleotides or even aminoacids: the real biological scale is not defined, it simply models at a level in which information can be transmitted through reproduction. It is therefore unfair – and maybe meaningless – to ask whether the model as a whole is real or not. Nonetheless, it also does not mean it does not show important results and insights that can be applied to the interpretation of real evolutionary systems. We shall never forget George Box's line *"All models are wrong, but some are useful"* (1979) [220], and I argue that the Derrida-Higgs model can be a useful one. However, the comparison of theory and data is still a task to be done.

We focused here on the mathematical framework only, although having (hopefully) grounded the models we dealt with on evolutionary biology. There are questions I would still like to delve into, concerning the model, especially the role of generation overlap to the species formation and the distinction between pre and post-zygotic reproductive barriers, but these are left for future work. With some luck, we completed, in this text, the first theory of the Derrida-Higgs process.

In what concerns our epidemiology investigation, we have introduced a computational tool that led to some theoretical results on the spread of RNA viruses throughout contact networks. To frame an outbreak over networks instead of considering mean-field models is not a new idea, but to have access to the genome sequences during the spread and to analyze the viral diversity over its many possible configurations is the novelty we address here.

We have of course not exhausted the possibilities of our epidemic framework, and we kindly invite the reader to pursue further generalizations on the model – which is not hard to do! Our study on quarantines quantifies important strategies for the mitigation of an epidemic, as the study on the viral diversity on modular networks stresses the im-

portance of taking the connectivity of different communities into account when managing epidemiological policies. But what someone could conclude after including differential fitness among strains or the wane of immunity over time is also of great importance to epidemiology.

My hope is that the reader could get at least familiarized with different concepts in biomathematics: population dynamics, evolutionary biology, genetics, ecology, stochastic processes, network theory, numerical simulations, agent-based modeling, epidemiology, and any other hidden layer of knowledge someone can find. Apart from the contribution to science, specifically to biomathematics, as the sum up of my research over the last five years, I hope this text can also be my contribution to education in this beautiful interdisciplinary field.

With no more to add, I acknowledge the reader for your patience and invaluable curiosity.

Bibliography

[1] Mark Ridley. *Evolution, 3rd Edition*. 3rd ed. Blackwell Science Ltd, 2004.

[2] John Wilkins. "Species, kinds, and evolution". In: *Reports of the National Center for Science Education* 26.4 (2006), pp. 36–45.

[3] Warren John Ewens. *Mathematical population genetics: theoretical introduction*. Vol. 27. Springer, 2004.

[4] Ronald Aylmer Fisher. *The genetical theory of natural selection: a complete variorum edition*. Oxford University Press, 1999.

[5] John Burdon Haldane. *The causes of evolution*. Vol. 5. Princeton University Press, 1990.

[6] Sewall Wright. "Evolution in Mendelian populations". In: *Genetics* 16.2 (1931), p. 97.

[7] Julian Huxley et al. "Evolution. The modern synthesis." In: *Evolution. The Modern Synthesis.* (1942).

[8] Bruce Alberts. *Molecular biology of the cell*. Garland science, 2017.

[9] Carl R Woese, Otto Kandler, and Mark L Wheelis. "Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya." In: *Proceedings of the National Academy of Sciences* 87.12 (1990), pp. 4576–4579.

[10] Cody E Hinchliff et al. "Synthesis of phylogeny and taxonomy into a comprehensive tree of life". In: *Proceedings of the National Academy of Sciences* 112.41 (2015), pp. 12764–12769.

[11] Esteban Domingo, Julie Sheldon, and Celia Perales. "Viral quasispecies evolution". In: *Microbiology and Molecular Biology Reviews* 76.2 (2012), pp. 159–216.

[12] Alexander E Gorbalenya and Chris Lauber. "Phylogeny of viruses". In: *Reference Module in Biomedical Sciences* (2017).

[13] GE Fox et al. “The phylogeny of prokaryotes”. In: *Science* 209.4455 (1980), pp. 457–463.

[14] Madeline C Weiss et al. “The physiology and habitat of the last universal common ancestor”. In: *Nature microbiology* 1.9 (2016), pp. 1–8.

[15] Eugene V Koonin and Artem S Novozhilov. “Origin and evolution of the genetic code: the universal enigma”. In: *IUBMB life* 61.2 (2009), pp. 99–111.

[16] Manfred Eigen. “Selforganization of matter and the evolution of biological macromolecules”. In: *Naturwissenschaften* 58.10 (1971), pp. 465–523.

[17] Gustavo Caetano-Anollés et al. “The origin, evolution and structure of the protein world”. In: *Biochemical Journal* 417.3 (2009), pp. 621–637.

[18] Arianne J Matlin, Francis Clark, and Christopher WJ Smith. “Understanding alternative splicing: towards a cellular code”. In: *Nature reviews Molecular cell biology* 6.5 (2005), pp. 386–398.

[19] Frank E Zachos. *Species concepts in biology*. Vol. 801. Springer, 2016.

[20] Francis Galibert et al. “Toward understanding dog evolutionary and domestication history”. In: *Comptes rendus biologies* 334.3 (2011), pp. 190–196.

[21] Tomo Royama. *Analytical population dynamics*. Vol. 10. Springer Science & Business Media, 2012.

[22] James D Murray. *Mathematical biology I: an introduction*. 2002.

[23] Alfred J Lotka. “Analytical note on certain rhythmic relations in organic systems”. In: *Proceedings of the National Academy of Sciences* 6.7 (1920), pp. 410–415.

[24] Vito Volterra. “Fluctuations in the abundance of a species considered mathematically”. In: *Nature* 118.2972 (1926), pp. 558–560.

[25] Martin Feinberg. “Foundations of chemical reaction network theory”. In: (2019).

[26] Gui Araujo. “A framework of population dynamics from first principles”. In: *arXiv preprint arXiv:2304.12378* (2023).

[27] Guilherme David Araujo. “A Bayesian framework of reaction networks for dynamical population models”. PhD thesis. Universidade de São Paulo.

[28] Lei Ying. “On the approximation error of mean-field models”. In: *ACM SIGMETRICS Performance Evaluation Review* 44.1 (2016), pp. 285–297.

[29] Charles M Macal and Michael J North. "Agent-based modeling and simulation". In: *Proceedings of the 2009 winter simulation conference (WSC)*. IEEE. 2009, pp. 86–98.

[30] Volker Grimm et al. "A standard protocol for describing individual-based and agent-based models". In: *Ecological modelling* 198.1-2 (2006), pp. 115–126.

[31] Marcos Nascimento Magalhães. *Probabilidade e variáveis aleatórias*. Edusp, 2006.

[32] Øystein Ore. *Cardano: The gambling scholar*. Vol. 5063. Princeton University Press, 2017.

[33] Prakash Gorroochurn. "Some laws and problems of classical probability and how Cardano anticipated them". In: *Chance* 25.4 (2012), pp. 13–20.

[34] Lokenath Debnath and Kanadpriya Basu. "A short history of probability theory and its applications". In: *International Journal of Mathematical Education in Science and Technology* 46.1 (2015), pp. 13–39.

[35] Andreĭ Kolmogorov. *Foundations of the theory of probability: Second English Edition*.

[36] William Feller. "An Introduction to Probability Theory and its Applications, Volume 1". In: (1968).

[37] Crispin W Gardiner et al. *Handbook of stochastic methods*. Vol. 3. springer Berlin, 1985.

[38] Stephen M Stigler. "Who discovered Bayes's theorem?" In: *The American Statistician* 37.4a (1983), pp. 290–296.

[39] J Martin Bland and Douglas G Altman. "Bayesians and frequentists". In: *Bmj* 317.7166 (1998), pp. 1151–1160.

[40] Frederick Reif. *Fundamentals of statistical and thermal physics*. 1998.

[41] Lawrence T DeCarlo. "On the meaning and use of kurtosis." In: *Psychological methods* 2.3 (1997), p. 292.

[42] Sheldon M Ross. *Introduction to probability models*. Academic press, 2014.

[43] Nicolaas Godfried Van Kampen. *Stochastic processes in physics and chemistry*. Vol. 1. Elsevier, 1992.

[44] Tânia Tomé and Mário J De Oliveira. *Stochastic dynamics and irreversibility*. Springer, 2015.

[45] Marco Bartolozzi and Anthony William Thomas. "Stochastic cellular automata model for stock market dynamics". In: *Physical review E* 69.4 (2004), p. 046112.

[46] Joseph Klafter, Michael F Shlesinger, and Gert Zumofen. "Beyond brownian motion". In: *Physics today* 49.2 (1996), pp. 33–39.

[47] Rahman Farnoosh et al. "A stochastic perspective of RL electrical circuit using different noise terms". In: *COMPEL-The international journal for computation and mathematics in electrical and electronic engineering* 30.2 (2011), pp. 812–822.

[48] PH Leslie and JC Gower. "The properties of a stochastic model for the predator-prey type of interaction between two species". In: *Biometrika* 47.3/4 (1960), pp. 219–234.

[49] David G Kendall. "Stochastic processes and population growth". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 11.2 (1949), pp. 230–282.

[50] TD Frank. "Nonlinear Markov processes". In: *Physics Letters A* 372.25 (2008), pp. 4553–4555.

[51] Iris Sandler and Laurence Sandler. "On the origin of Mendelian genetics". In: *American Zoologist* 26.3 (1986), pp. 753–768.

[52] Ulrich Kutschera. "A comparative analysis of the Darwin-Wallace papers and the development of the concept of natural selection". In: *Theory in Biosciences* 122 (2003), pp. 343–359.

[53] John W Drake et al. "Rates of spontaneous mutation". In: *Genetics* 148.4 (1998), pp. 1667–1686.

[54] Michael Lynch. "Evolution of the mutation rate". In: *TRENDS in Genetics* 26.8 (2010), pp. 345–352.

[55] Michael Lynch et al. "Genetic drift, selection and the evolution of the mutation rate". In: *Nature Reviews Genetics* 17.11 (2016), pp. 704–714.

[56] Stylianos E Antonarakis et al. "Down syndrome". In: *Nature Reviews Disease Primers* 6.1 (2020), p. 9.

[57] Montgomery Slatkin. "Gene flow in natural populations". In: *Annual review of ecology and systematics* 16.1 (1985), pp. 393–430.

[58] Luis Boto. "Horizontal gene transfer in evolution: facts and challenges". In: *Proceedings of the Royal Society B: Biological Sciences* 277.1683 (2010), pp. 819–827.

[59] Roy Curtiss III. "Bacterial conjugation". In: *Annual review of microbiology* 23.1 (1969), pp. 69–136.

[60] Michael C Whitlock and David E McCauley. "Indirect measures of gene flow and migration: FST≠ 1/(4Nm+ 1)". In: *Heredity* 82.2 (1999), pp. 117–125.

[61] Joanna Masel. "Genetic drift". In: *Current Biology* 21.20 (2011), R837–R838.

[62] Donald A Darling and AJF58908 Siegert. "The first passage problem for a continuous Markov process". In: *The Annals of Mathematical Statistics* (1953), pp. 624–639.

[63] Tom Chou and Maria R D'Orsogna. "First passage problems in biology". In: *First-passage phenomena and their applications*. World Scientific, 2014, pp. 306–345.

[64] Charles Darwin's. "On the origin of species". In: *published on* 24 (1859), p. 1.

[65] Julian Davies and Dorothy Davies. "Origins and evolution of antibiotic resistance". In: *Microbiology and molecular biology reviews* 74.3 (2010), pp. 417–433.

[66] Pascale Gerbault et al. "Evolution of lactase persistence: an example of human niche construction". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 366.1566 (2011), pp. 863–877.

[67] Ulrich G Mueller et al. "Artificial selection on microbiomes to breed microbiomes that confer salt tolerance to plants". In: *MSystems* 6.6 (2021), e01125–21.

[68] Malte Andersson and Yoh Iwasa. "Sexual selection". In: *Trends in ecology & evolution* 11.2 (1996), pp. 53–58.

[69] James W Foster. "Kin selection and gorilla reproduction". In: *American Journal of Primatology* 3.S1 (1982), pp. 27–35.

[70] Clive K Catchpole. "Bird song, sexual selection and female choice". In: *Trends in Ecology & Evolution* 2.4 (1987), pp. 94–97.

[71] Thomas M Williams and Sean B Carroll. "Genetic and molecular insights into the development and evolution of sexual dimorphism". In: *Nature Reviews Genetics* 10.11 (2009), pp. 797–804.

[72] Sarah Blaffer Hrdy. "Infanticide among animals: a review, classification, and examination of the implications for the reproductive strategies of females". In: *Ethology and Sociobiology* 1.1 (1979), pp. 13–40.

[73] Dieter Lukas and Elise Huchard. "The evolution of infanticide by males in mammalian societies". In: *Science* 346.6211 (2014), pp. 841–844.

[74] AWF Edwards. "GH Hardy (1908) and hardy–weinberg equilibrium". In: *Genetics* 179.3 (2008), pp. 1143–1150.

[75] Oliver Mayo. "A century of Hardy–Weinberg equilibrium". In: *Twin Research and Human Genetics* 11.3 (2008), pp. 249–256.

[76] Ricky Der, Charles L Epstein, and Joshua B Plotkin. "Generalized population models and the nature of genetic drift". In: *Theoretical population biology* 80.2 (2011), pp. 80–99.

[77] Jerry A Coyne and H. Allen Orr. "Speciation". In: (2004).

[78] Jody Hey. "On the failure of modern species concepts". In: *Trends in ecology & evolution* 21.8 (2006), pp. 447–450.

[79] Chung-I Wu. *What are species and how are they formed?* 2022.

[80] Leigh Van Valen. "Ecological species, multispecies, and oaks". In: *Taxon* (1976), pp. 233–239.

[81] Jerry A Coyne. "Ernst Mayr and the origin of species". In: *Evolution* 48.1 (1994), pp. 19–30.

[82] Michael C Orr et al. "Six steps for building a technological knowledge base for future taxonomic work". In: *National Science Review* 9.12 (2022), nwac284.

[83] Randall T Schuh. "The Linnaean system and its 250-year persistence". In: *The Botanical Review* 69.1 (2003), pp. 59–78.

[84] Diane Marie Beaudoin Dodd. *Behavioral corrleates of the addptive divergence of Drosophila (reproductive isolation, habit choice)*. Yale University, 1984.

[85] Diane MB Dodd. "Reproductive isolation as a consequence of adaptive divergence in Drosophila pseudoobscura". In: *Evolution* (1989), pp. 1308–1311.

[86] Sergey Gavrilets. "Perspective: models of speciation: what have we learned in 40 years?" In: *Evolution* 57.10 (2003), pp. 2197–2215.

[87] Norman A Johnson. "Speciation: genomic sequence data and the biogeography of speciation". In: *National Science Review* 9.12 (2022), nwac294.

[88] Daniel I Bolnick and Benjamin M Fitzpatrick. "Sympatric speciation: models and empirical evidence". In: *Annu. Rev. Ecol. Evol. Syst.* 38 (2007), pp. 459–487.

[89] Christopher H Martin. "Is sympatric speciation in nature only possible with micro-parapatry? A new case study of glacial lake cyprinid fishes". In: *National Science Review* 9.12 (2022), nwad006.

[90] Ole Seehausen and Jacques JM van Alphen. "The effect of male coloration on female mate choice in closely related Lake Victoria cichlids (Haplochromis nyererei complex)". In: *Behavioral Ecology and Sociobiology* 42 (1998), pp. 1–8.

[91] Marta Barluenga et al. "Sympatric speciation in Nicaraguan crater lake cichlid fish". In: *Nature* 439.7077 (2006), pp. 719–723.

[92] Andreas F Kautt, Gonzalo Machado-Schiaffino, and Axel Meyer. "Multispecies outcomes of sympatric speciation after admixture with the source population in two radiations of Nicaraguan crater lake cichlids". In: *PLoS genetics* 12.6 (2016), e1006157.

[93] Nicholas H Barton et al. *Evolution*. 2007.

[94] Paul G Higgs and Bernard Derrida. "Stochastic models for species formation in evolving populations". In: *Journal of Physics A: Mathematical and General* 24.17 (1991), p. L985.

[95] Réka Albert and Albert-László Barabási. "Statistical mechanics of complex networks". In: *Reviews of modern physics* 74.1 (2002), p. 47.

[96] Albert-Laszlo Barabasi. *Network Science*. Cambridge University Press, 2016.

[97] Sergey N Dorogovtsev and José FF Mendes. *The nature of complex networks*. Oxford University Press, 2022.

[98] Mark Newman, Albert-László Barabási, and Duncan J Watts. *The structure and dynamics of networks*. Princeton university press, 2011.

[99] Craig W Reynolds. "Flocks, herds and schools: A distributed behavioral model". In: *Proceedings of the 14th annual conference on Computer graphics and interactive techniques*. 1987, pp. 25–34.

[100] William Bialek et al. "Statistical mechanics for natural flocks of birds". In: *Proceedings of the National Academy of Sciences* 109.13 (2012), pp. 4786–4791.

[101] Matthew B Biggs et al. "Metabolic network modeling of microbial communities". In: *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* 7.5 (2015), pp. 317–334.

[102] Albert-László Barabási and Réka Albert. "Emergence of scaling in random networks". In: *science* 286.5439 (1999), pp. 509–512.

[103] Albert-László Barabási and Eric Bonabeau. "Scale-free networks". In: *Scientific american* 288.5 (2003), pp. 60–69.

[104] Douglas Brent West et al. *Introduction to graph theory*. Vol. 2. Prentice hall Upper Saddle River, 2001.

[105] Romualdo Pastor-Satorras and Alessandro Vespignani. "Epidemic spreading in scale-free networks". In: *Physical review letters* 86.14 (2001), p. 3200.

[106] Yamir Moreno, Romualdo Pastor-Satorras, and Alessandro Vespignani. "Epidemic outbreaks in complex heterogeneous networks". In: *The European Physical Journal B-Condensed Matter and Complex Systems* 26 (2002), pp. 521–529.

[107] Ray Solomonoff and Anatol Rapoport. "Connectivity of random nets". In: *The bulletin of mathematical biophysics* 13 (1951), pp. 107–117.

[108] Casper Goffman. "And what is your Erdös number?" In: *The American Mathematical Monthly* 76.7 (1969), pp. 791–791.

[109] American Mathematical Society. *Collaboration Distance*. Accessed on 07 Dec. 2023. URL: `https://mathscinet.ams.org/mathscinet/freetools/collab-dist`.

[110] Brian Hopkins. "Kevin Bacon and graph theory". In: *Problems, Resources, and Issues in Mathematics Undergraduate Studies* 14.1 (2004), pp. 5–11.

[111] Paul Erdős, Alfréd Rényi, et al. "On the evolution of random graphs". In: *Publ. math. inst. hung. acad. sci* 5.1 (1960), pp. 17–60.

[112] Stanley Milgram. "The small world problem". In: *Psychology today* 2.1 (1967), pp. 60–67.

[113] Duncan J Watts and Steven H Strogatz. "Collective dynamics of 'small-world' networks". In: *nature* 393.6684 (1998), pp. 440–442.

[114] Réka Albert, Hawoong Jeong, and Albert-László Barabási. "Diameter of the world-wide web". In: *nature* 401.6749 (1999), pp. 130–131.

[115] Derek de Solla Price. "A general theory of bibliometric and other cumulative advantage processes". In: *Journal of the American society for Information science* 27.5 (1976), pp. 292–306.

[116] Sergey N Dorogovtsev, José Fernando F Mendes, and Alexander N Samukhin. "Structure of growing networks with preferential linking". In: *Physical review letters* 85.21 (2000), p. 4633.

[117] Zoltán Dezső and Albert-László Barabási. "Halting viruses in scale-free networks". In: *Physical Review E* 65.5 (2002), p. 055103.

[118] Marcus AM De Aguiar. "Speciation in the Derrida–Higgs model with finite genomes and spatial populations". In: *Journal of Physics A: Mathematical and Theoretical* 50.8 (2017), p. 085602.

[119] Nick Lane. "On the origin of bar codes: genetic sequences in a cell's mitochondria can be used to accurately determine species. Could this be because they are responsible for creating what they identify?" In: *Nature* 462.7271 (2009), pp. 272–275.

[120] Débora Princepe and Marcus AM De Aguiar. "Modeling mito-nuclear compatibility and its role in species identification". In: *Systematic biology* 70.1 (2021), pp. 133–144.

[121] Bernard Derrida and Luca Peliti. "Evolution in a flat fitness landscape". In: *Bulletin of mathematical biology* 53.3 (1991), pp. 355–382.

[122] John C Avise et al. "Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics". In: *Annual review of ecology and systematics* 18.1 (1987), pp. 489–522.

[123] Michael Begon and Colin R Townsend. *Ecology: from individuals to ecosystems*. John Wiley & Sons, 2021.

[124] Brian J McGill et al. "Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework". In: *Ecology letters* 10.10 (2007), pp. 995–1015.

[125] RG Hughes. "Theories and models of species abundance". In: *The American Naturalist* 128.6 (1986), pp. 879–899.

[126] Frank W Preston. "The commonness, and rarity, of species". In: *Ecology* 29.3 (1948), pp. 254–283.

[127] John S Gray, Anders Bjørgesæter, and Karl. I. Ugland. "On plotting species abundance distributions". In: *Journal of Animal Ecology* (2006), pp. 752–756.

[128] Ronald A Fisher, A Steven Corbet, and Carrington B Williams. "The relation between the number of species and the number of individuals in a random sample of an animal population". In: *The Journal of Animal Ecology* (1943), pp. 42–58.

[129] WG Wilson et al. *Biodiversity and species interactions: extending Lotka–Volterra community theory*. 2003.

[130] Larissa Lubiana Botelho, Flavia Maria Darcie Marquitti, and Marcus AM de Aguiar. "Extinction and hybridization in a neutral model of speciation". In: *Journal of Physics A: Mathematical and Theoretical* 55.38 (2022), p. 385601.

[131] Heidi M McBride, Margaret Neuspiel, and Sylwia Wasiak. "Mitochondria: more than just a powerhouse". In: *Current biology* 16.14 (2006), R551–R560.

[132] Geoffrey E Hill. "Mitonuclear coevolution as the genesis of speciation and the mitochondrial DNA barcode gap". In: *Ecology and evolution* 6.16 (2016), pp. 5831–5842.

[133] Alexandra Pavlova et al. "Purifying selection and genetic drift shaped Pleistocene evolution of the mitochondrial genome in an endangered Australian freshwater fish". In: *Heredity* 118.5 (2017), pp. 466–476.

[134] EM Baptestini, L Kaufman, and Y Bar-Yam. "Global patterns of speciation and diversity". In: *Nature* 460.7253 (2009), pp. 384–387.

[135] Martin A Nowak. *Evolutionary dynamics: exploring the equations of life*. Harvard university press, 2006.

[136] Débora Princepe, Marcus AM de Aguiar, and Joshua B Plotkin. "Mito-nuclear selection induces a trade-off between species ecological dominance and evolutionary lifespan". In: *Nature Ecology & Evolution* 6.12 (2022), pp. 1992–2002.

[137] Na Zhu et al. "A novel coronavirus from patients with pneumonia in China, 2019". In: *New England journal of medicine* 382.8 (2020), pp. 727–733.

[138] Stanley Perlman. *Another decade, another coronavirus*. 2020.

[139] Di Wu et al. "The SARS-CoV-2 outbreak: what we know". In: *International journal of infectious diseases* 94 (2020), pp. 44–48.

[140] W Joost Wiersinga et al. "Pathophysiology, transmission, diagnosis, and treatment of coronavirus disease 2019 (COVID-19): a review". In: *Jama* 324.8 (2020), pp. 782–793.

[141] Domenico Cucinotta and Maurizio Vanelli. "WHO declares COVID-19 a pandemic". In: *Acta bio medica: Atenei parmensis* 91.1 (2020), p. 157.

[142] C Jessica E Metcalf, Dylan H Morris, and Sang Woo Park. "Mathematical models to guide pandemic response". In: *Science* 369.6502 (2020), pp. 368–369.

[143] Natalie M Linton et al. "Epidemiological characteristics of novel coronavirus infection: A statistical analysis of publicly available case data". In: *MedRxiv* (2020), pp. 2020–01.

[144] Benjamin Ivorra et al. "Mathematical modeling of the spread of the coronavirus disease 2019 (COVID-19) taking into account the undetected infections. The case of China". In: *Communications in nonlinear science and numerical simulation* 88 (2020), p. 105303.

[145] Neil Ferguson et al. "Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand". In: (2020).

[146] Dylan H Morris et al. "Optimal, near-optimal, and robust epidemic control". In: *Communications Physics* 4.1 (2021), p. 78.

[147] Seth Flaxman et al. "Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe". In: *Nature* 584.7820 (2020), pp. 257–261.

[148] Sarah Schaffer DeRoo, Natalie J Pudalov, and Linda Y Fu. "Planning for a COVID-19 vaccination program". In: *Jama* 323.24 (2020), pp. 2458–2459.

[149] Barney S Graham. "Rapid COVID-19 vaccine development". In: *Science* 368.6494 (2020), pp. 945–946.

[150] Guido Forni and Alberto Mantovani. "COVID-19 vaccines: where we stand and challenges ahead". In: *Cell Death & Differentiation* 28.2 (2021), pp. 626–639.

[151] James M Sanders et al. "Pharmacologic treatments for coronavirus disease 2019 (COVID-19): a review". In: *Jama* 323.18 (2020), pp. 1824–1836.

[152] Ângela Maria Bagattini et al. "PP25 Brazilian Collaborative Network For COVID-19 Modeling: Successful Experience Of Using Real-Time Science To Support Evidence-Based Decision-Making". In: *International Journal of Technology Assessment in Health Care* 38.S1 (2022), S48–S49.

[153] World Health Organization et al. "Impact of COVID-19 on people's livelihoods, their health and our food systems". In: *Joint statement by ILO, FAO, IFAD and WHO* 13 (2020).

[154] Youssef Miyah et al. "COVID-19 impact on public health, environment, human psychology, global socioeconomy, and education". In: *The Scientific World Journal* 2022 (2022).

[155] World Health Organization. *WHO Coronavirus (COVID-19) Dashboard*. Accessed December 5, 2023. 2023. URL: `https://covid19.who.int/`.

[156] Kimberly A Prather, Chia C Wang, and Robert T Schooley. "Reducing transmission of SARS-CoV-2". In: *Science* 368.6498 (2020), pp. 1422–1424.

[157] Andrew G Harrison, Tao Lin, and Penghua Wang. "Mechanisms of SARS-CoV-2 transmission and pathogenesis". In: *Trends in immunology* 41.12 (2020), pp. 1100–1115.

[158] Quan-Xin Long et al. "Clinical and immunological assessment of asymptomatic SARS-CoV-2 infections". In: *Nature medicine* 26.8 (2020), pp. 1200–1204.

[159] Daniel P Oran and Eric J Topol. "Prevalence of asymptomatic SARS-CoV-2 infection: a narrative review". In: *Annals of internal medicine* 173.5 (2020), pp. 362–367.

[160] Angela L Rasmussen and Saskia V Popescu. "SARS-CoV-2 transmission without symptoms". In: *Science* 371.6535 (2021), pp. 1206–1207.

[161] Wei Cao and Taisheng Li. "COVID-19: towards understanding of pathogenesis". In: *Cell research* 30.5 (2020), pp. 367–369.

[162] Reed Abelson. "COVID overload: US hospitals are running out of beds for patients". In: *The New York Times* (2020).

[163] Alexander T Toth et al. "Surge and mortality in ICUs in New York City's public healthcare system". In: *Critical Care Medicine* 49.9 (2021), pp. 1439–1450.

[164] JK Nature. "Coronavirus: the first three months as it happened". In: *Nature News* (2020).

[165] P Sol Hart, Sedona Chinn, and Stuart Soroka. "Politicization and polarization in COVID-19 news coverage". In: *Science communication* 42.5 (2020), pp. 679–697.

[166] Monica Malta, Steffanie A Strathdee, and Patricia J Garcia. "The brazilian tragedy: Where patients living at the 'Earth's lungs' die of asphyxia, and the fallacy of herd immunity is killing people." In: *EClinicalMedicine* 32 (2021).

[167] Jon Roozenbeek et al. "Susceptibility to misinformation about COVID-19 around the world". In: *Royal Society open science* 7.10 (2020), p. 201199.

[168] Daniel Jolley and Jenny L Paterson. "Pylons ablaze: Examining the role of 5G COVID-19 conspiracy beliefs and support for violence". In: *British journal of social psychology* 59.3 (2020), pp. 628–640.

[169] Salman Bin Naeem and Maged N Kamel Boulos. "COVID-19 misinformation online and health literacy: a brief overview". In: *International journal of environmental research and public health* 18.15 (2021), p. 8091.

[170] Simran Preet Kaur and Vandana Gupta. "COVID-19 Vaccine: A comprehensive status report". In: *Virus research* 288 (2020), p. 198114.

[171] Talha Burki. "The online anti-vaccine movement in the age of COVID-19". In: *The Lancet Digital Health* 2.10 (2020), e504–e505.

[172] Brandy Zadrozny and Ben Collins. "As vaccine mandates spread, protests follow — some spurred by nurses". In: *NBC News* (2021).

[173] Alistair Coleman and Shayan Sardarizadeh. "Anti-vax protests: 'Sovereign citizens' fight UK Covid vaccine rollout". In: *BBC News* (2022).

[174] RH Pfizer. *Pfizer and BioNTech achieve first authorization in the world for a vaccine to combat COVID-19*. 2020.

[175] Yvette N Lamb. "BNT162b2 mRNA COVID-19 vaccine: first approval". In: *Drugs* 81 (2021), pp. 495–501.

[176] Kate M Bubar et al. "Model-informed COVID-19 vaccine prioritization strategies by age and serostatus". In: *Science* 371.6352 (2021), pp. 916–921.

[177] Manuel Adrian Acuña-Zegarra et al. "COVID-19 optimal vaccination policies: A modeling study on efficacy, natural and vaccine-induced immunity responses". In: *Mathematical biosciences* 337 (2021), p. 108614.

[178] Jon Cohen and Kai Kupferschmidt. *As vaccines emerge, a global waiting game begins*. 2020.

[179] David P Fidler. *Vaccine nationalism's politics*. 2020.

[180] Houriiyah Tegally et al. "Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa". In: *medrxiv* (2020), pp. 2020–12.

[181] Deepa Vasireddy et al. "Review of COVID-19 variants and COVID-19 vaccine efficacy: what the clinician should know?" In: *Journal of Clinical Medicine Research* 13.6 (2021), p. 317.

[182] Daming Zhou et al. "Evidence of escape of SARS-CoV-2 variant B. 1.351 from natural and vaccine-induced sera". In: *Cell* 184.9 (2021), pp. 2348–2361.

[183] Renato Mendes Coutinho et al. "Model-based estimation of transmissibility and reinfection of SARS-CoV-2 P. 1 variant". In: *Communications Medicine* 1.1 (2021), p. 48.

[184] Emily Harris. "WHO Declares End of COVID-19 Global Health Emergency". In: *JAMA* 329.21 (2023), pp. 1817–1817.

[185] Monica Malta et al. “Political neglect of COVID-19 and the public health consequences in Brazil: The high costs of science denial”. In: *EClinicalMedicine* 35 (2021).

[186] Fred Brauer. “Compartmental models in epidemiology”. In: *Mathematical epidemiology* (2008), pp. 19–79.

[187] Edilson F Arruda et al. “Modelling and optimal control of multi strain epidemics, with application to COVID-19”. In: *PLoS One* 16.9 (2021), e0257512.

[188] Maira Aguiar and Nico Stollenwerk. “Mathematical models of dengue fever epidemiology: multi-strain dynamics, immunological aspects associated to disease severity and vaccines”. In: *Communication in Biomathematical Sciences* 1.1 (2017), pp. 1–12.

[189] Vitor M Marquioni and Marcus AM De Aguiar. “Quantifying the effects of quarantine using an IBM SEIR model on scalefree networks”. In: *Chaos, Solitons & Fractals* 138 (2020), p. 109999.

[190] Vitor M Marquioni and Marcus AM de Aguiar. “Modeling neutral viral mutations in the spread of SARS-CoV-2 epidemics”. In: *Plos One* 16.7 (2021), e0255438.

[191] Xue-Zhi Li, Junyuan Yang, and Maia Martcheva. *Age structured epidemic modeling*. Vol. 52. Springer Nature, 2020.

[192] Miles S Okino and Michael L Mavrovouniotis. “Simplification of mathematical models of chemical reaction systems”. In: *Chemical reviews* 98.2 (1998), pp. 391–408.

[193] Gordon R Conway. “Mathematical models in applied ecology”. In: *Nature* 269.5626 (1977), pp. 291–297.

[194] Eberhard O Voit, Harald A Martens, and Stig W Omholt. “150 years of the mass action law”. In: *PLoS computational biology* 11.1 (2015), e1004012.

[195] Otso Ovaskainen and Baruch Meerson. “Stochastic models of population extinction”. In: *Trends in ecology & evolution* 25.11 (2010), pp. 643–652.

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

[197] Steven C Bankes. "Agent-based modeling: A revolution?" In: *Proceedings of the National Academy of Sciences* 99.suppl_3 (2002), pp. 7199–7200.

[198] Haifeng Zhang et al. "Hub nodes inhibit the outbreak of epidemic under voluntary vaccination". In: *New Journal of Physics* 12.2 (2010), p. 023015.

[199] Marc Barthélemy et al. "Dynamical patterns of epidemic outbreaks in complex heterogeneous networks". In: *Journal of theoretical biology* 235.2 (2005), pp. 275–288.

[200] Haley E Randolph and Luis B Barreiro. "Herd immunity: understanding COVID-19". In: *Immunity* 52.5 (2020), pp. 737–741.

[201] Zhongming Zhao et al. "Moderate mutation rate in the SARS coronavirus genome and its implications". In: *BMC evolutionary biology* 4 (2004), pp. 1–9.

[202] Ying Liu et al. "The reproductive number of COVID-19 is higher compared to SARS coronavirus". In: *Journal of travel medicine* (2020).

[203] Joseph T Wu et al. "Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China". In: *Nature medicine* 26.4 (2020), pp. 506–510.

[204] Yan-Rong Guo et al. "The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak–an update on the status". In: *Military medical research* 7 (2020), pp. 1–10.

[205] Jantien A Backer, Don Klinkenberg, and Jacco Wallinga. "Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20–28 January 2020". In: *Eurosurveillance* 25.5 (2020), p. 2000062.

[206] Zijie Shen et al. "Genomic diversity of severe acute respiratory syndrome–coronavirus 2 in patients with coronavirus disease 2019". In: *Clinical infectious diseases* 71.15 (2020), pp. 713–720.

[207] Mark EJ Newman. "Modularity and community structure in networks". In: *Proceedings of the national academy of sciences* 103.23 (2006), pp. 8577–8582.

[208] Wolfram Research. *Epidemic Data for Novel Coronavirus COVID-19*. Wolfram Data Repository `https://doi.org/10.24097/wolfram.04123.data`. 2020.

[209] Ruiyun Li et al. "Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2)". In: *Science* 368.6490 (2020), pp. 489–493.

[210] Wolfram Research. *Genetic Sequences for the SARS-CoV-2 Coronavirus*. Wolfram Data Repository `https://doi.org/10.24097/wolfram.03304.data`. 2020.

[211] Wing-Kin Sung. *Algorithms in bioinformatics: A practical introduction*. CRC Press, 2009.

[212] Peter Forster et al. "Phylogenetic network analysis of SARS-CoV-2 genomes". In: *Proceedings of the National Academy of Sciences* 117.17 (2020), pp. 9241–9243.

[213] Yongsen Ruan et al. "On the founder effect in COVID-19 outbreaks: how many infected travelers may have started them all?" In: *National Science Review* 8.1 (2021), nwaa246.

[214] Carolina LN Costa et al. "Registering the evolutionary history in individual-based models of speciation". In: *Physica A: Statistical Mechanics and its Applications* 510 (2018), pp. 1–14.

[215] Kelvin Kai-Wang To et al. "Coronavirus disease 2019 (COVID-19) re-infection by a phylogenetically distinct severe acute respiratory syndrome coronavirus 2 strain confirmed by whole genome sequencing". In: *Clinical Infectious Diseases* 73.9 (2021), e2946–e2951.

[216] Katherine F Smith et al. "Global rise in human infectious disease outbreaks". In: *Journal of the Royal Society Interface* 11.101 (2014), p. 20140950.

[217] David M Morens, Gregory K Folkers, and Anthony S Fauci. "The challenge of emerging and re-emerging infectious diseases". In: *Nature* 430.6996 (2004), pp. 242–249.

[218] Neil M Vora et al. "Want to prevent pandemics? Stop spillovers". In: *Nature* 605.7910 (2022), pp. 419–422.

[219] Patrick M Tarwater and Clyde F Martin. "Effects of population density on the spread of disease". In: *Complexity* 6.6 (2001), pp. 29–36.

[220] George EP Box. "Robustness in the strategy of scientific model building". In: *Robustness in statistics*. Elsevier, 1979, pp. 201–236.

[221] World Health Organization. *Coronavirus disease 2019 (COVID-19) Situation Report 37*. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200226-sitrep-37-covid-19.pdf?sfvrsn=2146841e_2. 2020.

Appendix A

On the Derrida-Higgs model

A.1 The expected similarity value

The expression for the expected similarity value can be simplified if we introduce some matrices. Let us start by analyzing each term of the expression for the expected value

$$\mathbb{E}(q_{t+1}^{\alpha\beta}) = \frac{e^{-4\mu}}{4N^2}\sum_{p_1}\sum_{p_2}\sum_{p_1'}\sum_{p_2'}\frac{A_{p_1p_2}A_{p_1'p_2'}}{N_{p_1}N_{p_1'}}\left(\underbrace{q_t^{p_1,p_1'}}_{\text{(i)}}+\underbrace{q_t^{p_1,p_2'}}_{\text{(ii)}}+\underbrace{q_t^{p_2,p_1'}}_{\text{(iii)}}+\underbrace{q_t^{p_2,p_2'}}_{\text{(iv)}}\right),$$

we can analyze each one of them. Defining the matrices $(\mathbb{A}_n)_{ij} = \frac{A_{ij}}{N_i}$ and $(\mathbb{Q}_t)_{ij} = q_t^{ij}$,

(i)

$$\begin{aligned}\sum_{p_1}\sum_{p_2}\sum_{p_1'}\sum_{p_2'}\frac{A_{p_1p_2}A_{p_1'p_2'}}{N_{p_1}N_{p_1'}}q_t^{p_1,p_1'} &= \sum_{p_1}\sum_{p_1'}\frac{q_t^{p_1,p_1'}}{N_{p_1}N_{p_1'}}\left(\sum_{p_2}A_{p_1p_2}\right)\left(\sum_{p_2'}A_{p_1'p_2'}\right)\\ &= \sum_{p_1}\sum_{p_1'}q_t^{p_1,p_1'} = \sum_{i,j}(\mathbb{Q}_t)_{ij}\,; \end{aligned} \tag{A.1.1}$$

(ii)

$$\begin{aligned}\sum_{p_1}\sum_{p_2}\sum_{p_1'}\sum_{p_2'}\frac{A_{p_1p_2}A_{p_1'p_2'}}{N_{p_1}N_{p_1'}}q_t^{p_1,p_2'} &= \sum_{p_1}\sum_{p_1'}\sum_{p_2'}\frac{A_{p_1'p_2'}q_t^{p_1,p_2'}}{N_{p_1}N_{p_1'}}\left(\sum_{p_2}A_{p_1p_2}\right)\\ &= \sum_{p_1}\sum_{p_1'}\left(\sum_{p_2'}\frac{A_{p_1'p_2'}}{N_{p_1'}}q_t^{p_1,p_2'}\right) = \sum_{i,j}(\mathbb{A}_n\times\mathbb{Q}_t)_{ij}\,; \end{aligned} \tag{A.1.2}$$

(iii)

$$\sum_{p_1}\sum_{p_2}\sum_{p_1'}\sum_{p_2'}\frac{A_{p_1p_2}A_{p_1'p_2'}}{N_{p_1}N_{p_1'}}q_t^{p_2,p_1'} = \sum_{p_1}\sum_{p_1'}\sum_{p_2}\frac{A_{p_1p_2}q_t^{p_2,p_1'}}{N_{p_1}N_{p_1'}}\left(\sum_{p_2'}A_{p_1'p_2'}\right)$$
$$= \sum_{p_1}\sum_{p_1'}\left(\sum_{p_2}\frac{A_{p_1p_2}}{N_{p_1}}q_t^{p_2,p_1'}\right) = \sum_{i,j}\left(\mathbb{Q}_t\times\mathbb{A}_n^T\right)_{ij}; \tag{A.1.3}$$

(iv)

$$\sum_{p_1}\sum_{p_2}\sum_{p_1'}\sum_{p_2'}\frac{A_{p_1p_2}A_{p_1'p_2'}}{N_{p_1}N_{p_1'}}q_t^{p_2,p_2'} = \sum_{p_1}\sum_{p_2'}\sum_{p_1'}\left(\sum_{p_2}\frac{A_{p_1p_2}}{N_{p_1}}q_t^{p_2,p_2'}\right)\frac{A_{p_1'p_2'}}{N_{p_1'}}$$
$$= \sum_{p_1}\sum_{p_1'}\left(\sum_{p_2'}(\mathbb{A}_n\times\mathbb{Q}_t)_{p_1p_2'}(\mathbb{A}_n^T)_{p_2'p_1'}\right)$$
$$= \sum_{i,j}\left(\mathbb{A}_n\times\mathbb{Q}_t\times\mathbb{A}_n^T\right)_{ij}. \tag{A.1.4}$$

Thus,

$$\mathbb{E}(q_{t+1}^{\alpha\beta}) = \frac{e^{-4\mu}}{4N^2}\sum_{i,j}\left(\mathbb{Q}_t+\mathbb{A}_n\times\mathbb{Q}_t+\mathbb{Q}_t\times\mathbb{A}_n^T+\mathbb{A}_n\times\mathbb{Q}_t\times\mathbb{A}_n^T\right)_{ij}. \tag{A.1.5}$$

Now, let $\mathbb{I}_N$ be the identity matrix of order N and defining the matrix $\mathbb{C}\equiv(\mathbb{I}_N+\mathbb{A}_n)/2$, one can write

$$\mathbb{E}(q_{t+1}^{\alpha\beta}) = \frac{e^{-4\mu}}{N^2}\sum_{i,j}\left(\mathbb{C}\mathbb{Q}_t\mathbb{C}^T\right)_{ij}. \tag{A.1.6}$$

A simple property of the defined matrices $\mathbb{A}_n$ and $\mathbb{C}$ is that they are both stochastic to the right. To prove it, one needs to notice that each row i of $\mathbb{A}_n$ is normalized by the degree of node i,

$$\sum_{j=1}^{N}(\mathbb{A}_n)_{ij} = \sum_{j=1}^{N}\frac{A_{ij}}{N_i} = \frac{1}{N_i}\sum_{j=1}^{N}A_{i,j} = 1$$

and now for $\mathbb{C}$,

$$\sum_{j=1}^{N}(\mathbb{C}_n)_{ij} = \sum_{j=1}^{N}\frac{1}{2}(\mathbb{I}_N+\mathbb{A}_n)_{ij} = \frac{1}{2}\left(\sum_{j=1}^{N}(\mathbb{I}_N)_{ij}+\sum_{j=1}^{N}(\mathbb{A}_n)_{ij}\right) = 1.$$

Although interesting, this notation does not help to better understand the process, once the operation $\mathbb{C}\mathbb{Q}_t\mathbb{C}^T$ is followed by a sum over the whole matrix in order to find an average value, in such a way that these matrices do not define an evolution that could be studied, for instance, as a Markov chain.

A.2 Sums on three and four indexes

A common step during the calculation of the moments of the similarity distribution of the Derrida-Higgs model is the sum over many indexes. The problem concerning these sums is that some combinations of indexes may have different outcomes, therefore it is important to organize these sums over all these possible combinations, and that is what we show here.

$$\sum_{p_1,p_1',p_2'} f(p_1,p_1',p_2') = \sum_{p_1}\left[\sum_{p_1'=p_1}\sum_{p_2'=p_1} + \sum_{p_1'=p_1}\sum_{p_2'\neq p_1} + \sum_{p_1'\neq p_1}\sum_{p_2'=p_1} + \sum_{p_1'\neq p_1}\sum_{p_2'=p_1'} + \sum_{p_1'\neq p_1}\sum_{p_2'\neq p_1,p_1'}\right] f(p_1,p_1',p_2') \tag{A.2.1}$$

and

$$\begin{aligned}\sum_{p_1,p_2,p_1',p_2'} f(p_1,p_2,p_1',p_2') = \sum_{p_1}\Bigg[&\sum_{p_2=p_1}\sum_{p_1'=p_1}\sum_{p_2'=p_1} + \sum_{p_2=p_1}\sum_{p_1'=p_1}\sum_{p_2'\neq p_1} + \sum_{p_2=p_1}\sum_{p_1'\neq p_1}\sum_{p_2'=p_1} + \sum_{p_2=p_1}\sum_{p_1'\neq p_1}\sum_{p_2'=p_1'}\\ &+ \sum_{p_2=p_1}\sum_{p_1'\neq p_1}\sum_{p_2'\neq p_1,p_1'} + \sum_{p_2\neq p_1}\sum_{p_1'=p_1}\sum_{p_2'=p_1} + \sum_{p_2\neq p_1}\sum_{p_1'=p_1}\sum_{p_2'=p_2} + \sum_{p_2\neq p_1}\sum_{p_1'=p_1}\sum_{p_2'\neq p_2,p_1}\\ &+ \sum_{p_2\neq p_1}\sum_{p_1'=p_2}\sum_{p_2'=p_1} + \sum_{p_2\neq p_1}\sum_{p_1'=p_2}\sum_{p_2'=p_2} + \sum_{p_2\neq p_1}\sum_{p_1'=p_2}\sum_{p_2'\neq p_1,p_2} + \sum_{p_2\neq p_1}\sum_{p_1'\neq p_1,p_2}\sum_{p_2'=p_1}\\ &+ \sum_{p_2\neq p_1}\sum_{p_1'\neq p_1,p_2}\sum_{p_2'=p_2} + \sum_{p_2\neq p_1}\sum_{p_1'\neq p_1,p_2}\sum_{p_2'=p_1'} + \sum_{p_2\neq p_1}\sum_{p_1'\neq p_1,p_2}\sum_{p_2'\neq p_1,p_2,p_1'}\Bigg] f(p_1,p_2,p_1',p_2').\end{aligned} \tag{A.2.2}$$

A simple check on these expansions is to consider the f functions to be $f = 1$ and perform the sums, term by term

$$\begin{aligned}\sum_{p_1,p_1',p_2'} 1 &= N^3\\ &= N\left[1 + (N-1) + (N-1) + (N-1) + (N-1)(N-2)\right] = N^3 \text{ , q.e.d.}\end{aligned}$$

and

$$\begin{aligned}\sum_{p_1,p_2,p_1',p_2'} 1 &= N^4 \\ &= N\,[1 + (N-1) + (N-1) + (N-1) \\ &+ (N-1)(N-2) + (N-1) + (N-1) + (N-1)(N-2) \\ &+ (N-1) + (N-1) + (N-1)(N-2) + (N-1)(N-2) \\ &+ (N-1)(N-2) + (N-1)(N-2) + (N-1)(N-2)(N-3)] = N^4 \text{ , q.e.d.}\end{aligned}$$

A.3 Population averages

Many population averages have been identified during the calculations. We summarize them here.

$$\langle q_t^{ij}\rangle_P = \frac{1}{N(N-1)} \sum_{p_1} \sum_{p_1'\neq p_1} q_t^{p_1 p_1'}, \tag{A.3.1}$$

$$\langle (q_t^{ij})^2\rangle_P = \frac{1}{N(N-1)} \sum_{p_1} \sum_{p_1'\neq p_1} (q_t^{p_1 p_1'})^2, \tag{A.3.2}$$

$$\langle q_t^{ij} q_t^{kj}\rangle_P = \frac{1}{N(N-1)(N-2)} \sum_{p_1} \sum_{p_1'\neq p_1} \sum_{p_2'\neq p_1,p_1'} q_t^{p_1 p_1'} q_t^{p_1 p_2'}, \tag{A.3.3}$$

$$\langle q_t^{ij} q_t^{kl}\rangle_P = \frac{1}{N(N-1)(N-2)(N-3)} \sum_{p_1} \sum_{p_2\neq p_1} \sum_{p_1'\neq p_1,p_2} \sum_{p_2'\neq p_1,p_2,p_1'} q_t^{p_1 p_1'} q_t^{p_2 p_2'}, \tag{A.3.4}$$

$$\langle q_t^{ijkl}\rangle_P = \frac{1}{N(N-1)(N-2)(N-3)} \sum_{p_1} \sum_{p_2\neq p_1} \sum_{p_1'\neq p_1,p_2} \sum_{p_2'\neq p_1,p_2,p_1'} q_t^{p_1 p_1' p_2 p_2'}. \tag{A.3.5}$$

A.4 Simulations

The set of simulations behind this numerical investigation comprises 50 different runs, up to, approximately, generation 500. In order to calculate the probability distributions shown in figures 7.1, 7.2 and 7.7, all generations in a range t_{ini} to t_{end} were used. t_{ini} corresponds to a chosen generation after the equilibration time. Table A.1 shows these values for the different values of genome size B. It also shows how many speciation events were counted over this time range and across all the simulations and the inferred parameter a of Eq. (7.2.6). The richness in each case was calculated over all the generations in this range.

In order to make it faster, the histograms of Fig. 7.5 were made by considering a reduced number of generations. We calculated the autocorrelation function

$$\rho_X(\tau) = \frac{\langle X_t X_{t+\tau}\rangle - \langle X_t\rangle\langle X_{t+\tau}\rangle}{\sigma_t \sigma_{t+\tau}} \tag{A.4.1}$$

of the number of species $\mathcal{S}$ in each simulation and considered the time lag τ when it first

Table A.1: **Equilibrium intervals and other parameters**. The other parameters of this set of simulations are $N = 400$, $\mu = 0.0025$ and $q_{min} = 0.8$.

Genome Size B	t_{ini}	t_{end}	# Speciation Events	a ($\times 10^{-5}$)
2500	100	500	9519	3.00 ± 0.04
5000	80	500	10735	3.73 ± 0.04
7500	70	470	10536	4.22 ± 0.05
∞	80	500	11721	4.55 ± 0.05

happens that

$$\rho_{\mathcal{S}}(\tau) < \epsilon \tag{A.4.2}$$

with $\epsilon = 1/\sqrt{1 + t_{end} - t_{ini}}$. Then we sampled the distribution of abundances after every τ generations. This is a simple methodology to reduce the amount of data and still have enough unbiased data without the need to run many different uncorrelated simulations.

Appendix B

On the epidemics model

B.1 Recovery probability

Suppose that whenever an individual gets infected, after the onset of symptoms, it is going to remain sick for a period of time τ, distributed according to $\mathcal{P}(\tau)$. Hence, in a simulation, this period may be chosen according to $\mathcal{P}$ for every infected individual, or, at the end of every simulation step, the individual recovers with probability $r(t)$, where t is the time elapsed since the onset of symptoms.

In order to calculate $r(t)$ given $\mathcal{P}$, we consider that an individual is still infected up to time t. Then, the probability of the symptoms to continue after that time is given by

$$1 - r(t) = \mathcal{P}(\tau > t | \tau > t - 1) = \frac{\mathcal{P}(\tau > t, \tau > t - 1)}{\mathcal{P}(\tau > t - 1)} = \frac{\mathcal{P}(\tau > t)}{\mathcal{P}(\tau > t - 1)} \tag{B.1.1}$$

.

If we choose $\mathcal{P}$ as an exponential distribution, we get

$$\mathcal{P}(\tau) = \lambda e^{-\lambda \tau}, \tag{B.1.2}$$

which leads to

$$\mathcal{P}(\tau > s) = e^{-\lambda s}. \tag{B.1.3}$$

Therefore,

$$1 - r(t) = \frac{e^{-\lambda t}}{e^{-\lambda (t-1)}} = e^{-\lambda} = \mathcal{P}(\tau > 1), \tag{B.1.4}$$

which is a constant, as expected due to the *memoryless property* of the exponential distribution function.

Notwithstanding, the mean value of τ equals $1/\lambda$. Supposing $\lambda \sim 10$ days,

$$1 - r(t) \approx 1 - \lambda,$$

and then

$$r(t) = \lambda, \tag{B.1.5}$$

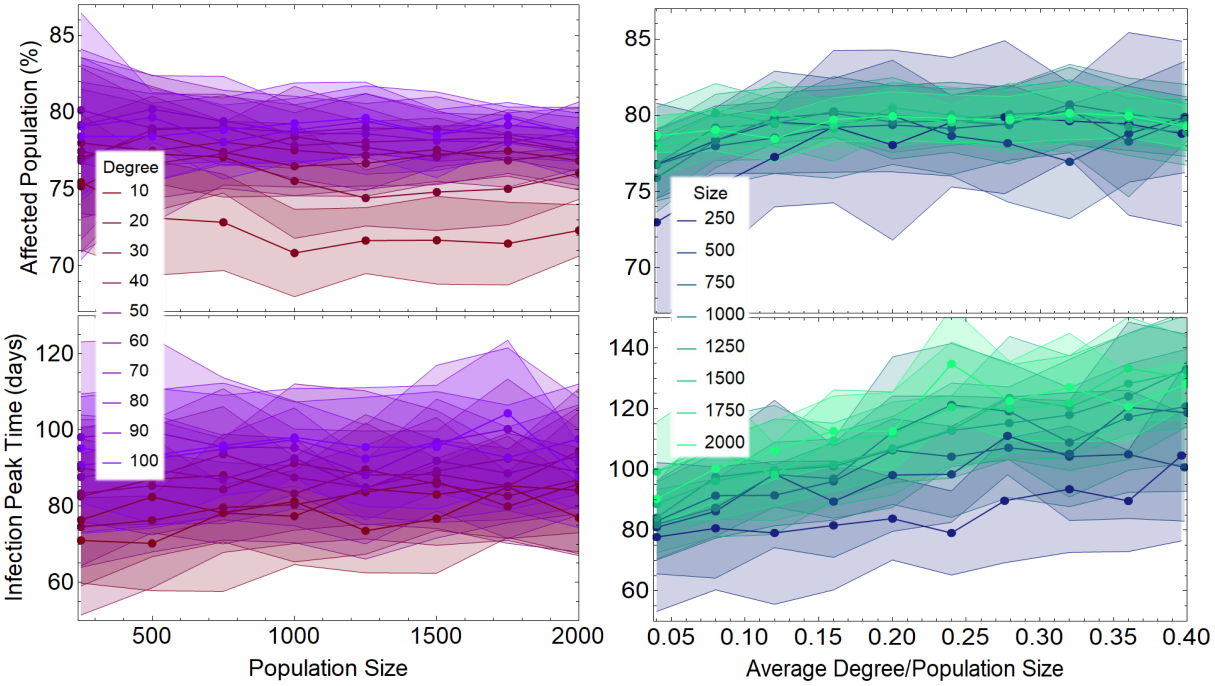


Figure B.1: **Outbreak dependence on network size and average degree.** The figure shows epidemic features for different Barabasi-Albert network parameters (network size and average degree). In purple, the different shades are for different average degrees, while in blue, different shades or for different network sizes. Each point is an average of 20 runs. The simulation parameters are displayed in Table 10.1. The
Source: Figure produced by the author and partially used in close correspondence with the referees of Marquioni and de Aguiar, 2020 [189].

which is what is considered in the model.

B.2 Network size and degree

Whenever the studies we conducted includes an average degree or population size variability, we kept these parameters fixed and equal to $N = 2000$ and $D = 100$, based on a preliminary investigation, whose results (for the spread over Barabasi-Albert networks) are shown in Fig.B.1. In purple shades, different curves have different average degree, while in blue shades, different curves have different population size. We observe that the population size does not significantly affect when the infection has its peak, or the infection size (proportionally to the total population). However, when the peak of infection happens is delayed for greater average degrees.

A reason why the infection peak time is not significantly affected by the population size is due to its rapid (exponential) initial spread, which accounts for many nodes in the network. Notwithstanding, it seems to be counterintuitive that the greater the average degree, the greater is the infection peak time, because in this case, also the number of possible infections from a single infected individual increases, then the spread seems to be faster. But this would happen if the probability of infection would be kept the same, which is not true, since this value is changed to keep R_0 constant across different parameters. Moreover, the more connections a node has, the slower the simulation gets.

Therefore, we chose $N = 2000$ and $D = 100$ because it maintains the simulation times feasible while keeping a proportionally large population (considering the range of parameters we have investigated), not changing in a meaningful way general features of the spread.

Appendix C

On viral diversity

From Marquioni and de Aguiar (Supplemental Material), *PLoS ONE*, 2021 [190].

C.1 Analytical calculations

Our goal is to derive a recurrence equation for the average genetic distance, i.e., given the distance d_t at time t, we aim to calculate the distance d_{t+1} at time $t+1$. The idea is to calculate d_{t+1} as a weighted average, where the weights are the number of pairs that are distanced by a certain amount. In a SEIR model, every iteration starts with a given number of recovered (R_t), infected (I_t) and exposed (E_t) individuals. When an individual recovers, its infecting virus stops to spread and to evolve, and we call it a *final virus*. There are R_t final viruses at the beginning of a given iteration. Viruses infecting Exposed individuals can mutate during this iteration. However, viruses in Infected individuals can either evolve and mutate in this time step or not, since their hosts might recover. The latter become final and are counted as r_t. Infected individuals can also spread the virus, which replicate before evolving or becoming final. Such *offspring* (x_t) increase the number of viruses in Exposed individuals in the next iteration, when they will be allowed to evolve.

At the beginning of iteration $t+1$, there are $(R_t+E_t+I_t)(R_t+E_t+I_t-1)/2$ pairs of viruses sharing an average distance equal to d_t, but along the iteration some of the distances may increase by a certain amount to be calculated, as also new viruses may arise. Therefore,

$$d_{t+1} = \frac{1}{Z'_t}\left(d_t\frac{(R_t+E_t+I_t)(R_t+E_t+I_t-1)}{2} + \text{Increases} + \text{Offspring}\right), \tag{C.1.1}$$

where Z'_t is a normalization factor, which counts the total number of pairs at the end of iteration $t+1$,

$$Z'_t = \frac{(R_t+E_t+I_t+x_t)(R_t+E_t+I_t+x_t-1)}{2}. \tag{C.1.2}$$

If the mutation rate is zero and no new infections occur ($x_t = 0$) the "Increases" term and the "Offspring" term are equal to zero, and $d_{t+1} = d_t$, as expected.

In the following two subsections, we shall calculate the "Increases" term and the "Offspring" term, which accounts for the evolution and for the spread, respectively.

C.1.1 Increases

Genetic distances between evolving viruses increase over time. In order to calculate how much these distances increase we first consider that mutations occurring in the same locus of different genomes are unlikely, as well as more than one mutation per locus on a single genome. This approximation holds as long as the epidemic duration T remains sufficiently small, $\mu T \ll 1$. Thus, after one time step, an evolving genome acquires, on average, $B\mu$ mutations. The distance between two evolving genomes will increase, on average, by $2B\mu$ nucleotides after one time step. The distance between viruses in exposed individuals, for example, increases by $2B\mu$ and because there are $E_t(E_t-1)/2$ pairs of exposed individuals, their evolution along the iteration $t+1$ contributes $2B\mu E_t(E_t-1)/2$ to the Increases term. On the other hand, the distance between viruses in an exposed and a recovered individual, or an infected individual that recovers, is only $B\mu$, because the latter two do not evolve. There are $E_t(R_t+r_t)$ pairs among these viruses, and thus their contribution to Increases is $E_t(R_t+r_t)B\mu$. We recall that the updates in our model occur in the order "Transmission", "Attempt to Recovery" and lastly, "Genome Evolution". Thus, if an infected individual recovers its virus does not have the chance to mutate.

Therefore, in order to compute the Increases term, we must calculate the average increase in distance between all pairs of viruses and how many pairs of these viruses exist. Table C.1 summarizes this information. We obtain

$$\begin{aligned}\text{Increases} =& E_tR_tB\mu + E_tr_tB\mu + (I_t-r_t)r_tB\mu + (I_t-r_t)R_tB\mu + (I_t-r_t)E_t2B\mu \\ &+\frac{E_t(E_t-1)}{2}2B\mu + \frac{(I_t-r_t)(I_t-r_t-1)}{2}2B\mu. \end{aligned} \tag{C.1.3}$$

Table C.1: **Increases in average distance and number of pairs of viruses.** From Marquioni and de Aguiar, (Supplemental Material), 2021 [190].

Viruses	Number of Pairs	Average Distance Increase
(E_t) and (R_t)	E_tR_t	$B\mu$
(E_t) and (r_t)	E_tr_t	$B\mu$
(I_t-r_t) and (r_t)	$(I_t-r_t)r_t$	$B\mu$
(I_t-r_t) and (R_t)	$(I_t-r_t)R_t$	$B\mu$
(I_t-r_t) and (E_t)	$(I_t-r_t)E_t$	$2B\mu$
(E_t) and (E_t)	$E_t(E_t-1)/2$	$2B\mu$
(I_t-r_t) and (I_t-r_t)	$(I_t-r_t)(I_t-r_t-1)/2$	$2B\mu$
(R_t) and (R_t)	$R_t(R_t-1)/2$	0
(r_t) and (r_t)	$r_t(r_t-1)/2$	0
(r_t) and (R_t)	r_tR_t	0

C.1.2 Offspring

The contribution of the new infections to the average distance d_{t+1}, the Offspring term, is more tricky. To simplify matters we will assume that an infected individual infects only one susceptible per time step, which is a good assumption if the basic reproduction number R_0 is small compared to the average duration of symptoms. Thus, x_t is also the number of individuals who infected a susceptible within the time step $t+1$, which will be called *ancestors* from now on. Let D_1 be the average distance between ancestors and the other viruses at time t, and D_2, the distance between the exposed and the other viruses. Note that an ancestor may recover and, therefore, not mutate in this time step. The Offspring term is a sum of different contributions between offspring and the other viruses in the population, as explained in detail below.

1. *Genetic distance between offspring and recovered.* The number of pairs is $x_t R_t$. Because offspring do not evolve in the time step they appear, their average distance is D_1. Then, its contribution to the Offspring term is $\boldsymbol{x_t R_t D_1}$.

2. *Genetic distance between offspring and exposed.* The number of pairs is $x_t E_t$. Because the exposed evolve, these pairs contribute with $\boldsymbol{x_t E_t (D_2 + B\mu)}$ to the Offspring term.

3. *Genetic Distance between offspring of an infected (ancestor) that does not recover (there are $(I_t - r_t)$ of these individuals) and infected*:

   (a) The distance between an offspring and its ancestor is $B\mu$, since the ancestor evolves. There are $x_t(I_t - r_t)/I_t$ new infections of this type, contributing with $\boldsymbol{x_t((I_t - r_t)/I_t)B\mu}$ to the distance.

   (b) For each offspring there are $I_t - r_t - 1$ infected individuals that did not recover and are not its ancestral. The distance between the offspring and these individuals is $(D_1 + B\mu)$, adding $\boldsymbol{x_t((I_t - r_t)/I_t)(I_t - r_t - 1)(D_1 + B\mu)}$ to the Offspring term.

   (c) The distance between the offspring and individuals that recover is D_1, because neither of these viruses evolve in this time step. There are $x_t((I_t - r_t)/I_t)r_t$ pairs of these viruses, adding $\boldsymbol{x_t((I_t - r_t)/I_t)r_t D_1}$ to the Offspring term.

4. *Genetic distance between offspring of infected (ancestor) that recover in this iteration (there are r_t of these individuals) and infected*:

   (a) The distance between offspring and its ancestor is zero, because none of them evolve.

   (b) The distance between the offspring and the other viruses of type is D_1. There are $x_t r_t / I_t$ new infections of this type, contributing $\boldsymbol{(x_t r_t/I_t)(r_t - 1)D_1}$ to the Offspring term.

   (c) The distance between offspring and the other infected individuals is $\boldsymbol{(x_t r_t/I_t)(I_t - r_t)(D_1 + B\mu)}$, since the other infected viruses evolve..

5. *Genetic distance between offspring.* Because each ancestor gives rise to only one new infection, this distance equals D_1, and once there are $x_t(x_t - 1)/2$ pairs of offspring, this contribution is $\boldsymbol{(x_t(x_t - 1)/2)D_1}$.

6. By summing everything up, we get

$$\begin{aligned}\text{Offspring} = {} & x_t R_t D_1 + x_t E_t (D_2 + B\mu) \\ & + x \frac{(I_t - r_t)}{I_t} B\mu + x_t \frac{(I_t - r_t)}{I_t}(I_t - r_t - 1)(D_1 + B\mu) + x_t \frac{(I_t - r_t)}{I_t} r_t D_1 \\ & + x_t \frac{r_t}{I_t} 0 + x_t \frac{r_t}{I_t}(r_t - 1) D_1 + x_t \frac{r_t}{I_t}(I_t - r_t)(D_1 + B\mu) \\ & + \frac{x_t(x_t - 1)}{2} D_1. \end{aligned} \tag{C.1.4}$$

Putting all these terms together and defining $Z_t \equiv 2Z'_t$ we obtain

$$\begin{aligned} d_{t+1} = \frac{1}{Z_t} \big(& d_t (R_t + E_t + I_t)(R_t + E_t + I_t - 1) \\ & + x_t D_1 (x_t - 3 + 2R_t + 2I_t + 2E_t D_2 / D_1) \\ & + 2B\mu (E_t + I_t - r_t)(E_t + I_t + R_t + x_t - 1) \big). \end{aligned} \tag{C.1.5}$$

The reason for assigning the distance D_1 between infected and other viruses, instead of d_t, is that infected individuals represent only a fraction of the viruses in the population, and the distance between them and other viruses grows over time, therefore being above the average d_t. The same holds for the exposed individuals.

Although we were not able to analytically find an expression for D_1 and D_2, we can approximate them as follows. First we assume that $D_2 \approx D_1$. When the epidemic begins, all viruses are infected, so that $D_1 = d_t$. However, the ratio between infected and recovered individuals decreases to zero along the epidemic, making D_1 larger than d_t. Thus, to first order, it is possible to approximate $D_1 \approx d_t(1 + \epsilon)$, with ϵ a function of the number of recovered individuals, $R_t/(I_t + E_t + R_t)$ and the average number of mutations $B\mu$. Our simulations showed that the linear function $D_1 = d_t(1 + 2B\mu R_t/(I_t + E_t + R_t))$ works well (considering the parameters in Appendix A), leading to the theoretical result expressed by Eq.(12.1.1) from the main text.

C.1.3 Continuum Limit

To achieve the continuum limit we start by substituting $r_t = R_{t+1} - R_t$ and $x_t = E_{t+1} - E_t + I_{t+1} - I_t + R_{t+1} - R_t$ in Eq.(12.1.1) from the main text and subtracting d_t from both sides of this equation:

$$\begin{aligned} d_{t+1} - d_t = \frac{1}{Z_t} \big\{ & 2 d_t \left(E_{t+1} - E_t + I_{t+1} - I_t + R_{t+1} - R_t \right) \times \\ & \times \left[-1 + B\mu \frac{R_t}{I_t + E_t + R_t} \left(R_{t+1} + R_t + I_{t+1} + I_t + E_{t+1} + E_t - 3 \right) \right] \\ & + 2B\mu \left(E_t + I_t + R_t - R_{t+1} \right) \left(E_{t+1} + I_{t+1} + R_{t+1} - 1 \right) \big\} \end{aligned} \tag{C.1.6}$$

with

$$Z_t = (E_{t+1} + I_{t+1} + R_{t+1})(E_{t+1} + I_{t+1} + R_{t+1} - 1). \tag{C.1.7}$$

Then, we consider the first order approximations

$$\begin{aligned} f_t &\approx f(t) \\ f_{t+1} &\approx f(t) + \dot{f}(t)\Delta t, \end{aligned}$$

and once $B\mu$ in the last line of Eq.(C.1.6) is the number of mutations per time step, we replace it by $B\mu\Delta t$

$$\dot{d}(t)\Delta t = \tag{C.1.8}$$

$$\begin{aligned} &\frac{1}{Z_t}\Big\{2d(t)\Delta t\left(\dot{E}(t) + \dot{I}(t) + \dot{R}(t)\right) \times \\ &\times\left[-1 + B\mu\frac{R(t)}{I(t)+E(t)+R(t)}\left(2R(t) + 2I(t) + 2E(t) + \Delta t(\dot{E}(t) + \dot{I}(t) + \dot{R}(t)) - 3\right)\right] \\ &+ 2B\Delta t\mu\left(E(t) + I(t) - \dot{R}(t)\Delta(t)\right)\left(R(t) + I(t) + E(t) + \Delta t(\dot{E}(t) + \dot{I}(t) + \dot{R}(t)) - 1\right)\Big\} \end{aligned} \tag{C.1.9}$$

with

$$\begin{aligned} Z_t = &(R(t) + I(t) + E(t) + \Delta t(\dot{E}(t) + \dot{I}(t) + \dot{R}(t)))\times \\ &\times (R(t) + I(t) + E(t) + \Delta t(\dot{E}(t) + \dot{I}(t) + \dot{R}(t)) - 1). \end{aligned} \tag{C.1.10}$$

Finally, by taking the limit $\Delta t \to 0$ we obtain the continuous time equation.

C.1.4 Multiple Infections

The average distance $d^{(i)}_{root,t}$ between viruses from a lineage and its root is calculated using the same technique discussed above, however it is much simpler, once we only need to calculate the average distance from a kind of virus and the root (a single virus which does not evolve). Using the same notation, but now with a super-index to denote the lineage, we obtain

$$d^{(i)}_{root,t+1} = \frac{1}{Z_t}\left[\left(R^{(i)}_t + E^{(i)}_t + I^{(i)}_t\right) d^{(i)}_{root,t} + E^{(i)}_t B\mu + \left(I^{(i)}_t - r^{(i)}_t\right) B\mu + x^{(i)}_t D^{(i)}_{1,root}\right] \tag{C.1.11}$$

with $Z_t = (E_t^{(i)} + I_t^{(i)} + R_t^{(i)} + x_t^{(i)})$ and $D_{1,root}^{(i)}$ being the average distance between infected and the root, which is given (similarly to D_1) by

$$D_{1,root}^{(i)} = d_{root,t}^{(i)} \left(1 + 4B\mu \frac{R_t^{(i)}}{E_t^{(i)} + I_t^{(i)} + R_t^{(i)}} \right).$$

The factor 4 is a fit from numerical investigations. The continuum limit is obtained by subtracting $d_{root,t}^{(i)}$ from both sides of Eq.(12.1.5) from the main text, applying the continuous approximation for each epidemic curve and taking the limit $\Delta t \to 0$.

C.2 Real genetic evolution algorithm

In order to estimate the real (from genetic data) genetic evolution, we used 55 complete genome sequences collected in China[210]. First, these sequences were ordered and numbered by its collection date and a matrix of genetic distances d_{ij} between genomes i and j has been constructed. Each pair of sequences were alligned with the Needleman-Wunsch algorithm, with score $+1$ for match and -1 for mismatch[211]. Then, the distance between two genomes was computed counting the number of substitutions between the sequences, neglecting *indels*.

We defined a time window $\tau_W = 14 = \tau_0$ days. Thus, every genome collected within τ_W are considered infected, and the genomes collected before this time window are considered recovered. Now, we calculate the average distance among the infected $d_{I,t}$, recovered $d_{R,t}$ and among infected and recovered $d_{IR,t}$ at the time t. Fig.C.1 shows an example of a distance matrix with a specific time window. Finally, the average distance at time t can be computed as

$$d_t = \frac{d_{I,t} I_t (I_t - 1) + 2 d_{IR,t} I_t R_t + d_{R,t} R_t (R_t - 1)}{(R_t + I_t)(R_t + I_t - 1)} \tag{C.2.1}$$

where I_t and R_t are respectively given by $I(t)$ and $R(t)$ described evaluated in the supplemental material.

With this algorithm, we obtained 20 non-overlapping sets of infected genomes. One of these sets contained only one sequence and was not usable; a second set was too far from all other data and was also discarded. Thus, we were able to calculate 18 points (that appear in Fig.12.3 from the main text) with error bars given by the standard deviation of each set of distances (between infected, recovered and between infected and recovered) at each time t.

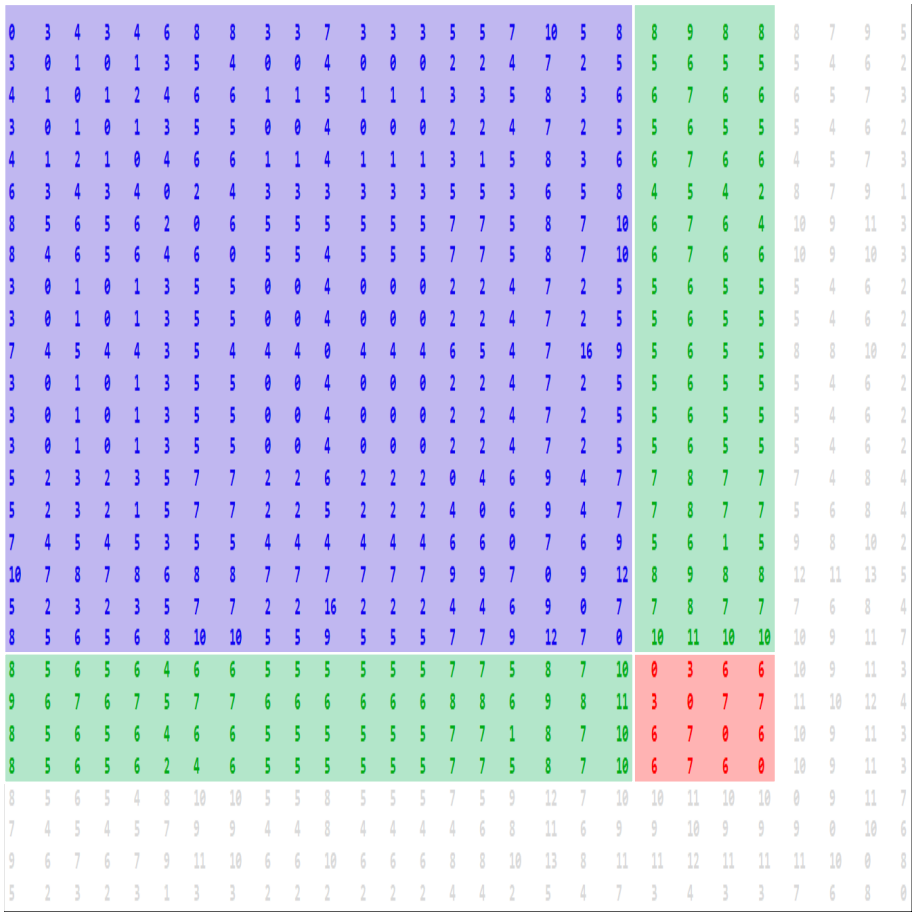


Figure C.1: **Example of distance matrix to illustrate the algorithm to infer the genetic evolution.** Every genome collected within a time window τ_W is considered to belong to an infected individual. The red block shows distances between these viruses. The blue block shows viruses that appeared before the present time window, whose individuals are considered to have recovered. Green blocks are distances between infected and recovered individuals. The remaining entries are distances from viruses that have not appeared yet at that considered time, i.e., they appeared after the considered time window.
Source: Figure from Marquioni and de Aguiar, (Supplemental Material), 2021 [190].

C.3 The COVID-19 data from China

We got the Chinese epidemic data from the dataset “Epidemic Data for Novel Coronavirus COVID-19" from Wolfram data repository[208]. Unfortunately, this dataset starts on 22 January (going up to 18 August by the date of our analysis), lacking the previous data. Another concern is about the change in the notification protocols adopted by the Chinese government. On 13 February, the Hubei province started to report not only the positive laboratory tests, but also the clinically diagnosed cases as infected too, appearing a sudden increase in infected curve[144]. We also need to correct the data by including undetected cases.

Firstly, in order to correct the notification problem, we smoothly distribute the sudden increase number of cases among the previous dates. Following reference [144], the

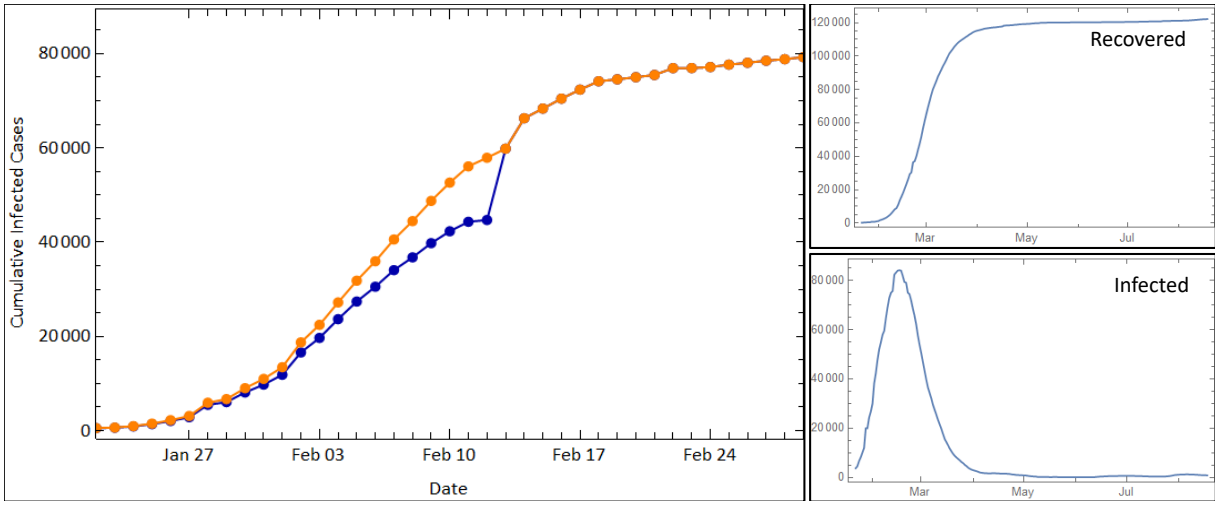


Figure C.2: **Chinese epidemic curves after corrections.** The left chart shows the cumulative number of infections in China. The blue curve is the reported number of cases before the smoothness procedure of Eq.(C.3.1) and the orange curve is the result of this procedure. The right charts are the recovered and infected curves $R(t)$ and $I(t)$.
Source: Figure from Marquioni and de Aguiar, (Supplemental Material), 2021 [190].

corrected accumulated number of cases $I_{a,c}(t)$ is given by

$$I_{a,c}(t) = I_a(t) + 15133 \frac{\sum_{i=22\text{ Jan}}^{t} I_a(t)}{\sum_{i=22\text{ Jan}}^{13\text{ Feb}} I_a(t)} \tag{C.3.1}$$

for $t \in \{22\text{ Jan}, \ldots, 12\text{ Feb}\}$, where $I_c(t)$ is the accumulated number of cases at date t, and $15133 = I_a(13\text{ Feb}) - I_a(12\text{ Feb})$ is the sudden increase due to the changes in the notification protocol.

Now, the undetected cases in China were estimated in reference [209], and also following reference [144], we get

$$I_{a,c'}(t) = \frac{I_{a,c}(t)}{1 - \theta(t)} \tag{C.3.2}$$

for the estimated total number of cases at time t, where θ is the undetected fraction,

$$\theta(t) = \begin{cases} 0.86, \text{ for } t \leq 24\text{ Jan} \\ \textit{linear decrease}, \text{ for } 24\text{ Jan} \leq t \leq 08\text{ Feb} \\ 0.31, \text{ for } t \geq 08\text{ Feb} \end{cases} \tag{C.3.3}$$

This correction is also applied to the recovered curve. However, the Wolfram data distincts recovered $Rec(t)$ from deaths $Dea(t)$, while our theory does not differentiates these numbers. Thus, the number of recovered individuals we must consider is

$$R(t) = \frac{Rec(t) + Dead(t)}{1 - \theta(t)} \tag{C.3.4}$$

and the infected curve is now given as

$$I(t) = I_{a,c'}(t) - R(t) \tag{C.3.5}$$

Fig.C.2 shows the curves after these corrections. Once we do no have directly access to exposed data, we did not consider exposed individuals, meaning, at this point, that we are dealing with a SIR model without any prejudice to the present theory. However, bad data is an important source of error.

Finally, we fit an exponential curve to a few initial data points of $I(t)$ and $R(t)$ and extrapolate it to previous dates. For the I-curve, we have adjusted the exponential $e^{a(t-t_0)}$, with fit parameters a and t_0, on the first $n_I = 10$ data points and extrapolated it up to the first case t_0 days before. With this approach, we found $t_0 = 11$ Dec, which is close to the first case reported by WHO, 08 Dec [221]. For the $R(t)$-curve, we have used the first $n_R = 13$ data points. The numbers n_I and n_R were chosen in order to make the exponential extrapolation makes sense according to WHO estimates of the first case, as also to make $R(t) < I(t)$ in a plausible way.

Now, the curves $R(t)$ and $I(t)$ can be implemented in the recurrence equation and the distance evolution can be estimated, with the first distance d_0 equalling zero.

Appendix D

Genome Data

From Marquioni and de Aguiar (Supplemental Material), *PLoS ONE*, 2021 [190].

D.1 Data table 1

D.1 Table. All Chinese genome sequences. All genomes registered in Wolfram Repository "Genetic Sequences for the SARS-CoV-2 Coronavirus" with complete *NucleotideStatus* and human *Host* from China (data accessed 19/08/2020). From Marquioni and de Aguiar, (Supplemental Material), 2021 [190].

Accession Number	Collection Date	Length	Geographic Location	Included?	Justification
MN908947	26 Dec 2019*	29903	Wuhan, Hubei**	Y	
MN938384	10 Jan 2020	29838	Shenzen, Guangdong	Y	
MN975262	11 Jan 2020	29891	Wuhan, Hubei**	Y	
MN988668	02 Jan 2020	29881	Wuhan, Hubei**	Y	
MN988669	02 Jan 2020	29881	Wuhan, Hubei**	Y	
MN996527	30 Dec 2019	29825	Wuhan, Hubei	Y	
MN996528	30 Dec 2019	29891	Wuhan, Hubei	Y	
MN996529	30 Dec 2019	29852	Wuhan, Hubei	Y	
MN996530	30 Dec 2019	29854	Wuhan, Hubei	Y	
MN996531	30 Dec 2019	29857	Wuhan, Hubei	Y	
MT019529	23 Dec 2019	29899	Wuhan, Hubei	Y	
MT019530	30 Dec 2019	29889	Wuhan, Hubei	N	MT19530 to MT19532: Might be biased with no other informations data (sequences from the same researchers, collected at the same day and with uite the same length, up to the date we have made the analysis).*
MT019531	30 Dec 2019	29899	Wuhan, Hubei	N	
MT019532	30 Dec 2019	29890	Wuhan, Hubei	N	

MT019533	01 Jan 2020	29883	Wuhan, Hubei	Y	
MT034054	03 Jan 2020	29885	Beijing	Y	
MT039873	20 Jan 2020	29833	Hangzhou, Zhejiang	Y	
MT039874	22 Jan 2020	29858	Hangzhou, Zhejiang**	Y	
MT049951	17 Jan 2020	29903	Kunming,† Yunnan	Y	
MT079843	22 Jan 2020	29915	Wuhan, Hubei**	Y	MT079843 to MT079854:
MT079844	22 Jan 2020	29910	Wuhan, Hubei**	N	Might be biased data
MT079845	22 Jan 2020	29955	Wuhan, Hubei**	N	(probable nosocomial
MT079846	22 Jan 2020	29903	Wuhan, Hubei**	N	transmission).**
MT079847	22 Jan 2020	29872	Wuhan, Hubei**	N	Then we have included
MT079848	22 Jan 2020	29880	Wuhan, Hubei**	N	only one genome.
MT079849	22 Jan 2020	29904	Wuhan, Hubei**	N	
MT079850	22 Jan 2020	29885	Wuhan, Hubei**	N	
MT079851	22 Jan 2020	30018	Wuhan, Hubei**	N	
MT079852	22 Jan 2020	29891	Wuhan, Hubei**	N	
MT079853	22 Jan 2020	29766	Wuhan, Hubei**	N	
MT079854	22 Jan 2020	29897	Wuhan, Hubei**	N	
MT093631	08 Jan 2020	29860	Beijing†	N	No detailed geographic information available.
MT121215	02 Feb 2020	29945	Shanghai	Y	
MT123290	05 Feb 2020	29891	Guangzhou, Guangdong	Y	
MT123291	29 Jan 2020	29882	Guangzhou, Guangdong	Y	
MT123292	27 Jan 2020	29923	Guangzhou, Guangdong	Y	
MT123293	29 Jan 2020	29871	Guangzhou, Guangdong	Y	
MT135041	26 Jan 2020	29903	Beijing	N	MT135041 to MT135044:
MT135042	28 Jan 2020	29903	Beijing	N	Might be biased data
MT135043	28 Jan 2020	29903	Beijing	N	(the lengths are all
MT135044	28 Jan 2020	29903	Beijing	Y	the same). Then we have included only one genome.
MT226610	20 Jan 2020	29899	Kunming, Yunnan†	N	No detailed geographic

					information available.
MT253696	23 Jan 2020	29781	Hangzhou, Zhejiang	N	MT253696 to MT253710:
MT253697	23 Jan 2020	29781	Hangzhou, Zhejiang	N	Might be biased data
MT253698	23 Jan 2020	29781	Hangzhou, Zhejiang	N	(cluster of cases*;
MT253699	24 Jan 2020	29781	Hangzhou, Zhejiang	N	also they all have
MT253700	25 Jan 2020	29781	Hangzhou, Zhejiang	N	the same length).
MT253701	21 Jan 2020	29781	Hangzhou, Zhejiang	N	Then we have included
MT253702	21 Jan 2020	29781	Hangzhou, Zhejiang	N	only one genome.
MT253703	25 Jan 2020	29781	Hangzhou, Zhejiang	N	
MT253704	25 Jan 2020	29781	Hangzhou, Zhejiang	N	
MT253705	22 Jan 2020	29781	Hangzhou, Zhejiang	N	
MT253706	22 Jan 2020	29781	Hangzhou, Zhejiang	N	
MT253707	25 Jan 2020	29781	Hangzhou, Zhejiang	N	
MT253708	21 Jan 2020	29781	Hangzhou, Zhejiang	N	
MT253709	21 Jan 2020	29781	Hangzhou, Zhejiang	N	
MT253710	21 Jan 2020	29781	Hangzhou, Zhejiang	Y	
MT259226	10 Jan 2020	29868	Wuhan, Hubei	Y	
MT259227	26 Jan 2020	29863	Wuhan, Hubei	Y	
MT259228	26 Jan 2020	29861	Wuhan, Hubei	Y	
MT259229	26 Jan 2020	29864	Wuhan, Hubei	Y	
MT259230	25 Jan 2020	29866	Wuhan, Hubei	Y	
MT259231	25 Jan 2020	29865	Wuhan, Hubei	Y	
MT281577	10 Mar 2020	29903	Fujyang, Anhui	Y	
MT291826	30 Dec 2019	29807	Wuhan, Hubei	Y	
MT291827	30 Dec 2019	29858	Wuhan, Hubei	Y	
MT291828	30 Dec 2019	29858	Wuhan, Hubei	Y	
MT291829	30 Dec 2019	29774	Wuhan, Hubei	Y	
MT291830	30 Dec 2019	29807	Wuhan, Hubei	Y	
MT291831	24 Jan 2020	29872	Beijing	Y	
MT291832	25 Jan 2020	29828	Beijing	Y	
MT291833	28 Jan 2020	29821	Beijing	Y	

MT291834	28 Jan 2020	29865	Beijing	Y	
MT291835	27 Jan 2020	29834	Beijing	Y	
MT291836	29 Jan 2020	29860	Beijing	Y	
MT407649	22 Jan 2020	29833	Hangzhou,† Zhejiang	Y	
MT407650	22 Jan 2020	29821	Hangzhou,† Zhejiang	Y	
MT407651	22 Jan 2020	29822	Hangzhou,† Zhejiang	Y	
MT407652	26 Jan 2020	29835	Hangzhou,† Zhejiang	Y	
MT407653	26 Jan 2020	29835	Hangzhou,† Zhejiang	Y	
MT407654	24 Mar 2020	29817	Hangzhou,† Zhejiang	Y	
MT407655	24 Mar 2020	29817	Hangzhou,† Zhejiang	Y	
MT407656	24 Mar 2020	29835	Hangzhou,† Zhejiang	Y	
MT407657	24 Mar 2020	29776	Hangzhou,† Zhejiang	Y	
MT407658	24 Mar 2020	29770	Hangzhou,† Zhejiang	Y	
MT407659	24 Mar 2020	29828	Hangzhou,† Zhejiang	Y	
MT412134	24 Feb 2020	29867	Zhengzhou, Henan†	N	No detailed geographic information available.
MT446312	05 Feb 2020	29879	Guangzhou, Guangdong	Y	
MT510727	15 Feb 2020	29903	Meizhou, Guangdong†	N	MT510727 and MT510728:
MT510728	13 Feb 2020	29903	Meizhou, Guangdong†	N	Might be biased data (data from familial cluster*). There is also no detailed geographic information available.
MT534630	26 Jan 2020	29845	Changzhou,Jiangsu	Y	
MT568634	25 Feb 2020	29861	Guangzhou, Guangdong	N	MT568634 to MT568641:
MT568635	25 Feb 2020	29854	Guangzhou, Guangdong	N	data from a work
MT568636	27 Feb 2020	29858	Guangzhou, Guangdong	N	presenting different
MT568637	25 Feb 2020	29860	Guangzhou, Guangdong	N	approaches for genome
MT568638	25 Feb 2020	29854	Guangzhou, Guangdong	N	sequencing.** Then,
MT568639	25 Feb 2020	29861	Guangzhou, Guangdong	N	this data might have

MT568640	25 Feb 2020	29858	Guangzhou, Guangdong	N	more errors than the
MT568641	25 Feb 2020	29868	Guangzhou, Guangdong	N	others.
MT622319	23 Jan 2020	29889	Shanghai†	N	No detailed geographic information available.
MT627325	28 Feb 2020	29859	Shanghai†	N	No detailed geographic information available.
NC045512	Dec 2019	29903	Wuhan, Hubei**	N	Identical to MN908947.*

**GenBank* information.
**Publication Information.
† Laboratory address.

D.2 Data table 2

D.2 Table. Included sequences sorted by Collection Date. All informations according to D.1 Table. From Marquioni and de Aguiar, (Supplemental Material), 2021 [190].

Number	Accession Number	Collection Date	Length	Geographic Location
#1	MT019529	23 Dec 2019	29899	Wuhan, Hubei
#2	MN908947	26 Dec 2019	29903	Wuhan, Hubei
#3	MT291829	30 Dec 2019	29774	Wuhan, Hubei
#4	MT291826	30 Dec 2019	29807	Wuhan, Hubei
#5	MT291830	30 Dec 2019	29807	Wuhan, Hubei
#6	MN996527	30 Dec 2019	29825	Wuhan, Hubei
#7	MN996529	30 Dec 2019	29852	Wuhan, Hubei
#8	MN996530	30 Dec 2019	29854	Wuhan, Hubei
#9	MN996531	30 Dec 2019	29857	Wuhan, Hubei
#10	MT291827	30 Dec 2019	29858	Wuhan, Hubei
#11	MT291828	30 Dec 2019	29858	Wuhan, Hubei
#12	MN996528	30 Dec 2019	29891	Wuhan, Hubei
#13	MT019533	01 Jan 2020	29883	Wuhan, Hubei
#14	MN988668	02 Jan 2020	29881	Wuhan, Hubei
#15	MN988669	02 Jan 2020	29881	Wuhan, Hubei
#16	MT034054	03 Jan 2020	29885	Beijing
#17	MN938384	10 Jan 2020	29838	Shenzhen, Guangdong
#18	MT259226	10 Jan 2020	29868	Wuhan, Hubei
#19	MN975262	11 Jan 2020	29891	Wuhan, Hubei
#20	MT049951	17 Jan 2020	29903	Yunnan
#21	MT039873	20 Jan 2020	29833	Hangzhou, Zhejiang
#22	MT253710	21 Jan 2020	29781	Hangzhou, Zhejiang
#23	MT407650	22 Jan 2020	29821	Zhejiang
#24	MT407651	22 Jan 2020	29822	Zhejiang
#25	MT407649	22 Jan 2020	29833	Zhejiang
#26	MT039874	22 Jan 2020	29858	Hangzhou, Zhejiang
#27	MT079843	22 Jan 2020	29915	Wuhan, Hubei
#28	MT291831	24 Jan 2020	29872	Beijing
#29	MT291832	25 Jan 2020	29828	Beijing
#30	MT259231	25 Jan 2020	29865	Wuhan, Hubei
#31	MT259230	25 Jan 2020	29866	Wuhan, Hubei
#32	MT407652	26 Jan 2020	29835	Zhejiang
#33	MT407653	26 Jan 2020	29835	Zhejiang
#34	MT534630	26 Jan 2020	29845	Changzhou, Jiangsu
#35	MT259228	26 Jan 2020	29861	Wuhan, Hubei
#36	MT259227	26 Jan 2020	29863	Wuhan, Hubei
#37	MT259229	26 Jan 2020	29864	Wuhan, Hubei
#38	MT291835	27 Jan 2020	29834	Beijing
#39	MT123292	27 Jan 2020	29923	Guangzhou, Guangdong
#40	MT291833	28 Jan 2020	29821	Beijing
#41	MT291834	28 Jan 2020	29865	Beijing

#42	MT135044	28 Jan 2020	29903	Beijing
#43	MT291836	29 Jan 2020	29860	Beijing
#44	MT123293	29 Jan 2020	29871	Guangzhou, Guangdong
#45	MT123291	29 Jan 2020	29882	Guangzhou, Guangdong
#46	MT121215	02 Feb 2020	29945	Shanghai
#47	MT446312	05 Feb 2020	29879	Guangzhou, Guangdong
#48	MT123290	05 Feb 2020	29891	Guangzhou, Guangdong
#49	MT281577	10 Mar 2020	29903	Fuyang, Anhui
#50	MT407658	24 Mar 2020	29770	Zhejiang
#51	MT407657	24 Mar 2020	29776	Zhejiang
#52	MT407654	24 Mar 2020	29817	Zhejiang
#53	MT407655	24 Mar 2020	29817	Zhejiang
#54	MT407659	24 Mar 2020	29828	Zhejiang
#55	MT407656	24 Mar 2020	29835	Zhejiang

D.3 Data table 3

D.3 Table. Genome information used to calculate points in Fig.12.3. We have used a 14 days time window, i.e., every sequenced genome within an interval of 14 days were considered as infected ones, while the previous were considered to be recovered. From Marquioni and de Aguiar, (Supplemental Material), 2021 [190].

Point Number*	Infected Genomes	Recovered Genomes	Date Interval
#1	#03 → #19	#1 → #02	30 Dec 2019 → 12 Jan 2020
#2	#13 → #19	#1 → #12	01 Jan 2020 → 14 Jan 2020
#3	#14 → #19	#1 → #13	02 Jan 2019 → 15 Jan 2020
#4	#16 → #19	#1 → #15	03 Jan 2019 → 16 Jan 2020
#5	#17 → #27	#1 → #16	10 Jan 2019 → 23 Jan 2020
#6	#19 → #28	#1 → #18	11 Jan 2019 → 24 Jan 2020
#7	#20 → #45	#1 → #19	17 Jan 2019 → 30 Jan 2020
#8	#21 → #46	#1 → #20	20 Jan 2019 → 02 Feb 2020
#9	#22 → #46	#1 → #21	21 Jan 2019 → 03 Feb 2020
#10	#23 → #46	#1 → #22	22 Jan 2019 → 04 Feb 2020
#11	#28 → #48	#1 → #27	24 Jan 2019 → 06 Feb 2020
#12	#29 → #48	#1 → #28	25 Jan 2019 → 07 Feb 2020
#13	#32 → #48	#1 → #31	26 Jan 2019 → 08 Feb 2020
#14	#38 → #48	#1 → #37	27 Jan 2019 → 09 Feb 2020
#15	#40 → #48	#1 → #39	28 Jan 2019 → 10 Feb 2020
#16	#43 → #48	#1 → #42	29 Jan 2019 → 11 Feb 2020
#17	#46 → #48	#1 → #45	02 Feb 2019 → 15 Feb 2020
#18	#47 → #48	#1 → #46	05 Feb 2019 → 18 Feb 2020
#19**	#49 → #49	#1 → #48	
#20†	#50 → #55	#1 → #49	24 Mar 2019 → 06 Apr 2020

*In Fig.12.3 from the main text, points are numbered from left to right.
**Since there is only one genome in this time window, we cannot estimate a distance among the infected population, so genome #49 was not used.
† This point was not included in Fig.12.3 because it is lacking more than one month of genetic information between points #18 and #19, therefore the distance among the recovered population cannot be well inferred.