Università degli Studi di Trieste

Dipartimento di Fisica
Corso di Laurea Magistrale in Fisica

GENETICA STATISTICA E INFERENZA
DCA OLTRE IL QUASI-LINKAGE
EQUILIBRIUM

Laureando:
Vito Dichio

Matricola:
SM2300430

Relatore:
prof. Erik Aurell

Correlatrice:
prof.ssa Hongli Zeng

Anno accademico 2019/2020
Vito Dichio

Statistical Genetics and DCA Inference Beyond the Quasi-Linkage Equilibrium

Titleback
This document was written with \LaTeX on Mac using ArsClassica, a reworking of the classicthesis v4.6 style designed by André Miede, inspired to the masterpiece The Elements of Typographic Style by Robert Bringhurst.

Contacts
dichio.vito@gmail.com
A man sets out to draw the world. As the years go by, he peoples a space with images of provinces, kingdoms, mountains, bays, ships, islands, fishes, rooms, instruments, stars, horses, and individuals. A short time before he dies, he discovers that the patient labyrinth of lines traces the lineaments of his own face.

— Jorge Luis Borges
Consideriamo qui la teoria NS, un modello statistico per la genetica delle popolazioni introdotto da Neher & Shraiman (2011) che descrive il processo evolutivo sotto l’azione di mutazioni, ricombinazioni, deriva genetica, selezione naturale (epistasi con interazioni a coppie). In particolare, consideriamo il Quasi-Linkage Equilibrium (QLE), che si ha per un tasso di ricombinazione $r$ sufficientemente alto rispetto alla forza della selezione $\sigma$. Combinando i risultati della teoria NS con le tecniche di Direct Coupling Analysis (DCA) per la fisica statistica inversa, Zeng & Aurell (2020) hanno provato in silico che in QLE è possibile inferire l’epistasi a partire dalla distribuzione (dinamica) dei genomi. In questo lavoro, tre diverse linee di ricerca sono seguite. In primo luogo, sviluppiamo e testiamo una nuova teoria (Gaussian Closure) che permette di inferire l’epistasi in un dominio più ampio rispetto al regime QLE, ovvero quando il tasso di mutazioni $\mu$ è comparabile o superiore ad $r$. In secondo luogo, studiamo il comportamento di una popolazione in un’evoluzione dove la forza della selezione è sufficientemente alta rispetto a $r, \mu$, una fase inesplorata che chiamiamo NRC-phase. Infine, usiamo la DCA e un’assunzione di QLE per classificare i contributi epistatici alla fitness dei loci polimorfici del Sars-Cov-2 coronavirus.

**ABSTRACT**

We here consider the NS theory, a statistical model for population genetics studied by Neher & Shraiman (2011) that describes a population evolving due to the mutations, recombination, genetic drift, natural selection (pairwise epistatic fitness). In particular, we consider the Quasi-Linkage Equilibrium (QLE), which is found for a sufficiently high recombination rate $r$ with respect to the selection strength $\sigma$. By combining the results of the NS theory with the techniques of the Direct Coupling Analysis (DCA) for the Inverse Statistical Physics, Zeng & Aurell (2020) have proved in silico that in QLE it is possible to infer the epistasis from the knowledge of the (dynamical) distribution of the genotypes. Here, three different lines of research are pursued. Firstly, we develop and test a new theory (Gaussian Closure) that allows inference of epistasis in a broader domain than the QLE regime i.e. when the mutation rate $\mu$ is comparable or larger than $r$. Second, we study the behaviour of an evolving population for sufficiently high selection strength with respect to $r, \mu$, an unexplored scenario that we name NRC-phase. Lastly, we use DCA together with an assumption of QLE to epistatic contributions to fitness from polymorphic loci in the Sars-Cov-2 coronavirus.
## CONTENTS

On Models, Typical Behaviours and Spherical Cows xi  
Modelling Nature ............................................ xi  
The Unreasonable Effectiveness of Statistical Mechanics ................. xiii  
Biological Complexity for Physicists ................................... xiv  
Structure and Contents of this Thesis .................................. xvi

### Background

1 Introduction to Statistical Genetics  
1.1 Subject Matter: Population Genetics in a Nutshell ................. 3  
1.2 Model: NS Theory ........................................ 5  
  1.2.1 NS Hypotheses ........................................ 6  
  1.2.2 Fitness .................................................. 7  
  1.2.3 Mutations ............................................... 8  
  1.2.4 Recombination ......................................... 9  
  1.2.5 Dynamics of Genotype Distribution ...................... 10  
1.3 Quasi-Linkage Equilibrium: KNS Theory ........................... 15  
  1.3.1 The Central Result of KNS theory ....................... 17  
1.4 Random Genetic Drift ....................................... 19  
1.5 KNS Theory for Categorical Data ................................ 21  
1.6 Contours .................................................. 23  
  1.6.1 Biological Realism. .................................... 24  
  1.6.2 Beyond the QLE phase. ................................ 25  
  1.6.3 Different Approaches to the Evolutionary Problem .......... 26

2 Simulating Evolution: FFPopSim ................................ 29  
2.1 Generalities ............................................. 29  
  2.1.1 Initialization .......................................... 30  
  2.1.2 Evolution ............................................. 31  
2.2 Observables ............................................... 32  
2.3 The Elephant in the Room ................................... 35
# Contents

3 Inverse Statistical Physics: DCA

3.1 An Inverse Paradigm .................................................. 37
3.1.1 Statement of the IIP .............................................. 37
3.1.2 Thermodynamics of the IIP ...................................... 39
3.2 Maximum Likelihood .................................................. 41
3.3 Approximate solutions for the IIP ................................. 43
3.3.1 Mean Field Approximation ..................................... 43
3.3.2 Pseudolikelihood Maximization ............................... 44
3.3.3 MF vs PLM ......................................................... 46
3.4 Inference in the NS theory .......................................... 47
3.5 Datacentrism ............................................................. 50

Advances ........................................................................ 53

4 A Gaussian Closure for the NS Theory ......................... 55
4.1 A Gaussian Ansatz ..................................................... 55
4.1.1 The Logic Of The Gaussian Ansatz ......................... 55
4.1.2 Closed Equations for Cumulants Dynamics .............. 57
4.1.3 New Inference Formula for Epistasis .................... 59
4.2 Gaussian Closure Under Examination .......................... 61
4.2.1 Experimental Layout ............................................. 61
4.2.2 Results & Comments ............................................ 64
4.3 Directions ................................................................. 66

5 QLE Rupture: a New Phase ............................................. 67
5.1 Hallmarks of the NRC-phase ...................................... 67
5.1.1 Localization in the Parameter Space ....................... 67
5.1.2 7 Observations ..................................................... 68
5.1.3 What If(s) .......................................................... 74
5.2 Understanding the NRC-phase: Insights ...................... 76
5.2.1 QLE → NRC: Intuition ......................................... 76
5.2.2 NRC → QLE: Muller’s Ratchet ............................ 78
5.2.3 Escape Times ..................................................... 79
5.2.4 Speculation: A NRC-phase for E.Coli? ................. 82
5.3 Spectrum ................................................................. 84

6 Detecting Epistasis in SARS-CoV-2 ............................... 85
6.1 Data journey from bio labs to DCA scores ................. 86
6.1.1 MSA Refinement .................................................. 86
6.1.2 PLM Inference ..................................................... 87
CONTENTS

6.1.3 Randomization Sieve ........................................ 90
6.2 Results: Spotlight on Strong Couplings .................... 92

Appendices .............................................................. 95
A Moments vs Cumulants ........................................... 97
  A.1 Example: Gaussian case ....................................... 98
B The Fisher-Wright Model ......................................... 100
C Code Bones .......................................................... 103
D Maximum Entropy and Information ........................... 104
Glossary ................................................................. 107
Bibliography ........................................................... 110
Acknowledgements .................................................... 115
ON MODELS, TYPICAL BEHAVIOURS AND SPHERICAL COWS

What is science about? What, exactly, is a model? When and why does probability come into play? What are the fundamental mechanisms underlying statistical mechanics? Where does the interest in it beyond physics stem from? In particular, on what grounds we as physicist are allowed to invade the realms of biology?

There is material for more than a single book in these questions, indeed there are many out there. Here in this introductory chapter, we will not claim or try to give exhaustive answers, which would be beyond the scope of this work and of the writer’s competences. Yet we believe it is necessary to dedicate a few pages to grasp their most fundamental essence, just like at the beginning of each journey it is always worth wondering where we are, what we are doing, where we are going.

In this introduction we progressively narrow the focus of our questions down to the specific topic of this Thesis, a preview of which is also provided below.

MODELLING NATURE

An exceedingly popular aphorism among the statisticians is - Essentially, all models are wrong, but some are useful - by George E.P. Box (1919-2013) [Box, 1976]. What does it mean?

Let us start by pointing out the milestones of the process that, from a question about a natural phenomenon, leads to a quantitative (numerical) prediction; science, one may say, is what lies between these two extrema.¹

1. Nature. In the beginning, there is a question about a natural phenomenon i.e. something that is observed to happen or exist, typically characterized by an enormous number of details. Quantitative measurements are possible and/or experiments can be carried out.

2. Description. The problem is expressed in a language, with a text of finite length. We cannot take into account all the details and possible specifications of the phenomenon, instead we focus on a few of them, the relevant ones. The description would not be strictly correct, but hopefully capture the most important features of the problem.

3. Model. The implication is: mathematical, it always is a mathematical model. The translation of the description above in this formal language entails introducing the suitable variables and parameters, making the appropriate assumptions and approximations.

¹ Much of the material of this first chapter, draws from the inspiring lectures of the Information Theory course, held by prof. M. Marsili in Trieste, ICTP, academic year 2019/2020. Unfortunately, the lecture notes are available nowhere on the internet (to our knowledge).
4. Calculation. Reaching and answer involves a logical/mathematical calculation. The result of this process may or may not be in agreement with experiments and/or simulations: in the first case, we have reached some understanding of the phenomenon.

The fact that these steps are so well-established in the everyday scientific life must not mislead: a closer inspection tells us that the logic underlying them is by no means trivial. To illustrate this point, we refer to a marvellous piece of literature - *The Unreasonable Effectiveness of Mathematics in the Natural Sciences* - by E.P. Wigner [Wigner, 1960] and turn the spotlight on the following questions:

i. *Why should irrelevant details exist?* The very possibility of science relies on the fact that, in spite of the staggering complexity of each natural process, we can study the regularities that are observed for a set of few relevant variables and forget about the overwhelming number of unimportant details. A body falling from an height $h$ takes a time $t = \sqrt{\frac{2h}{g}}$ to reach the ground ($g = 9.81 \, m/s^2$), as was first discovered by G. Galilei (1564-1642) in Pisa. This law is true not only in Pisa, but everywhere; not only in the XVII century, but at any time; it is valid whatever the colour of the falling body, whoever the experimenter letting the body fall (...) There is no general principle by means of which these forms of invariance can be understood and the fact that laws of nature can be expressed in terms of few variables independent from all the rest is not at all obvious: it is, quoting E.P. Wigner, surprising.

ii. *Why does mathematics work?* Even more striking is the effectiveness of mathematics as a language that allows to formalize questions about nature and that provides a framework where to work out answers. It is true that many of the concepts of mathematics mimic nature (think of the elementary mathematics or geometry); there are also several examples of how mathematical theories have been later recognized to be useful in describing natural phenomena. However, it is difficult to argue that concepts like complex numbers or Hilbert spaces were inspired by nature, despite they have turned out to be essential for its formal description. In other words, mathematics does not need nature. So why do we use it? There is an uncanny number of cases in which answers based on mathematical calculations have turned out to be extremely accurate, which suggest that mathematics is not just a language, it is the correct language and there is no fundamental reason for this to be so. In the words of E.P. Wigner again - *The miracle of the appropriateness of the language of mathematics for the formulation of the laws of physics is a wonderful gift which we neither understand nor deserve.*

In the effort of being as general as possible in this section, we have avoided specifying the discussion to physics or any other branch of the natural sciences, we will deal below with (some of) them. Indeed, we did not need it as it is well-known that - *the unity of all science consists alone in its method not in its material* - (K. Pearson, The Grammar of Science, 1892).
The unreasonable effectiveness of statistical mechanics

From the previous section, we deduce that the laws of nature are *conditional* statements, since they are based on a small (relevant) subset of all possible information on a natural process. The notion of (in)dependence has nothing to do with physics; it is a probabilistic statement. However, the role of probability and statistics is much more fundamental than simply providing a quantitative relation between important and unimportant details.

The existence of laws of simple nature is remarkable for one more reason i.e. that in most phenomena, we perfectly know that there is a terrific number of interacting degrees of freedom at stake, examples being the paramagnetism, the Brownian motion, the diffusion process (...). All these systems are ensembles of a huge number of indistinguishable interacting entities from which collective phenomena emerge that do not depend on microscopic details. The ultimate reason why this happens is to be found in the weak Law of Large Numbers or, better, in its information theory analog, the Asymptotic Equipartition Property [Cover & Thomas, 2006]: we cannot fail to state it here, if only to show the simplicity and elegance of one of the most fundamental results of the scientific knowledge:

**Theorem 1. Asymptotic Equipartition Property (AEP):** let \(X_1, X_2, \ldots, X_n\) be discrete random variables with alphabet \(\mathcal{X}\), independent and identically distributed (i.i.d.) with \(X_i \sim p(x) \forall i\). Then

\[
\lim_{n \to \infty} \Pr\left[\left| -\frac{1}{n} \log p(X_1, X_2, \ldots, X_n) - H(X) \right| > \epsilon \right] = 0 \quad \forall \epsilon > 0
\]

where \(H(x) = -\sum_{x \in \mathcal{X}} p(x) \log p(x)\) is the Shannon entropy of the distribution \(p(x)\).

The idea is that, if we define the *typical set* \(A^{(n)}_{\epsilon}\) with respect to \(p(x)\) as the set of sequences \((x_1, \ldots, x_n) \in \mathcal{X}^n\) for which \(2^{-n(H(X) + \epsilon)} < p(x_1, \ldots, x_n) < 2^{-n(H(X) - \epsilon)}\), any property that is proved for the typical sequences will hold true with high probability and determine the collective behaviour of a large sample/system. The existence of typical behaviours is exactly what allows the emergence of *regularities* that usually depend on few variables and we call laws of nature.

Let us now consider the branch of physics concerned with the macroscopic behaviours of matter (*phases*) as they result from the properties of the interactions of its constituents i.e. statistical mechanics. The emergence of collective behaviours of large ensembles of interacting particles, independent from microscopic details, has little to do with physics and it is primarily due to the statistical nature of the underlying laws. This is, ultimately, the reason of the well-known *universality* of critical behaviours between phases of matter (*phase transition*) [Goldenfeld, 1992] and, in a broader perspective, it lays the foundations for translating these results much beyond

---

2 The results of experiments always come with *experimental errors*, which are basically due to our lack to ability in taking into account possible effects that might play a role in the result. In other words, in experimental errors we confine what we deem *irrelevant*. There is no surprise in the fact that the quantitative theory of errors is entirely based on statistics and probability theory.
the frontiers of the physics.

The connection between information theory and statistical mechanics as a microscopic theory of thermodynamics is particularly interesting in this regard, and it is exemplified by the concept of entropy, which is at the heart of both the disciplines. In statistical physics, it was defined by L. E. Boltzmann (1844-1906) for systems at thermodynamic equilibrium as $S = k_B \log \Omega$, where $k_B$ is a constant and $\Omega$ is the number of configurations of a system consistent with macroscopic constraints (e.g. volume, pressure). In light of the seminal work of C. E. Shannon (1916-2001) [Shannon, 1948] and of the development of the information theory, it seems more natural now to see the thermodynamic entropy as a special case of the aforementioned Shannon entropy, a measure of information or, more precisely, of the uncertainty of a random variable.

Very-roughly speaking, every time the scientist is interested in the collective behaviour at equilibrium of a large number of interacting entities, a "statistical mechanics" may be possible.

A review of all the fields of science in which ideas from statistical mechanics (and information theory) have been applied is out of reach here. Paradigmatic models like the Ising/Potts/Heisenberg model, spin-glass models (...); concepts like phase transitions, critical exponents, renormalization group (...); techniques like the high-temperature expansion, mean-field approximation (...); all these notions pop up everywhere in the scientific literature outside the domain of physics: examples come from statistics and inference [Zdeborová & Krzakala, 2016], network science [Albert & Barabási, 2002], large deviation theory [Touchette, 2009], economics [Yakovenko & Rosser, 2009], ecology [Azaele et al., 2016], sociology [Castellano et al., 2009] (...)

Many more can be found e.g. in [Sornette, 2006], one more will be the subject of this Thesis; for all the countless ones that we ignore and missed in this list, we can but apologize.

**BIOLOGICAL COMPLEXITY FOR PHYSICISTS**

Physics, one may say, is the art of approximation. As the joke goes, for a theoretical physicist a cow is a sphere\(^3\): indeed, if we are interested in the gravitational motion of a cow close to a black hole, this may be accurate, but have you ever tried to roll a cow down a hill? (Please, do not do that!)

Theoretical physicists are accustomed to make approximations; once again, this relies on the fact that irrelevant details exist in the description of phenomena. In many cases of interest in physics (not always!), details can be arranged in a hierarchy of relevance: the amount of them we will take into account depends on the level of accuracy we want to reach.

Above we have discussed the journey from a question about nature to a scientific answer, it is worth emphasizing that science (and physics in particular) is about questions at least as much as about answers: asking how a financial market behaves at the outbreak of an economic crisis

\(^3\) As reported in [https://en.wikipedia.org/wiki/Spherical_cow], here comes the story: Milk production at a dairy farm was low, so the farmer wrote to the local university, asking for help from academia. A multidisciplinary team of professors was assembled, headed by a theoretical physicist, and two weeks of intensive on-site investigation took place. (...) Shortly thereafter the physicist returned to the farm, saying to the farmer, “I have the solution, but it works only in the case of spherical cows in a vacuum.
has no such sharp answer as asking the magnitude of the force between two charged particle: neither the first nor the second are trivial questions, but we expect the answer to the latter to depend on few variables (as we know, magnitude of charges, distance between them) while we have no reason to believe that the same is true for the former.

The keyword is complexity: the more a system is complex, the more we should be careful with our assumptions. Statistical behaviours of large ensembles of identical interacting entities are undoubtedly helpful in pruning the number of variables, but the system may still too complex to be accurately described by our simple models from statistical physics. This discussion is particularly appropriate when we consider biology, life sciences and related disciplines: even the smallest unit of life, the cell, is of baffling complexity and distinguishing what is relevant from what is not in an intracellular process may seem a daunting task. Most of this Thesis, we anticipate, will address the problem of modelling the evolutionary process i.e. the evolution of a population of individuals under the effects of some mechanisms that we deem relevant (mutations, recombination, natural selection, genetic drift); but question whether or not evolution should be predictable has no easy answer [Lässig et al., 2017]. There is no doubt that some of the assumptions that physicists make when dealing with biology may be disconcerting: can we model a gene in a genome or a neuron in a brain as an Ising spin? Is an amino-acid in a polypeptide chain reasonably described by a Potts spin? Do these assumptions lead to accurate results?

In some cases, they do, or at least this is what data seem to suggest. As an example, a simple Potts model with pairwise interactions seems to be sufficient to design artificial proteins [Cocco et al., 2018; Socolich et al., 2005]. Indeed, there is one last point that we have not yet wrote about: big data (as it is fashionable to say nowadays). The deluge of data in recent years has flooded many branches of the scientific research, biology has certainly benefited from this scientific shock: thanks to high-throughput experiments we are now able to gain insights on processes of tremendous complexity; a remarkable example is, for instance, the study of the human brain from the physics perspective [Lynn & Bassett, 2019]. Not only we can validate our models on data, a growing enthusiasm is in the possibility learning from data. In other words, what seemed to be a hopeless task - describing and predicting complex systems with scientific rigour and method - is more and more realistic, thanks to a valuable ally: data are informed by models, models are informed by data.

As last words of the brief journey to the conceptual foundations of this Thesis, we borrow those of a great theoretical physicist, E. Schrödinger (1887-1961), who already in the 1944, puzzled with the dilemma of What is life? wrote:

(...) ‘naïve physicist’s ideas about organisms’ [are the] the ideas which might arise in the mind of a physicist who, after having learnt his physics and, more especially, the statistical foundation of his science, begins to think about organisms and about the way they behave

---

4 Some have gone as far as to claim that - Petabytes allow us to say: “Correlation is enough.” We can stop looking for models - (from Anderson, C.: The End of Theory: The Data Deluge Makes the Scientific Method Obsolete, Wired (2008) http://www.wired.com/2008/06/pb-theory/). This is clearly a fancy slogan, but deceitful in its deepest aspects; as pointed out by [Hosni & Vulpiani, 2018], we cannot “throw away equations”, the relationship between data science and the art of modelling has to be carefully considered beyond any enthusiastic pulse, not to raise false expectations.
and function and who comes to ask himself conscientiously whether he, from what he has learnt, from the point of view of his comparatively simple and clear and humble science, can make any relevant contributions to the question. It will turn out that he can. [Schrödinger, 1944]

STRUCTURE AND CONTENTS OF THIS THESIS

This work is organized in two main blocks.

I. The first part, BACKGROUND Ch.(1-3), illustrates the theoretical and computational tools that we will employ for our researches. In particular:

- Ch.(1) deals the Neher - Shraiman (NS) theory of population genetics i.e. the statistical description of a population of individuals (genomes) as it evolves under the effect of recombination, mutations, natural selection. In particular, the regime of Quasi-Linkage Equilibrium (QLE) of weak selection and strong recombination, where a way appears to infer from data the fitness landscape of an evolving population.
- Ch.(2) introduces the simulation tool FFPopSim, based on the NS theory, and shows how it can be used to simulate and visualize the evolutionary process in a QLE regime.
- Ch.(3) states and discusses the Inverse Statistical Physics, in particular the Inverse Ising Problem (IIP) i.e. the problem of finding the unknown parameters of the Ising Hamiltonian starting from a set of samples from the same distribution. Some of the techniques developed are applied to the aforementioned inference of the fitness landscape.

Each chapter in this first part is opened by a short introduction and preview of its content. The last section of each of them aims instead at giving some broader context and points out some methodological and conceptual aspects the discussion.

II. The second part, ADVANCES Ch.(4-6), illustrates the projects that we have worked for the Thesis work. In particular:

- Ch.(4) aims at improving the QLE description as it appears in Ch.(1) and providing a theory that works in a broader range of parameters. The approach is based on a Gaussian Ansatz of the probability distribution of the genotypes in the population. The results based on the NS theory and those based on the Gaussian Ansatz are compared following our recent contribution [Zeng, Mauri, et al., 2020], currently under peer review.
○ Ch.(5) focuses on what seems to be a new, previously undetected phase of the evolutionary process, that emerges for sufficiently high selection with respect to mutations and recombination, we name it Non-Random Coexistence (NRC) phase. After studying and characterizing its behaviour as it appears in simulations using FFPopSim Ch.(2), we go through the literature hunting for useful hints for its theoretical understanding that still is missing.

○ Ch.(6), finally, is a summary of the content of [Zeng, Dichio, et al., 2020], also under peer review. At the outbreak of the Covid-19 pandemic, labs all around the world have started sequencing the Sars-Cov-2, the virus responsible for the disease. This information have been collected in open-access repositories, from which they can be downloaded and analyzed. This is precisely what we have done, by exploiting the theoretical/computational tools of Ch.(3), in order to determine the most important biological couplings between different loci of the viral genome.

Each of these three last chapter has, more or less, the structure of a paper: after an introduction, the main body follows with derivations and results; the last section always contains a critical discussion of the results and a preview of future directions of investigation.

In addition, four appendices have been added: App.(A) defines moments and cumulants and points out the relationship between them; App.(B) discusses the Fisher-Wright model of genetic drift; App.(C) briefly illustrate the structure of the code employed in this work; App.(D), finally, is dedicated to the Maximum Entropy principle and to the definition basic concepts from information theory.

Finally, the reader not yet accustomed to biology will find in the last pages a glossary where all the biological concepts that appear in this work are concisely defined.
BACKGROUND
1 INTRODUCTION TO STATISTICAL GENETICS

The first pillar of this work is statistical genetics: in the broad area of theoretical biology, it is concerned with the development of statistical methods for describing, among others, population genetics. The chapter will open with a description of the subject matter in Sec.(1.1), followed by an introduction to the Neher-Shraiman theory of statistical genetics in Sec.(1.2). In Sec.(1.3) we describe the regime of Quasi-Linkage Equilibrium and the Kimura-Neher-Shraiman theory, some extensions are described in Sec.(1.4 - 1.5). Finally in Sec.(1.6) an overall discussion will be found.

1.1 SUBJECT MATTER: POPULATION GENETICS IN A NUTSHELL

This section is dedicated to a brief introduction to all the biology we need to know and nothing more. It will be summerized in a nutshell, we will mention several technical terms, the reader not yet familiar with them will find corresponding entries in the glossary at the end of this Thesis. As general references for the biological side we mention [Klug et al., 2012].

*It is interesting to contemplate an entangled bank, clothed with many plants of many kinds, with birds singing on the bushes, with various insects flitting about, and with worms crawling through the damp earth, and to reflect that these elaborately constructed forms, so different from each other, and dependent on each other in so complex a manner, have all been produced by laws acting around us ... Thus, from the war of nature, from famine and death, the most exalted object which we are capable of conceiving, namely, the production of the higher animals, directly follows. There is grandeur in this view of life, with its several powers, having been originally breathed into a few forms or into one; and that, whilst this planet has gone cycling on according to the fixed law of gravity, from so simple a beginning endless forms most beautiful and most wonderful have been, and are being, evolved. [Darwin, 1859]*

The key element of evolution is heredity: since the dawn of biological life, information has flowed ceaselessly from parents to children, in an ongoing process that has shaped the astonishing variety of complex organisms that inhabit this planet. Several questions arise: how and where is information stored? how is information transmitted? what is the origin of the observed diversity of life?

Life on the planet falls under the into three domains: Eukaryotes, whose cells are endowed with a defined nucleus and organelles; Prokaryotes, cells with no nucleus; Archaea, prokaryotic organisms that have distinct molecular characteristics that differentiate them from the
previous two domains. According to the central dogma of molecular biology (with few exceptions), all the biological information is encoded in the DNA, a macro-molecule present in each cell that consists in two sugar-phosphate ribbon-like stands that coil around to form a double helix and whose horizontal rungs are pairs of complementary nucleobases: A-T, G-C, see Fig. 1.1.

The information lies in the exact sequence of nucleobases of each strand: by means of the transcription - translation processes, a fragment of such sequence (gene) is translated into a polypeptide chain, whose folding will eventually result in a protein. Such protein in turn will perform one of the uncountable functions needed to sustain life. Complex regulatory mechanisms of the gene expression weave an intricate and largely unknown network of interactions within genes. In eukaryotes the DNA is often condensed in the form of chromosomes: the set of chromosomes is the genetic heritage of an individual. A population in which each cell has one complete set of chromosomes is named haploid, if there are two such sets, diploid.

Let us now broaden the perspective and consider an entire population of individuals. In general, each gene is present in a population with different alleles, in different amounts. The goal of population genetics is to study the genetic composition and dynamics of biological populations. Indeed several processes can alter the genetic pool of a population, the most important of them being listed below.

- **Natural selection.** At the phenotype level, advantageous characteristics will enhance the probability for an individual to survive and reproduce (high fitness). This has consequences at the genotype level, even though the exact map between these two layers is overshadowed by biological complexity and mostly unknown.

- **Mutations.** They can arise by chance in a genomic sequence e.g. because of transcription errors. Mutations can have no consequences on the level of protein (synonymous) or cause alterations in the polypeptide chain they code for (non-synonymous). They represent a source of variability for the evolutionary process.
Recombination. Several mechanisms of interaction between individuals of a population can lead to the exchange of genetic material. In prokaryotes, transduction, transformation, conjugation. In eukaryotes, recombination happens during meiosis where the mixing between two chromosomes from each of the parents is enhanced by the crossing-over mechanism. Just like mutations, recombination fuels variability in the population.

Chance. Accidental events can dramatically alter the genetic pool of a population: population bottlenecks, genetic drift, hybridization...

Differently from population genetics, quantitative genetics deals with the genetics of continuously varying characters. While the former looks at the genotype scale (microstate) the latter focuses on statistics at the phenotype scale (macrostate).

1.2 Model: NS Theory

Now that we understand the subject matter, our first job is to set up a minimal theoretical framework to model it. Our task is indeed twofold: not only we have to describe the problem in a language that we understand, that we can handle and from which we can (hopefully) extract useful predictions, but also our description has to be minimal i.e. we would like to use the minimum possible number of variables that capture the whole of the relevant features of the phenomena we are modelling.

A number of authors have addressed the problem of the quantitative description of population genetics, following slightly different approaches. One of the very first results dating back to 1930 is the R.A. Fisher’s celebrated fundamental theorem of natural selection [Fisher, 1930] according to which in absence of mutations and in the limit of an infinite population the average fitness of the population cannot decrease in time, and becomes stationary only when all of the individuals in the population bear an optimal genome, corresponding to the maximum value of the fitness.¹ This deterministic picture of evolution, albeit simple, holds in some very specific cases but it is now regarded more as an exception than as the rule.

More recently, a considerable amount of literature has exploited methods from both equilibrium and non-equilibrium statistical mechanics to describe population genetics. The ensemble of these methods has later gained life of its own and it is collectively referred as statistical genetics (SG). We mention [Blythe & McKane, 2007; Peliti, 1997] for modern reviews on this topic and the seminal work of R.A. Neher and B.I. Shraiman [R. A. Neher & Shraiman, 2011], the main reference for this chapter: in what follows, we will indeed describe the Neher - Shraiman approach to statistical genetics (NS theory).

As a very first question, one should wonder why statistical mechanics should work at all in describing the evolution of a population of genomes. An illuminating observation is that thermodynamics is conceptually very close to quantitative genetics.

Let us try to draw some analogies between them:

¹ See [Peliti, 1997] for a concise proof.
**Thermodynamics** is a phenomenological description of observable physical properties of a large ensemble of particles. This is achieved by taking macroscopic averages over the random motion of individual particles at thermal equilibrium. Despite the complex and chaotic single-particle motion, deterministic laws of thermodynamics emerge thanks to the statistical properties of the ensemble of particles considered.

**Quantitative genetics** is a phenomenological description of observable phenotypic traits of a population of individuals (genomes). This is achieved by taking averages over the whole population, whose dynamics is governed by several drivers defined at the genotype level. Despite the largely unknown genotype-phenotype map and chaotic environmental effects, laws of quantitative genetics emerge thanks to the statistical properties of the ensemble of genomes considered.

The role of statistical mechanics is to provide a "bridge" between the physics at the microscopic scale and the thermodynamic quantities, the latter expressed as ensemble averaged functions of the coordinates and momenta of the constituent particles. We seek likewise a coarse-grained phenomenological theory bridging the gap between the genotype dynamics and the population averaged observables of quantitative genetics, hence the name **statistical genetics**.

Convinced of the possibility of a statistical theory of genome evolution, we shall now describe a first model that implements this key-idea and provides a fertile ground for implanting our theoretical building. First of all, we will set the stage i.e. point out the hypotheses of the model and clarify its boundary conditions, therefore providing an environment where our abstract genomes will live and act. Secondly, we shall describe the mechanisms that drive these entities’ behaviours and encode them in a master equation. Exploiting the latter, we will (hopefully) derive useful relations whose validity can be assessed on real data.

### 1.2.1 NS Hypotheses

We here set the stage of the NS theory, in the form of a list of hypotheses.

I. **Genomic structure.** Consider an haploid genome \( g = (s_1, \ldots, s_L) \) of \( L \) loci \( s_i \) where \( i = 1, \ldots, L \). The number \( L \) of loci is fixed and equal for all the individual genomes. A population is a collection \( \{ s_\alpha \} \alpha \in A \), where \( A \) is a set of indices. Each genome \( g \) appears in the population with probability \( P(g) \).

II. **Ising genes.** Suppose biallelic loci i.e. \( s_i = \pm 1 \ \forall i \); the genome space is represented by the \( 2^L \) vertices of the hypercube \( \{-1, 1\}^L \).

III. **Constant population.** The total number of individuals is fixed \( |A| = \text{cost} \). With this hypothesis we model e.g. the struggle for survival in an environment with limited re-

---

2 The identification of genes as spins in a lattice, heralded by this name, will become more and more clear as we build up the theory. At this stage, it is only a formal similarity.
sources. Until further notice, we make the further assumption of an infinite population 
\(|A| = \infty\), that allows us in a first approach to neglect stochastic effects.

IV. Evolution. The genome distribution \(P(g,t)\) evolves in time. The dynamics is driven by
the three operators representing natural selection Sec.(1.2.2), mutations Sec.(1.2.3) and
recombination Sec.(1.2.4). Their action is encoded in a master equation i.e. a phenomeno-
logical first-order differential equation

\[
\frac{d}{dt} P(g,t) = \frac{d}{dt} \mid_{fit} P(g,t) + \frac{d}{dt} \mid_{mut} P(g,t) + \frac{d}{dt} \mid_{rec} P(g,t) .
\] (1.1)

In order to provide an explicit expression for Eq.(1.1), we now turn to analyse one by one the
terms in the rhs.

1.2.2 Fitness

The approach we follow for a mathematical description of natural selection is based on the
definition of the fitness function \(F(g)\), that is proportional to the average number of offspring of an
individual whose genotype is \(g\). In other words, \(F(g)\) expresses the propensity of a genotype to
transfer its genomic material to the next generations. The explicit form of \(F(g)\) is said to define
the fitness landscape of the population.

Even if very broad, this definition is an ansatz:

\* It implicitly makes the assumption that the fitness function depends only on \(g\) and this
in not true in general: the reproductive rate of a given genome (or genomic trait) may
depend in principle on its frequency in the population e.g. because of some feedback
regulation system etc.

\* It gives up all hopes to describe biological mechanisms as cyclical dominance, where "\(A
> B\)", "\(B > C\)" but "\(C > A\)".

\* It also sets aside issues related to a possible fluctuating environment (fitness seas-
capes) and the related time-dependence of selection.

The fitness term in Eq.(1.1) can be written formally as

\[
\frac{d}{dt} \mid_{fit} P(g,t) = [F(g) - \langle F \rangle]P(g,t) ;
\] (1.2)

\(\langle F \rangle(t) = \sum_g F(g)P(g,t)\) is the average fitness that ensures the normalisation of \(P(g,t)\). The fit
individuals (high fitness) will grow, the unfit ones will decrease.
We need to provide an explicit expression for $F(g)$; previous researchers have explored different possibilities\(^3\), in this work we choose a fitness function with pairwise interactions:

$$F(g) = \bar{F} + \sum_i f_i s_i + \sum_{i<j} f_{ij} s_i s_j . \quad (1.3)$$

$F$ is a constant, irrelevant for Eq.(1.2). The first order contribution $f_i$ represent the additive fitness at locus $i$, independent of all other loci in the genome while higher order terms $f_{ij}$ (and $f_{ijk}, f_{ijkl},...$ if they were present) represent the genetic interactions between loci, also called epistasis. For future needs, we characterize the total fitness by

$$\sigma(f) = \sqrt{\sum_i f_i^2 + \sum_{i<j} f_{ij}^2}, \quad (1.4)$$

the epistatic fitness by $\sigma_e$ which has the same definition as Eq.(1.4) except that only the epistatic contribution appears and the additive fitness by $\sigma_a$ analogously.

As it is clear from Eq.(1.2), $F(g)$ has dimension $[t^{-1}]$, by consequence the same is true for all the coefficients $f_i, f_{ij},...$ and for $\sigma, \sigma_e, \sigma_a$, too.

As a final remark, it is worth noting that fitness in statistical genetics plays the same role of energy (modulo a minus sign) in statistical mechanics. The terminology of fitness landscape underlines the analogy with an energy landscape: $F(g)$ would be an Hamiltonian $-\mathcal{H}$ in statistical physics. We can further exploit this analogy to depict the evolutionary process of a population as an erratic motion of a point on the fitness landscape: contrary to what happens in an energy landscape, where the systems slides toward the valleys (min. energy), we will see here our point particle climbing the hill as much as possible (read highest number of offspring).

### 1.2.3 Mutations

The definition of mutation includes a multitude of different biological mechanism whose effect is to modify the genome sequence. For the modelling purpose, we choose a very simple mechanism: the single-locus swap $s_i \rightarrow -s_i$.

In a mathematical language, it is sufficient to introduce an operator $M_i$ whose action on a genomic sequence is to swap the $i$-th biallelic gene i.e.

$$M_i(s_1, ..., s_i, ..., s_L) = (s_1, ..., -s_i, ..., s_L).$$

Let also $\mu$ be the tunable mutation parameter we need to weight the mutation mechanism in our model. We assume it to be constant in time and the same for all loci; analogously to $\sigma$ it has

\(^3\) See for example [Peliti, 1997] for a detailed discussion of the cases $F(g) = \sum s_i$ (Fujiyama landscape) or $F(g) = \epsilon \delta_{g,0}$ (sharp-peak landscape).
dimensions \([t^{-1}]\), it is a rate. The mutation term in the master equation will simply take the form

\[
\frac{d}{dt} P(g, t) \bigg|_{\text{mut}} = \mu \sum_{i=1}^{L} [P(M_i, g, t) - P(g, t)].
\]  
(1.5)

1.2.4 Recombination

Through mating and recombination, two parents \(g^{(1)}, g^{(2)}\) mix their genomic sequences giving birth to two new individuals \(g, g'\). The mechanism we here have in mind is the crossing over of homologous gametes during meiosis where haploid individuals produce an haploid offspring.

Albeit relatively simple, this is sufficient to describe some forms of bacterial recombination (transformation, transduction where material goes in both ways) as well as recombination in several RNA viruses including HIV, influenza. In contrast, it cannot model some other biological mechanisms of genes-mixing, for instance bacterial conjugation.\(^4\)

The NS description of recombination makes use of a set of random variables \(\{\xi_i\}\) that define the crossover pattern as follows: consider the gene at locus \(i\) of the new individual \(g\), if it has been inherited from \(g^{(1)}\) then \(\xi_i = 1\) while if it comes from \(g^{(2)}\) then \(\xi_i = 0\). The sequence \(g'\) is simply complementary to \(g\). In symbols \(g, g'\) can be written as

\[
g : s_i = \xi_i s_{i}^{(1)} + (1 - \xi_i) s_{i}^{(2)},
\]

\[
g' : s'_i = (1 - \xi_i)s_{i}^{(1)} + \xi_i s_{i}^{(2)}.
\]  
(1.6)

Each different crossover pattern \(\{\xi_i\}\) comes with a probability \(C(\xi)\). Similarly to the mutation case, let \(r\) be the tunable overall recombination parameter (dimensions \([t^{-1}]\)) which in this case can be accompanied by a relative parents-dependent rate \(Q(g^{(1)}, g^{(2)})\) (dimensionless). Collecting everything we have so far, it is possible to express the time variation of the genomic sequence due to recombination as

\[
\frac{d}{dt} P(g, t) \bigg|_{\text{rec}} = r \sum_{\xi \in \xi} C(\xi) \left[ Q(g^{(1)}, g^{(2)}) P_2(g^{(1)}, g^{(2)}, t) - Q(g, g') P_2(g, g', t) \right],
\]  
(1.7)

where the sum runs over all possible recombination patterns and all possible sequences \(g'\) and \(P_2\) is the two-genome distribution (read two-particle distribution). Two more steps:

* Until further notice, we set \(Q(g_{\alpha}, g_{\beta}) = 1 \forall g_{\alpha}, g_{\beta}\) which means that any genome pair has the same recombination rate \(r\); this entails the assumption of a panmictic population, where any individual is equally likely to interact with anyone else.

\(^4\) In [R. A. Neher & Shraiman, 2011], Neher-Shraiman understand recombination in a slightly different way: two parents each produce a mating body (copy of genome) and these two merge and give birth to a new individual while the remaining half of genetic material is simply wasted. Regardless of the interpretations, the resulting evolution turns out to be the same.
It is hard to handle eq.(1.7) without a closure, so we will also assume that

\[ P_2(g_\alpha, g_\beta) = P(g_\alpha)P(g_\beta), \]  

(1.8)

which is valid if genomic sequences undergoing recombination are uncorrelated. This assumption is crucial and we will come back to it. The reason of such interest in Eq.(1.8) is that it is never exactly true.\(^5\) In a realistic biological environment, several phenomena can introduce correlations between different individuals e.g. competition for limited resources, geographical separation, existence of classes of individuals, or phylogenetic effects... If not exactly true, we will assume that such correlations are weak enough for Eq.(1.8) to be approximately so, i.e. \( P_2(g_\alpha, g_\beta) \sim P(g_\alpha)P(g_\beta) \).

Implementing in Eq.(1.7) we get

\[ \frac{d}{dt} P(g, t) = E \sum_{g' \in G} C(\xi) \left[ P(g^{(1)}, t) P(g^{(2)}, t) - P(g, t) P(g', t) \right]. \]  

(1.9)

### 1.2.5 Dynamics of Genotype Distribution

The aim now is to parameterize the distribution \( P(g, t) \) \( \forall g \) by its cumulants\(^6\). The cumulants of first and second order are \( \chi_i = \langle s_i \rangle \) and \( \chi_{ij} = \langle s_i s_j \rangle - \langle s_i \rangle \langle s_j \rangle \). It is worth highlighting that according to our definition, \( \chi_{ii} = 1 - \chi_i^2 \).

Defining the frequency \( \nu_i(\alpha) \) of the allele \( \alpha \) at locus \( i \) and the element of the covariance matrix \( M_{ij}(\alpha, \beta) \) relative to the alleles \( \alpha, \beta \) at loci \( i, j \) as

\[ \nu_i(\alpha) = \langle \delta_{a,s_i} \rangle, \]

\[ M_{ij}(\alpha, \beta) = \langle \delta_{a,s_i} \delta_{\beta,s_j} \rangle - \langle \delta_{a,s_i} \rangle \langle \delta_{\beta,s_j} \rangle, \]

(1.10)

it is easy to find relations with the \( \{ \chi_{ij}, \chi_{ij} \} : \nu_i(1) = \frac{1}{2}(1 + \chi_i), \nu_i(-1) = \frac{1}{2}(1 - \chi_i) \) and \( M_{ij}(1,1) = -M_{ij}(1,-1) = -M_{ij}(-1,1) = M_{ij}(-1,-1) = \frac{1}{2} \chi_{ij} \).

As a first step, plugging Eq.(1.2, 1.5, 1.9) into Eq.(1.4), we can write the full explicit master equation for the evolution of \( P(g) \):

\[ \frac{d}{dt} P(g, t) = [F(g) - \langle F \rangle]P(g, t) + \mu \sum_{i=1}^{L} [P(M_i g, t) - P(g, t)] + \]

\[ + r \sum_{g' \in G} C(\xi) \left[ P(g^{(1)}, t) P(g^{(2)}, t) - P(g, t) P(g', t) \right], \]

(1.11)

---

5 The recombination process is akin to a collision process in the kinetic theory of gases where Eq.(1.8) is the celebrated Boltzmann’s molecular chaos hypothesis (or Stoßzahlansatz): the velocities of colliding particles are uncorrelated, and independent of position. But this is not more than a formal analogy, since in our framework there are no such discussions as coarse graining of the phase space, time arrow, Boltzmann H theorem (...) [Grandy Jr, 2012].

6 For an introduction and discussion on cumulants, see App.(A).
valid in the limit $N \to \infty$. Now that the time dependencies are clear, we also drop the $t$ in order to lighten the notation as much as possible. It is straightforward to obtain dynamical equations for $\chi_i$ and $\chi_{ij}$, as shown below.

$\dot{\chi}_i$. We observe that the recombination term has no effect on the dynamics of $\chi_i$ as can be verified explicitly; this can be understood in light of the fact that recombination does not create nor destroy alleles, it simply reshuffles them. Neglecting the last term in eq.(1.11) one has

\[
\dot{\chi}_i = \frac{d}{dt} \left( \sum_g s_i P(g) \right) = \sum_g s_i \frac{d}{dt} P(g) = \sum_g \left( s_i [F(g) - \langle F \rangle] P(g) + \mu \sum_j [P(M_{j,g}) - P(g)] \right) = \langle s_i [F(g) - \langle F \rangle] \rangle - 2\mu \langle s_i \rangle ,
\]

where the last line follows from $\sum_s s_i P(M_{j,g}) = (-1)^{\delta_{ij}} \langle s_i \rangle$.

$\dot{\chi}_{ij}$. The dynamics of the second order cumulant (and in general all higher order cumulants) involves the recombination term. As a preliminary result, let us evaluate the time derivative of $\langle s_is_j \rangle$ under the recombination term alone:\footnote{The reader will easily notice how, following the same steps of this derivation but in the easier case of a single $\langle s_i \rangle$, one gets $\frac{d}{dt}\langle s_i \rangle = 0$, which is exactly what we have foreseen when deriving eq.(1.12) from a consistency argument.}

\[
\frac{d}{dt} \left. \langle s_is_j \rangle \right|_{rec} = \sum_{\xi,\xi',\xi''} C(\xi) s_i s_j \left[ P(g^{(1)}) P(g^{(2)}) - P(g) P(g') \right] \\
= \sum_{\xi,\xi',\xi''} C(\xi) \left[ (\xi s_i^{(1)} (1 - \xi) s_j^{(2)}) + (1 - \xi) s_i^{(1)} s_j^{(2)} \right] + \\
\times \left[ \sum_{\xi, s_i, s_j} C(\xi) s_i s_j P(g) P(g') \right] \\
= \sum_{\xi} C(\xi) \left[ \xi s_i \langle s_is_j \rangle + \xi_i (1 - \xi_j) \langle s_i \rangle \langle s_j \rangle \right] + \\
\times \left[ (1 - \xi_i) \xi_j \langle s_i \rangle \langle s_j \rangle + (1 - \xi_j) (1 - \xi_j) \langle s_i s_j \rangle - \langle s_i s_j \rangle \right] \\
= \langle s_i s_j \rangle \sum_{\xi} C(\xi) \left[ (1 - \xi_i - \xi_j) - (1 - \xi_i) (1 - \xi_j) \right] + \\
\times \langle s_i \rangle \langle s_j \rangle \sum_{\xi} C(\xi) \left[ \xi_i (1 - \xi_j) - (1 - \xi_i) \xi_j \right] \\
= c_{ij} \dot{\chi}_{ij} .
\]
In (a) we have used Eq. (1.6) and changed the first sum over $g, g'$ in a sum over $g^{(1)}, g^{(2)}$; in (b) we have exploited the definition of $\chi_{ij}$ and introduced a new definition
\[
c_{ij} = \sum_{\xi} C(\xi)(\xi_{i}(1 - \xi_{j}) + (1 - \xi_{i})\xi_{j}).
\] (1.14)

We now have all that is needed to evaluate, for $i \neq j$,
\[
\dot{\chi}_{ij} = \frac{d}{dt} (s_i s_j - \chi_{ij}) \\
= \frac{d}{dt} (s_i s_j) - \dot{\chi}_{ij} - \chi_{ij} \\
\overset{(a)}{=} \langle s_i s_j \rangle \frac{d}{dt} (F(g) - \langle F \rangle) + \mu \sum_g s_i s_j \sum_{k=1}^{L} [P(M_k g) - P(g)] + r \frac{d}{dt} \left| \text{rec} (s_i s_j) \right| \\
\overset{(b)}{=} \langle s_i s_j \rangle \frac{d}{dt} (F(g) - \langle F \rangle) - 4\mu \langle s_i s_j \rangle + r \frac{d}{dt} \left| \text{rec} (s_i s_j) \right| + 4\mu \chi_{ij} \\
\overset{(c)}{=} \langle (s_i - \chi_i) (s_j - \chi_j) \rangle \frac{d}{dt} (F(g) - \langle F \rangle) - 4\mu \chi_{ij} - r \text{cij} \chi_{ij}. \quad (1.15)
\]

In (a) we have used Eq. (1.11 - 1.12); in (b) we exploited $\sum_g s_i s_j P(M_k g) = (-1)^{\delta_{ik} + \delta_{jk}} \langle s_i s_j \rangle$ and added $\chi_{ij} \langle F(g) - \langle F \rangle \rangle = 0$; (c) comes again from the definition of $\chi_{ij}$ and from Eq. (1.13).

The Eq. (1.12, 1.15) are at the core of the NS theory of statistical genetics. It is important to underline that so far we have not made any specific choice for $P(g)$, we will deal with it later in this chapter. However, before we go any further, let us add a few comments on the results we got so far.

\begin{itemize}
  \item $c_{ij}$. Given the definition eq. (1.14), $c_{ij}$ can be easily interpreted as the probability that, in the offspring, the alleles at the two loci $i, j$ come from different parents. When recombinations are completely random, we clearly expect $c_{ij} = 1/2 \forall i, j$. Let us discuss how it is possible to guess the value of $c_{ij}$.
  \item Crossover rate. One easy way is to get $c_{ij}$ is to assume that, if there is recombination between two genomes, then each locus undergoes a crossover with probability $\omega$, which we call crossover rate. We will show now that this completely determines $c_{ij}$.
  Consider for simplicity a population of only two individuals $g^{(A)} = \{s^{(A)}_i\}_{i=1}^{L}$ and $g^{(B)} = \{s^{(B)}_i\}_{i=1}^{L}$. A crossover event in the k-th site means $s^{(A)}_k \leftrightarrow s^{(B)}_k$ (exchanged). Note that the crossover rate $\omega$ for each site $i$ does not depend on the present value of the spin $s^{(A)}_i = \pm 1$ but it only refers to the labels $A, B$; explicitly, if there is a recombination, then $P(A \rightarrow B \land B \rightarrow A) = \omega$ and $P(A \rightarrow A \land B \rightarrow B) = 1 - \omega$.
In order to emphasise this, we temporarily clean the notation in the following way
\[ g^{(A)} = \{ A_i \}_{i=1}^L \quad \text{and} \quad g^{(B)} = \{ B_i \}_{i=1}^L \]. Let us now focus on the loci \( i, j \) and suppose that there is a recombination event, one possible result is for instance:

\[
\begin{align*}
g^{(A)}; g^{(B)} : (\ldots A_i \ldots A_j \ldots) ; (\ldots B_i \ldots B_j \ldots) \\
g^{(D_1)}; g^{(D_2)} : (\ldots A_i \ldots B_j \ldots) ; (\ldots B_i \ldots A_j \ldots)
\end{align*}
\]

where \( g^{(D_1)} \) is the genome of the descendent \( D_1 \) and analogously for \( g^{(D_2)} \). We do not loose information if we focus on just one of them, say \( g^{(D_1)} \). With respect to the alleles in the loci \( i, j \) of \( g^{(D_1)} \), it is possible to define two states:

\[
T = (A_i A_i) \cup (B_i B_i) \quad \text{if it is true that those allele come from the same parent}
\]

\[
F = (A_i B_i) \cup (B_i A_i) \quad \text{otherwise.}
\]

It is easy to evaluate the truth table, see Tab.1.1, and read off from it the probability \( c_{ij} = 2\omega(1-\omega) \) that after a recombination event \((ar)\) the state is found to be \( F \) (second column, shaded). The state is obviously \( T \) before recombination \((br)\).

\[
\begin{array}{c|c|c}
\text{State} & \text{Transition Probability} \\
\text{T} & \text{T}_{ar} & \omega^2 + (1-\omega)^2 \\
\text{F} & \text{F}_{ar} & 2\omega(1-\omega)
\end{array}
\]

Table 1.1: Truth table.
Transition probabilities from the states at \( t \) to those at \( t+1 \).

We stress that in the present framework the coefficient \( c_{ij} \) ignores any spatial effect, it is the same for all pairs \( i, j \), regardless of their position in the genome (could they be neighbours or very far apart). The great advantage of this approach is its straightforward algorithmic implementation when trying to simulate biological data, as we will see in Ch.(2). As a conclusive remark, it should be noted that if \( \omega = \frac{1}{2} \), max. uncertainty, then \( c_{ij} = \frac{1}{2} \).

○ Neighbouring variability. A second approach is the one adopted in [Zeng & Aurell, 2020b], more realistic from a biological point of view. We do expect that if two loci are very far apart then they are mostly uncorrelated, so we need to introduce a penalization for the distance.

Let us consider a single recombination event and forget about the crossover rate; in its place we assume a fixed probability \( \rho \) that the recombination causes a crossover \textit{in only one of two neighboring (!) loci}, where \( \rho \) is uniform for all possible pairs of neighbouring loci. Consider two loci \( i, j \) that are \(|i-j|\) loci far apart \(8\). Keeping a similar notation as in the previous case, before recombination one has

\[
\begin{align*}
g^{(A)} &= (\cdots -A_i - A_{i+1} - A_{i+2} - \cdots -A_{j-2} - A_{j+1} - A_j - \cdots) \\
g^{(B)} &= (\cdots - B_i - B_{i+1} - B_{i+2} - \cdots - B_{j-2} - B_{j+1} - B_j - \cdots)
\end{align*}
\]

\(8\) We emphasise that fixing \( \rho \), we are fixing a probability in the space direction rather than in time direction as in the previous case. Moreover, \( \rho \) refers to the variability of a pair of (neighbouring) loci, not to a single locus (previous case).
After the recombination event, focusing again on just one of the new chains, one possible result is:

$$g^{(D_1)} = (\cdots - A_i - T A_{i+1} - F B_{i+2} - T \cdots - A_{j-2} - T B_{j-1} - B_j \cdots)$$

Where the $T, F$ indicate whether it is true/false that the two neighbouring loci come from the same parent, there are $n = |i - j|$ of them. According to our hypothesis, the outcome $F$ has probability $P(F) = \rho$ and $P(T) = 1 - \rho$. As a first approximation, the number of $F$ can be assumed to be binomial distributed: the probability that we find a number $k$ of $F$ is $P(k; n, \rho) = \binom{n}{k} \rho^k (1 - \rho)^{n-k}$.

To get the total probability that alleles at $i, j$ are different after a recombination, it is sufficient to sum $P(k; n, \rho)$ over all odd $k$:

$$c_{ij} = \sum_{k \text{ odd}} \binom{n}{k} \rho^k (1 - \rho)^{n-k}$$

$$= \frac{1}{2} \left[ \sum_{k=0}^{n} \binom{n}{k} \rho^k (1 - \rho)^{n-k} - \sum_{k=0}^{n} \binom{n}{k} (-\rho)^k (1 - \rho)^{n-k} \right]$$

$$= \frac{1}{2} \left[ (\rho + (1 - \rho))^n - (-\rho + (1 - \rho))^n \right]$$

$$= \frac{1}{2} \left[ 1 - (1 - 2\rho)^n \right]$$

$$\simeq \frac{1}{2} \left[ 1 - e^{-2\rho|i-j|} \right], \quad (1.16)$$

where the last line holds if $\rho|i-j| \sim 1$ and $|i-j| \gg 1$. In the limit $|i-j| \to \infty$ we find (consistently) $c_{ij} \sim 1/2$.

**Linkage (dis)equilibrium.** The words linkage disequilibrium (LD) stand for a non-random association of alleles at two or more loci [Slatkin, 2008]. Contrary to what the name may suggest, LD does not ensure either linkage or a lack of equilibrium. Focusing on LD for a pair of loci, several definitions have been proposed, for instance as $M_{ij}(\alpha, \beta)$ defined in Eq.(1.10), but all of them are obviously related to the quantity $\chi_{ij} = \langle s_i s_j \rangle - \chi_i \chi_j$.

The NS model provides us with a law for the evolution of this last quantity, namely Eq.(1.15), from which we see how selection drives $\chi_{ij}$ away from zero, while mutations and recombination act in the opposite way. It is always worth wondering: is this consistent? It is. The very aim of natural selection is to fix the most fit alleles in a population, and this is exactly the meaning of LD. On the other hand, random mutations and recombinations through modifying and mixing the genome sequences tend to destroy patterns

---

9 How is it possible? One has a binomial distribution if the $T, F$ are independent “draws from an urn”, which is apparently in sharp contrast with their definition. In practice, the loci that here appear as neighbours are not physically so: when dealing with a real genomic chain, typically we will analyze only a small subset of all the loci (for instance, those that show substantial variability), that we here “tie together” in our ideal model. Since in general these loci are physically separated by a number of bases, it is possible to assume that $T/F$ at the two sides of the same locus (as above) are independent.
and prevent specific configurations from fixating. The absence of correlation \( \chi_{ij} = 0 \ \forall i,j \) is naturally termed linkage equilibrium (LE) and implies no correlations between loci. It is also of great interest an intermediate phase of weak and steady correlations between loci, it is called quasi-linkage equilibrium (QLE) and we will discuss it in the next section.

One last word to mention that what we have just discussed about LD is only part of the story: there are several other mechanism outside the framework of our model that can cause LD in nature, for instance population subdivision, abrupt changes in population size (bottlenecks), exchange of individuals from different population strains...

1.3 Quasi-Linkage Equilibrium: KNS Theory

The last ingredient we need to master Eq.(1.11) is a specific mathematical form for the distribution \( P(g) \).

The fundamental insight here comes from the pioneering work of M. Kimura back in 1965, when he showed that in a population genetics model which includes recombination, if selection is weak on the time scale of recombination \( \sigma \ll r \) then the allele frequencies change slowly and the selection-induced correlations are weak, steady and can be treated as a perturbation; such a state for a population has been baptised quasi-linkage equilibrium (QLE).

10 Kimura provided a proof only for QLE in a simple two-locus systems, but later several authors have generalised QLE for general multilocus systems, see [R. A. Neher & Shraiman, 2011] and references therein.

in a population genetics model which includes recombination, if selection is weak on the time scale of recombination \( \sigma \ll r \) then the allele frequencies change slowly and the selection-induced correlations are weak, steady and can be treated as a perturbation; such a state for a population has been baptised quasi-linkage equilibrium (QLE).

The NS theory offers a suitable framework to implement this assumption, let us see how.

In [R. A. Neher & Shraiman, 2011] the authors propose the following ansatz for the parametrization of the one-genome probability distribution:

\[
P(g, t) = \frac{1}{Z(t)} \exp \left( \sum_i h_i(t) s_i + \sum_{i<j} J_{ij}(t) s_i s_j \right),
\]

the factor \( Z(t) = \sum_g \exp \left( \sum_i h_i(t) s_i + \sum_{i<j} J_{ij}(t) s_i s_j \right) \) is a normalization, \( h_i(t) \) and \( J_{ij}(t) \) are time dependent single-site and pairwise coefficients, respectively; we will drop again the \( t \) from now on.

We here implement the QLE hypothesis: the second order terms \( \{ J_{ij} \} \) capture to the leading order the correlations induced by selection, hence they are assumed they are small. We will refer to the NS theory in the QLE regime as the KNS theory.

This choice of notation is not accidental, we build one more bridge between the two worlds of Statistical Mechanics and Statistical Genetics: the distribution \( P(g) \) is drawn from an exponential family well-known in physics, the one associated to the generalized Ising model in the canonical ensemble. In fact, here the \( \{ h_i \} \) and \( \{ J_{ij} \} \) play the role of the Ising external magnetic fields and interactions, \( Z \) is nothing but the canonical partition function.

The distribution in Eq.(1.17) can be understood invoking the maximum entropy principle [Cover

---

10 Kimura provided a proof only for QLE in a simple two-locus systems, but later several authors have generalised QLE for general multilocus systems, see [R. A. Neher & Shraiman, 2011] and references therein.
as the one that maximizes the uncertainty under the constraints of having \{\chi_i\} and \{\chi_{ij}\} as first and second order cumulants, see Ch. (3) and App. (D). The following relations hold:
\[
\chi_i = \frac{\partial \log Z}{\partial h_i}, \quad \chi_{ij} = \frac{\partial^2 \log Z}{\partial h_i \partial h_j},
\]  

(1.18)
as can be easily verified. Let us evaluate these quantities in the QLE regime: we first evaluate perturbatively the partition function,
\[
Z = \sum_{g} e^{\sum_{i} h_{\phi_{i}} + \sum_{i<j} J_{ij} \phi_{i} \phi_{j}} \quad \overset{(a)}{=} \sum_{g} e^{\sum_{i} h_{\phi_{i}} \left(1 + \sum_{k<j} J_{kj} \phi_{k} \phi_{j}\right)} \\
= \sum_{g} e^{\sum_{i} h_{\phi_{i}}} + \sum_{k<j} J_{kj} \sum_{g} e^{\sum_{i} h_{\phi_{i}} \phi_{k} \phi_{j}} \\
= \prod_{i} 2 \cosh h_{i} + \sum_{k<j} J_{kj} \left(\prod_{i \neq k} 2 \cosh h_{i}\right) (2 \sinh h_{k})(2 \sinh h_{j}) \\
= 2^L \left(1 + \sum_{k<j} J_{kj} \tanh h_{k} \tanh h_{j}\right) \prod_{i} \cosh h_{i},
\]  

(1.19)
where in \((a)\) we have used \(J_{ij} \sim 0 \forall i \neq j\) and expanded to the first order in \(|J_{ij}|\). Using Eq. (1.18) we get
\[
\chi_{i} \overset{(a)}{=} \frac{1}{Z} \frac{\partial}{\partial h_{i}} \left[2^L \left(1 + \sum_{k<j} J_{kj} \tanh h_{k} \tanh h_{j}\right) \prod_{i} \cosh h_{i}\right] \\
= \frac{2^L}{Z} \left[ \left(1 + \sum_{k<j} J_{kj} \tanh h_{k} \tanh h_{j}\right) \prod_{i \neq k} \cosh h_{i} \sinh h_{j} + \left(1 + \sum_{i \neq j} \frac{\tanh h_{j}}{\cosh^2 h_{i}} \right) \prod_{i} \cosh h_{i}\right] \\
\overset{(a)}{=} \tanh h_{i} + \frac{1}{1 + \sum_{k \neq i} J_{kj} \tanh^2 h_{k} \tanh h_{j} \sum_{i \neq j} J_{ij} (1 - \tanh^2 h_{i}) \tanh h_{j}} \\
\sim \tanh h_{i} + \sum_{i \neq j} J_{ij} (1 - \tanh^2 h_{i}) \tanh h_{j}
\]  

(1.20)
\[
\chi_{ii} \overset{(a)}{=} 1 - \tanh^2 h_{i} \sim 1 - \chi_{i}^2
\]  

(1.21)
\[
\chi_{ij} \overset{(a)}{=} J_{ij} (1 - \tanh^2 h_{i}) (1 - \tanh^2 h_{j}) \sim J_{ij} (1 - \chi_{i}^2) (1 - \chi_{j}^2)
\]  

(1.22)
where in \((a)\) we have derived Eq. (1.20) w.r.t. \(h_{i}\) or \(h_{j}\), while in \((b)\) we have used again Eq. (1.20), upon moving the sum to the LHS. All of the Eq. (1.20 - 1.22) are correct to the first order in \(|J_{ij}|\). One final remark to stress that QLE is very reminiscent of the high-temperature expansion, as we will discuss in Sec. (1.6.2).
1.3.1 The Central Result of KNS theory

The goal now is to understand the dynamics of the parameters in the genotype distribution as a function of the KNS-theory parameters \( \mu, r, \{f_i\}, \{f_{ij}\} \) and the observable first and second order cumulants. Before delving into calculations, let us recap the assumptions we have already made:

i. In Eq.(1.8) a closure: the two-genome distribution factorizes.

ii. In Eq.(1.17) a one-particle distribution function of the generalized Ising form.

iii. Couplings \( \{f_{ij}\} \) small and treated as perturbations in the QLE regime of small correlations.

In addition, following [R. A. Neher & Shraiman, 2011] we will introduce one last assumption:

iv. The mutational contribution can be omitted, the mutation rate is sufficiently small for the approximation \( \mu \sim 0 \) to be valid.

In fact, it cannot be \( \mu = 0 \) otherwise QLE in an infinite population would only be a long-lived transient as the population drifts towards fixation, see Eq.(1.12). Non-zero mutations are necessary to maintain any variability, yet we are assuming that their rate is small enough to forget their contribution in what comes next.

If we now rewrite Eq.(1.11) as an equation for \( \log P(g) \) and substitute Eq.(1.17) we get:

\[
\frac{Z}{\mathcal{Z}} + \sum_i h_i s_i + \sum_{i<j} f_{ij} s_i s_j = F(g) - \langle F \rangle + r \sum_{\xi, g'} C(\xi) P(g') \left[ \frac{P(g^{(1)}) P(g^{(2)})}{P(g) P(g')} - 1 \right] \quad (1.23)
\]

Let us analyze separately the last term of the \textit{rhs}. For simplicity, we set \( \xi_i = 1 - \xi_i \):

\[
\sum_{\xi, g'} C(\xi) P(g') \left[ \frac{P(g^{(1)}) P(g^{(2)})}{P(g) P(g')} - 1 \right] = \quad \text{(a)}\\
= \sum_{\xi, g'} C(\xi) P(g') \left( e^{\sum_{i<j} f_{ij} [\xi_i \xi_j + \xi'_i \xi'_j] - s_i s_j} - 1 \right) \quad \text{(b)}\\
= \sum_{\xi, g'} C(\xi) P(g') \sum_{i<j} \left[ (\xi_i \xi_j + \xi'_i \xi'_j - 1) (s_i s_j + s'_i s'_j) + (\xi_i \xi_j + \xi'_i \xi'_j) (s_i s'_j + s'_i s_j) \right] \quad \text{(c)}\\
\approx \sum_{\xi} C(\xi) \sum_{i<j} \left[ (\xi_i \xi_j + \xi'_i \xi'_j - 1) (s_i s_j + s'_i s'_j) + (\xi_i \xi_j + \xi'_i \xi'_j) (s_i s'_j + s'_i s_j) \right] \quad \text{(d)}\\
= \sum_{i<j} c_{ij} [s_i \langle s_j \rangle + (s_i s_j) - (s_i s_j) - (s_i s_j)]
\]

(1.24)
In (a) we have used Eq. (1.17), inverted the relations Eq. (1.6) to express $s_i^{(1)} = \tilde{c}_i s_i + \tilde{c}_i' s_i'$, $s_i^{(1)} = \tilde{c}_i s_i + \tilde{c}_i' s_i'$, and $s_i^{(1)} + s_i^{(2)} = s_i + s_i' = 0$ that cancel for each field $h_i$; in (b) we have expanded to the first order in $|f_{ij}|$; in (c) we have averaged over $P(g')$; in (d), finally, we have used $c_{ij} = \sum\xi C(\xi)(\tilde{c}_i\tilde{c}_j + \tilde{c}_i\tilde{c}_j') = \sum\xi C(\xi)(1 - \tilde{c}_i\tilde{c}_j - \tilde{c}_i\tilde{c}_j')$.

Substituting Eq. (1.24) into Eq. (1.23) and using Eq. (1.3) we get:

$$-\frac{\dot{Z}}{Z} + \sum_i h_i s_i + \sum_{i<j} j_i s_i s_j =$$

$$= \bar{F} - \langle F \rangle + \sum_i f_i s_i + \sum_{i<j} f_{ij} s_i s_j + r \sum_{i<j} c_{ij} l_{ij} [s_i(s_j) + \langle s_i \rangle s_j - \langle s_i s_j \rangle]$$

(1.25)

Dynamical equations for $\{h_i, f_{ij}\}$ emerge when collecting together terms with the same monomials in $s_i$:

$$h_i = f_i + r \sum_{j \neq i} c_{ij} l_{ij} \chi_j$$

(1.26)

$$f_{ij} = f_{ij} - rc_{ij} l_{ij}$$

(1.27)

In the case where the recombination rate is very high $\sigma/r \ll 1$ the $\{l_{ij}\}$ rapidly reach the steady state, which deserves to be emphasized:

$$f_{ij} = l_{ij} \cdot rc_{ij}$$

(1.28)

This is a crucial result for the KNS theory because it paves the way for the inference of the epistatic fitness landscape encoded in the $\{f_{ij}\}$ from the couplings $\{l_{ij}\}$ that, even if not observable, can be inferred from data. We will extensively discuss and exploit this observation in Ch. (3). Finally, substituting the steady-state Eq. (1.28) in Eq. (1.26), we find

$$h_i = f_i + \sum_{j \neq i} f_{ij} \chi_j = \bar{f}_i ,$$

(1.29)

where $\bar{f}_i$ is the effective strength of selection on locus $i$ and it collects the contribution of every $\chi_i$ by means of the epistatic interaction. From Eq. (1.21 - 1.22) we note that while the $\chi_{ii}$ depends exclusively on the frequency $\chi_i$, the off-diagonal correlations are determined by the trade-off between epistasis and recombination i.e. $\sigma_e/r$.

In conclusion of our discussion on the QLE phase, let us look at its consequences for the dynamics of the first and second order cumulants.

In view of Eq. (1.28) and considering Eq. (1.15) with $\mu \sim 0$, the correlations will rapidly approach the expression:

$$\chi_{ij} = \frac{f_{ij}}{rc_{ij}} (1 - \chi_i^2)(1 - \chi_j^2) , \quad i \neq j .$$

(1.30)
Neglecting all terms but the one due to recombination in Eq. (1.15), we can write the dynamics of the second order cumulants $\chi_{ij}$ as an exponential decay to the asymptotic value $\bar{\chi}_{ij}$ i.e.

$$\chi_{ij} = \bar{\chi}_{ij}(1 - e^{-rc_{ij}t}) ,$$  \hspace{1cm} (1.31)

whose dynamic is indeed driven by the ordinary differential equation

$$\dot{\chi}_{ij} = rc_{ij} \bar{\chi}_{ij} e^{-rc_{ij}t} = rc_{ij} (\bar{\chi}_{ij} - \chi_{ij}) = f_{ij}(1 - \chi_i^2)(1 - \chi_j^2) - rc_{ij}\chi_{ij} ,$$  \hspace{1cm} (1.32)

where in (a) we have used Eq. (1.31) and in (b) Eq. (1.30).

The first order cumulants instead evolve according to the following dynamical equations:

$$\dot{\chi}_i = \langle s_i F \rangle - \chi_i \langle F \rangle = \partial_{h_i} \langle F \rangle \overset{(b)}{=} \sum_j \partial_{\phi_j} \chi_{ij} \partial_{\chi_j} \langle F \rangle = \sum_j \chi_{ij} \partial_{\chi_j} \langle F \rangle ,$$  \hspace{1cm} (1.33)

where in (a) we have used Eq. (1.12) with $\mu = 0$; in (b) the chain rule of differentiation; in (c) the fact that $\chi_{ij} = \partial \chi_i / \partial h_j$, see Eq. (1.18). We see from the rhs of Eq. (1.33) that the allele means evolve so to maximize $\langle F \rangle$, there are $L$ such equations and they are all coupled by the correlations $\{\chi_{ij}\}$.

This is nevertheless an enormous simplification with respect to the $2^L$ ordinary differential equations for each possible $g$ that we would have in general: in the QLE regime where $\sigma / r \ll 1$ and the correlations rapidly approach their steady state, the $L$ Eq. (1.33) are the only relevant dynamical equations and they define the $L$-dimensional QLE manifold. As long as the QLE holds, the genotype distribution (hence the population average of any trait) is confined on such manifold i.e. it can be parametrized by the set of time-dependent first cumulants $\{\chi_i(t)\}$.

### 1.4 Random Genetic Drift

Our discussion has considered hitherto the infinite population limit $N \to \infty$ and neglected any issue arising from taking into account the stochastic nature of the birth/death processes in a finite population. In the case $N < \infty$, the element of chance introduced by the random sampling of the individuals that survive from one generation to another is alone capable of driving an allele to fixation or extinction. Such an effect is well known in population genetics: as mentioned in Sec. (1.1), it is named genetic drift.

The goal of this section is to implant this last phenomenon into the framework of the KNS-theory. Before dealing with this discussion, tough, we advise the reading of App. (B), that we dedicate to a short detour on the simplest class of models of genetic drift, the Fisher-Wright (FW) models. Aside from helping the intuition on the mechanism of the genetic drift, it is also useful to get a broader picture of the historical context of population genetics at its roots (1930-1970).

We will limit ourselves to describing the framework and the main results, details will be found
We will not go further into this discussion, but let us show the way. Once the autocorrelations where the autocorrelation of the last term is found to be

As an example, let us consider the \( \{X_{ij}\} \). In the QLE, as we have discussed, these quantities relax much faster than the \( \{X_i\} \), hence we can assume the latter constant when studying the dynamics of the former. In particular, it is possible to decompose the solution into a deterministic part, which is nothing but Eq.\((1.30)\), and a stochastic contribution:

where the autocorrelation of the last term is found to be

apart from the already mentioned dependence on the system size, we observe the dependence on the recombination rate \( r \) that dampens not only the deterministic part but also the stochastic effect, as it should be.

We will not go further into this discussion, but let us show the way. Once the autocorrelations Eq.\((1.38)\) are known, the stochastic contribution to Eq.\((1.37)\) is determined; however, in order to use this result, one has to distinguish the case when the stochastic component dominates from the opposite one. With this proviso, it is possible to study the Langevin equations for \( \{X_i\} \) Eq.\((1.34)\) using Eq.\((1.37)\). The aforementioned distinction between the two regimes is concretely made by comparing the deterministic part of the \( \chi_i(t) \) with the stochastic term upon averaging the latter over the timescale of the dynamics of the \( \{X_i\} \), given by the inverse of \( \partial\chi_i\langle F\rangle \), see Eq.\((1.33)\). It is also possible to restore and take into account the mutational...
contribution $\mu \neq 0$ into Eq. (1.34), in which case one finds the deterministic component of $\chi_{ij}(t)$ to dominate when $N\mu \gg 1$ and $f_{ij} \gg \mu$, the stochastic one otherwise [R. A. Neher & Shraiman, 2011].

## 1.5 KNS Theory for Categorical Data

One of the major limitations of the discussion we have held so far apparently is the assumption of biallelic loci $s_i = \pm 1$: this is never true, since in real genetic data one always finds (at least) four different alleles, namely A, C, G, T. Fortunately, the framework we have set up and most importantly the results we have obtained can be smoothly generalized to the case of multi-allelic loci. Following [Gao et al., 2019], we here briefly revisit some aspects of the NS theory and extend it to such categorical data.

We consider again an infinite population, each individual is a genomic chain which consists in $L$ loci $g = \{z_1, \ldots, z_L\}$; each locus can in turn take $q_i$ values (alleles) i.e. $z_i = 1, \ldots, q_i$. $P(g,t)$ is the probability of finding the genotype $g$ at time $t$. Let us generalize Eq. (1.10) and define

$$v_i(\alpha) = \langle \delta_{z_i,\alpha} \rangle ,$$

$$M_{ij}(\alpha, \beta) = \langle \delta_{z_i,\alpha} \delta_{z_j,\beta} \rangle - v_i(\alpha)v_j(\beta) :$$

here the averages are over $P(g)$, $\delta$ indicates the $\delta$-Kronecker, $v_i(\alpha)$ stands for the frequency of the allele $\alpha$ at the $i$-th locus, while $M_{ij}(\alpha, \beta)$ is the element of the covariance matrix between loci $i$ and $j$ relative to the alleles $\alpha$ in $i$ and $\beta$ in $j$. The following normalizations hold: $\sum_{\alpha=1}^{q_i} v_i(\alpha) = 1$, $\sum_{\beta=1}^{q_j} M_{ij}(\alpha, \beta) = \sum_{\alpha=1}^{q_i} M_{ij}(\alpha, \beta) = 0 \forall i, j$.

We can still write formally the Eq. (1.11) but some more notational effort is required, let us consider one by one the drivers of evolution.

- **Fitness.** The formal expression for the fitness term is the same as in Eq. (1.2)

$$\frac{d}{dt} \left| f_{it} \right| P(g) = (F(g) - \langle F \rangle)P(g) .$$

We only need to rewrite the fitness function in a form appropriate for the multi-allelic case:

$$F(g) = \hat{F} + \sum_i f_i(z_i) + \sum_{ij} f_{ij}(z_i, z_j) ,$$

where $f_i(z_i)$, $f(z_i, z_j)$ are functions of the alleles $z_i$, $z_j$ and in general $f_{ij}(z_i, z_j) \neq f_{ji}(z_i, z_j)$. We stress that, unlike the biallelic case, there is no straightforward way to write them explicitly.
• **Mutations.** Let \( \mu_{\alpha,\beta}^{(i)} \) be the mutation rate at which the allele \( \alpha \) mutates into \( \beta \) at locus \( i \). In the most general framework, \( \mu_{\alpha,\beta}^{(i)} \neq \mu_{\beta,\alpha}^{(i)} \). We can generalize Eq.(1.5) into the following

\[
\left. \frac{d}{dt} \right|_{\text{mut}} P(g) = \sum_i \sum_{\alpha,\beta} \delta_{\alpha,\beta} \left( \mu_{\beta,\alpha}^{(i)} P(t^{(i)}_{\alpha,\beta}g) - \mu_{\alpha,\beta}^{(i)} P(g) \right).
\] (1.42)

The operator \( t^{(i)}_{\alpha,\beta} \) acts on the genomic sequence \( g \) as follows: if in the \( i \)-th locus of \( g \) there is the allele \( \alpha \) then it is changed to \( \beta \), otherwise nothing happens.

• **Recombination.** No change is needed for the contribution due to recombination, that we copy-paste from Eq.(1.7)

\[
\left. \frac{d}{dt} \right|_{\text{rec}} P(g) = r \sum_{\xi \neq g'} C(\xi) \left[ Q(g^{(1)},g^{(2)}) P_{2}(g^{(1)},g^{(2)}) - Q(g,g') P_{2}(g,g') \right].
\] (1.43)

In Eq.(1.6), we only need to formally substitute \( s_i \rightarrow z_i \).

All other explanations, observations, limitations hold true as described in Sec.(1.2.2 - 1.2.4). Following the same reasoning as in the first part of this chapter, we now assume that the two-genome distribution factorizes Eq.(1.8) in a simple product of two one-genome Potts-like distributions:

\[
P(g,t) = \frac{1}{Z(t)} \exp \left( \sum_i h_i(z_i,t) + \sum_{i,j} J_{ij}(z_i,z_j,t) \right),
\] (1.43)

which is the natural generalization of Eq.(1.17): \( Z \) is the partition function, \( h_i(z_i,t), J_{ij}(z_i,z_j,t) \) are functions of the alleles \( z_i, z_j \) and of time \( t \), we will drop the latter dependence in the next formulae. We can again justify Eq.(1.43) by resorting a max. entropy argument but in view of the normalizations imposed to Eq.(1.39-1.40) it is clear that we do not need all of the parameters \( \{h_i, J_{ij}\} \): in fact, there is an over parametrization issue in the previous Potts distribution as introduced here and one possible way to fix this gauge invariance is to impose the so called Ising gauge: \( \sum_a h_i(a) = \sum_a J_{ij}(a,\beta) = \sum_\beta J_{ij}(a,\beta) = 0 \ \forall i,j \).

The stage is set at this point to introduce the QLE assumption: in the case when recombination is prominent, selection is weak we can assume the couplings \( J_{ij} \) to be small. Neglecting the mutational contribution and following exactly the same steps shown in Sec.(1.3.1) with some more notational burden we get to generalize Eq.(1.25) to the following:

\[
\frac{\partial}{\partial t} \mathcal{Z} + \sum_{i,a} c_{i,a} \delta_{z_i,a} + \sum_{i,j,a,\beta} f_{ij}(a,\beta) \delta_{z_i,a} \delta_{z_j,\beta} =
\]

\[
\sum_{i,j,a,\beta} c_{ij} J_{ij}(a,\beta) \left( \delta_{z_i,a} E_Q[\delta_{z_j,\beta}] + E_Q[\delta_{z_i,a}] \delta_{z_j,\beta} - \langle Q \rangle \delta_{z_i,a} \delta_{z_j,\beta} - E_Q[\delta_{z_i,a} \delta_{z_j,\beta}] \right) +
\]

\[
\sum_{i,a} f_{i}(a) \delta_{z_i,a} + \sum_{i,j,a,\beta} f_{ij}(a,\beta) \delta_{z_i,a} \delta_{z_j,\beta}.
\] (1.44)
where \( c_{ij} \) has been introduced in Eq.\((1.14)\) and the following definitions have been employed:

\[
\langle Q \rangle = \sum_{g'} Q(g,g')P(g')
\]  \hspace{1cm} (1.45)

\[
E_Q[\delta z'_{i,\alpha}] = \sum_{g'} \delta z'_{i,\alpha}Q(g,g')P(g')
\]  \hspace{1cm} (1.46)

\[
E_Q[\delta z'_{i,\alpha}\delta z'_{j,\beta}] = \sum_{g'} \delta z'_{i,\alpha}\delta z'_{j,\beta}Q(g,g')P(g')
\]  \hspace{1cm} (1.47)

We can compare terms in \textit{LHS} and \textit{RHS} of Eq.\((1.44)\) and extract a dynamical equation for the couplings of the Potts distribution:

\[
J_{ij}(\alpha, \beta) = f_{ij}(\alpha, \beta) - r\langle Q \rangle c_{ij}J_{ij}(\alpha, \beta).
\]  \hspace{1cm} (1.48)

Finally, we impose the stationary condition and read off the steady state

\[
J_{ij}(\alpha, \beta) = \frac{f_{ij}(\alpha, \beta)}{r\langle Q \rangle c_{ij}},
\]  \hspace{1cm} (1.49)

In the case where the relative rate of recombination \( Q(g,g') \) is constant for all pairs \( g, g' \) or approximately so, it can be absorbed in \( r \). We see that the analogy between the last formula and the one for the biallelic case is complete: Eq.\((1.28)\) generalizes to the multi-allelic case in the simplest possible way. Here too, we underline that even if recombination cannot influence the frequencies \( \nu_i(\alpha) \), it couples the dynamics of each \( h_i(\alpha) \) to single locus statistics in every other site by means of the couplings \( J_{ij}(\alpha, \beta) \), as it can easily read off from Eq.\((1.25)\).

We finally highlight that this framework has been used in [2019] to reconstruct epistasis from a database of 3,000 genotypes (100,000 loci each) of the human pathogen \textit{S. pneumoniae}: the first necessary result that they use has been treated here, Eq.\((1.49)\); the second one, DCA reconstruction, will be discussed in Ch.\((3)\).

\section*{1.6 Contours}

\textit{On the 25th Sept. 1264, at the break of the day, the Duke of Auge appeared at the summit of the keep of his castle, there to consider, be it ever so little, the historical situation.}

It is widely recognized as a good practice to end a scientific discussion by clearly stating which the contours are, which the perspectives, which the limitations and, most importantly, the state of the art. The iconic opening words of the \textit{R. Queneau}’s masterpiece, \textit{The blue flowers}, gift us an elegant metaphor that precisely encodes the aim of these last pages; similar sections will be found at the end of each chapter.

Let us briefly recap what we have discussed this far. The NS theory has provided a stage where to set a quantitative description for population genetics. The approach adopted is based
on the modelling of genotypes as Ising spin chains and a master equation for the genotype distribution \( P(g) \), which is capable to grasp basic mechanisms for those that we believe to be the drivers of the evolutionary process: mutations, recombination, natural selection (encoded in a fitness landscape). Among the most remarkable results, the possibility of writing explicit dynamical equations for the first and second order cumulants of the distribution. Assuming factorization for the two-genotypes distribution, an Ising-like \( P(g) \) and in the QLE regime of high recombination - weak selection (KNS theory) a simple relation can be found between the epistatic components of the fitness landscape and the Ising couplings. Modifications and extents can be added to take into account finite population effects or categorical data.

We will now cast a critical eye on our results, zooming out from the particular to the general.

### 1.6.1 Biological Realism.

When designing a model, one always have to neglect details to an extent that is a trade-off between the realism of the framework and the solvability of the equations. Unfortunately, when it comes to evolutionary models the theorist is forced to make enormous simplifications. Some of them have already been pointed out throughout this chapter for the case of the NS theory, among the most important:

- **Natural selection.** We have neglected the possibility for the fitness function \( F(g) \) to change in time (fitness seascape), to depend or instantaneous statistics of the population (e.g. frequency-dependent selection);

- **Mutations.** Our pointlike mutations are an extreme idealization, we would like to describe for instance insertions and deletions (which entails allowing \( L \) to take different values), mutational hotspots...

- **Recombination.** In a recombination event, crossovers swap fragments of genetic material, not individual loci. Different mechanisms could be implemented to mimic other forms of bacterial recombination (e.g. conjugation). Furthermore, we would like to understand the consequences of recombination hotspots in our genomic sequences (i.e. non uniform crossover rate).

- **Ad libitum.** Many other phenomena affect natural evolution: catastrophes which lead to population bottlenecks, geographical separation (island models), hybridization ...

Some of these mechanisms can be implanted in the NS-framework, some will probably require a different theoretical approach, some others a complete paradigm shift.
1.6.2 Beyond the QLE phase.

Calculations based on the QLE assumption are meaningful when there is strong recombination among individuals but they are likely to yield nonsense results otherwise \( (r \sim 0 \text{ or } r = 0) \). For instance, Eq.\( (1.28) \) has been tested on suitable biological data (\textit{S. pneumoniae}) but several other organisms are out of reach for this analysis (\textit{Streptococcus pyogenes, Vibrio parahaemolyticus}...), see [2019] and references therein. The question arises of what happens when, for instance, selection is prominent and in general when we step out from the QLE assumptions. Large part of this thesis will address this question Ch.\( (4-5) \), in this section we briefly discuss what is known as the clonal - competition (CC) phase (or clonal interference).

In Sec.\( (1.3.1) \) we have learnt how the assumption of high recombination constrains the evolutionary dynamics on the QLE manifold, which means in particular that \( P(g) \) can be parametrized by the allele frequencies \( \{\chi_i\} \). As a consequence of the high recombination rate, in the QLE phase the fitness of individual genotypes that appear by chance does not matter, the target of the evolutionary process are the alleles (specifically, the allele frequencies).

Turning the picture upside down, we now consider the opposite case of weak recombination and strong selection. In such a case very fit genotypes will be amplified by selection when appearing by chance during the evolutionary process, genetic diversity will dramatically decrease and a large fraction of the population will condense in a group of few large clones in competition with each other, see also [R. A. Neher & Shraiman, 2009]. The dynamics of the system will "fly out" of the QLE manifold and the statistics of the population, e.g. cumulants, will be slaved to the fate of the largest clones. In other words, selection will now act upon genotypes rather than on alleles.

We can draw a parallel with the framework of statistical physics. The claim is that the transition QLE\( \rightarrow \)CC phase is related to the spin-glass transition in disordered physical systems and magnets [Castellani & Cavagna, 2005]. Indeed the distribution Eq.\( (1.17) \) is very reminiscent of

\[
P(\{s_i\}) \sim e^{-\frac{1}{T} \sum_i k_i s_i + \sum_{ij} J_{ij} s_i s_j + ...}
\]

which has been widely studied and can be found in three states:

- **Paramagnetic.** For high temperatures \( T \rightarrow \infty \) when correlations are weak throughout the system, analogous to our QLE phase \( (\sigma / r \sim 1/T) \);

- **Ferromagnetic.** For low temperatures \( T \rightarrow 0 \) and \( \{J_{ij}\} \) that all have the same sign: a global energy minimum exists and corresponds to an ordered state where the spins are all aligned. In a genomic population, this would correspond to a single fit genotype taking the lead and expanding in a huge clone.

- **Glassy.** For low temperatures \( T \rightarrow 0 \) and \( \{J_{ij}\} \) that have erratic signs. Many comparable local minima exist, the (frustrated) system falls in one of them and there is a strong dependence of the fate of the system from its initial realization. This is the analogous of our CC - phase.
The spin-glass approach to the description of the clonal condensation is pursued in [R. Neher & Shraiman, 2012] assuming no mutations, therein the analogous of the Parisi order parameter is found to be the probability \( Y \) that two random individuals sampled from the population have the same genotype.

Finally, interesting insights can be found in [Desai & Fisher, 2007] on the evolutionary dynamics beyond those just discussed and in particular about the balance between selection and mutations in absence of recombination.

### 1.6.3 Different Approaches to the Evolutionary Problem

We have so far described a forward approach to the problem of modelling evolution, the NS theory aims at predicting evolution, given initial conditions.

Yet the possibility of observing evolution with an experiment so test directly the theory is restricted up to now to very few cases, the most important issue being time (we would not wait millions of years to sadly see our theory being rejected).

Typically the evolutionary biologists have (a subset of) data of a specific realization of evolution, the one that we observe, and do not know the initial parameters of the population. Interesting questions could then be: when was the most recent common ancestor for an observed group of individuals? How many individuals of a specific lineage where in the population at a time \( t \) in the past?

Such questions are better addressed exploiting a backward approach, in some sense dual to the one we have learnt about. A simple model is that of the Kingman coalescent: suppose a population of fixed size \( N \), no recombination, no selection (neutral model) and no subdivisions of the population for whatever reason. Starting from now and going back in time, suppose that any two lineages merge at random with rate \( k(k-1)\sigma^2/2N \) where \( k \) is the number of remaining lineages and \( \sigma^2 \) is the variance of the offspring number. For instance one finds that the time to the most recent ancestor for a large population is approximately \( 2N/\sigma^2 \).

For a more sophisticated example in this class of backward models that includes selection, see for instance [Walczak et al., 2012].

Several different approaches exist on the evolutionary problem besides those discussed in these
chapter, each of them looking at the same problem but from different perspectives, casting a
different light, answering different questions. We do not claim to have discussed everything
worth to be discussed and refer to [Manrubia et al., 2020; R. A. Neher & Walczak, 2018] for
recent relevant reviews of the state-of-art in evolutionary dynamics.

In The blue flowers by R. Queneau, the next sentence after the ones quoted at the beginning of
this discussion is:

\[ It \text{ was somewhat confused. } \]

Whether or not this is also our case, the reader will tell.
This short chapter is dedicated to the description of the simulation tool we have extensively employed and exploited throughout this work. In Ch.(1) it has been discussed how the scarcity of forward experiments of population genetics may be a serious issue for the purpose of testing our theories. Several simulation packages have been written to compensate for this absence and are also widely used for practical purposes such as studying the evolution of drug resistance, selecting seasonal vaccines... Obviously, the value of simulations for scientific purposes extends much beyond the mere assessments of pre-existing theories: aside from epistemological issues [Winsberg, 2019], simulations can be predictive, they can help us explore uncharted regions, shed light when we are left in the dark.

For all these purposes and in the context of population genetics, we will use FFPopSim, the software developed by F. Zanini and R.A. Neher, based on the NS theory, see [Zanini & Neher, 2012] with Supplementary data. It is implemented in C++ with a Python2 wrapper. In Sec.(2.1) we illustrate the general structure of the simulation package and the features of our interest, details can be found in the reference indicated and in the package documentation. In Sec.(2.2) we will list and show some useful charts that can be used to visualize evolution in silico. Finally, in Sec.(2.3) we discuss critically the role of simulations in the study of the evolutionary problem.

### 2.1 Generalities

The purpose of FFPopSim is to simulate the evolutionary process for a population of haploid individuals, identified by their genomes \( g = (s_1, \ldots, s_L) \) with biallelic loci \( s_i = \pm 1 \).

There are two classes of forward simulations of population genetics, the genotype-based that track the number of individuals of each possible genotype and the individual-based that track the genotype of every existing individual, see Fig.(2.1). The crucial factor is the length of the genomes \( L \): one has \( 2^L \) possible genotypes and it is clear that a genotype-based approach soon becomes unfeasible from a computational point of view. On the contrary, the individual-based simulations are in general more efficient when the number of loci is high. FFPopSim implements both classes but we will exclusively focus on the second one, therein named haploid_highd.

In this case, a discrete generation scheme is implemented, in which every individual at every generation undergoes each of the processes that drive evolution with tunable probabilities. FFPopSim keeps track of the distribution \( P(g) \) that changes under the effect of mutation, recom-

---

1 Except for the mathematical symbols, we will use the teletype-font for everything related to simulations.
Figure 2.1: Strategies for forward simulations: The left panel illustrates a scheme that tracks the abundance of each possible genotype, encoded as a bit string. This is feasible up to $L \sim 20$ and is implemented in FFPopSim as the class `haploid_lowd`. The right panel illustrates individual-based simulations that track existing genotypes only. FFPopSim provides individual-based simulations through the class `haploid_highd`. From [2012].

Combination, natural selection. Random drift is taken into account by resampling each individual at each generation from a Poisson distribution with mean $NP(g)$, which results in a population of fluctuating size $N \pm O(\sqrt{N})$. We mention that the algorithms for storing and handling genomic sequences are based on a Fast Fourier Transform (FFT) of the genotype space.

2.1.1 Initialization

The class `haploid_highd` is instantiated by specifying the structure of the evolving population, the rates of the evolutionary mechanisms and initial conditions. In Tab. (2.1) we show a typical set of parameters that we will initialize when running a simulation. For future reference, we call this specific set QLE Simulation (QLES).

<table>
<thead>
<tr>
<th>FFPopSim</th>
<th>QLES</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
<td>1.000</td>
<td>carrying capacity</td>
</tr>
<tr>
<td>$L$</td>
<td>25</td>
<td>n. of loci</td>
</tr>
<tr>
<td>$T$</td>
<td>5.000</td>
<td>n. of generations</td>
</tr>
<tr>
<td>$r^*$</td>
<td>0.5</td>
<td>outcrossing rate</td>
</tr>
<tr>
<td>$\omega$</td>
<td>0.5</td>
<td>crossover rate</td>
</tr>
<tr>
<td>$\mu$</td>
<td>0.3</td>
<td>mutation rate</td>
</tr>
<tr>
<td>${f_i}$</td>
<td>0.0</td>
<td>additive fitness</td>
</tr>
<tr>
<td>${f_{ij}}$</td>
<td>$N(0,\sigma_e)$</td>
<td>epistatic fitness</td>
</tr>
</tbody>
</table>

Table 2.1: Parameters of the QLE Simulation (QLES). Low mutation - epistasis, high recombination. Random initial configuration. No additive fitness. Epistatic fitness coefficients are Gaussian distributed with zero mean and width $\sigma_e = 0.004$; moreover, $f_{ij} = f_{ji}$ and $f_{ii} = 0 \forall i,j$. Random initial conditions.

In the first place, the structure of the population is specified: the number of loci $L$ and the carry-
These probabilities are indeed consequences of the discreteness of the computer simulation. The outcrossing rate \( r^* \) is the probability of sexual reproduction. The crossover rate \( \omega \) and the mutation rate \( \mu \) are understood as per site per generation. Finally, \( \{f_i, f_j\} \) are the coefficients of the fitness function Eq. (1.3).

In an individual-based model like the one we are considering, the fundamental object undergoing evolution is not the genotype, but the clone \( c_i = (g_i, n_i) \) i.e. the pair of a genotype \( g_i \) and the number \( n_i(t) \) of individuals in the population that have that genotype at time \( t \). The initial conditions therefore are specified by setting \( \{c_i\} \) i.e. \( N \times L \) Boolean values for the \( \{g_i\} \) and \( N \) integers for the \( \{n_i(0)\} \). We will typically set random initial condition for the genotypes and \( n_i(0) = 1 \) \( \forall i \).

### 2.1.2 Evolution

From the point of view of the class `haploid_highd` of `FFPopSim`, the population \( \mathcal{P} \) is a set of clones. At each generation, the size of each clone is updated and new clones that emerge because of mutations or recombinations are added. A discrete time step (generation) \( \Delta t = 1 \) is implemented by enforcing mutations, recombination, selection, we illustrate them one by one below.

- **Mutations.** Mutations are bit-flip operations in a genotype. In the first place the number of individuals that undergo mutation is determined and marked, each of them mutates with probability \( 1 - e^{L\mu} \). Every individual that has been selected suffers at least one mutation, the exact number \( K \) being drawn from a Poisson distribution \( \mathcal{P}_{L,\mu}(K) \) with mean \( L\mu \). Target loci are chosen randomly.

- **Selection.** Let \( n_i(t) \) be the size of the clone \( i \) at time \( t \). We enforce selection by updating \( n_i(t) \to n_i(t + 1) \sim \mathcal{P}_\lambda \) where

\[
\lambda = \frac{1}{\langle e^F \rangle} e^{F(g_i)} + 1 - \frac{1}{\langle e^{F} \rangle} \sum n_j(t).
\]

In words, we draw the size \( n_i(t + 1) \) of the \( i \)-th clone at time \( t + 1 \) from a Poisson distribution with mean \( \lambda \) as in Eq. (2.1), where \( F(g) \) is the fitness function Eq. (1.3) and the average \( \langle e^F \rangle \) is over the entire population. We note that for \( F(g) \ll 1 \), \( e^F(g) / \langle e^F \rangle - 1 \sim \)

\footnote{These probabilities are indeed consequences of the discreteness of the computer simulation. The rate \( \mu \) as introduced in Sec.(1.2.3) is referred to a continuous-time formulation of the evolution. Let \( \mathcal{E} \) be the event that a mutation appears in an individual; suppose such events are independent and that the probability of two of them happening at the same time is negligible. If their average rate is \( \mu \) then the number \( k \) of events \( \mathcal{E} \) in the time interval \( \Delta t \) is \( \sim \mathcal{P}_{\mu \Delta t} \) where \( \mathcal{P}_\lambda(k) = \lambda^k e^{-\lambda} / k! \) is the Poisson distribution. The number of such mutations in a genome of length \( L \) in the interval \( \Delta t \) is the random variable \( K = \sum_{j=1}^{L} k_j \) that, being the sum of \( L \) i.i.d. Poisson random variables, is again Poisson distributed, with mean \( L\mu \Delta t \) i.e. \( K \sim \mathcal{P}_{L,\mu \Delta t} \). Finally, the probability that there is at least one mutation is \( 1 - \mathcal{P}_{L,\mu \Delta t}(0) = 1 - e^{L\mu \Delta t} \).}
\( F(g) - \langle F \rangle \), so that we retrieve Eq.(1.2). The growth rate adjustment \( \exp(1 - \sum_j n_j(t)/N) \) is implemented to constrain the population close to the carrying capacity \( N \).

- **Recombination.** A fraction \( r^* \) of the offspring at the previous step are designated for sexual reproduction.\(^3\) They are shuffled and randomly paired. For each pair a crossover pattern \( \{\xi_i\} \) as in Sec.(1.2.4) is created and the recombination is implemented by discarding parents and replacing them with two new individuals accordingly. Crossovers are assumed to happen independently between any two loci with rate \( \omega \).

### 2.2 OBSERVABLES

Understood the machinery of FFPopSim, we are ready to exploit it. The general structure of the Python program we have used to wrap this package and handle input/output of information is sketched in App.(C), we here focus on the description of some useful charts that can be drawn in order to observe and understand the evolutionary process in silico.

We initialize a simulation for illustrative purposes with the parameters as in Tab.(2.1), they are tuned so to set the system in a QLE regime i.e. rare mutations, weak fitness effects, high recombination.

- **\( \chi_i \) : First Order Cumulants.** The most straightforward statistics we can evaluate and track as a function of \( t \) is population average of the alleles in each locus i.e. \( \chi_i = \langle s_i \rangle \). In a QLE phase with no additive fitness as in Tab.(2.1), we expect the the strong recombination rate to prevent any fixation of an allele in the population and to constrain the first order cumulants around the random-state value \( \chi_i \sim 0 \). Fig.(2.2) shows the time evolution of all the \( L \) quantities \( \{\chi_i(t)\}_{i=1}^L \).

- **\( \chi_{ij} \) : Second Order Cumulants.** A crucial set of quantities for the NSK theory has been that of the second order cumulants \( \{\chi_{ij}(t)\} \) which can be easily computed. In Fig.(2.3) we show \( L \) of these \( L^2 \) quantities, in particular \( \{\chi_{ij}(t)\}_{i=1}^L \), analogous results are obtained for all other \( j \). In a QLE state, we expect \( \chi_{ij} \sim 0 \quad \forall j \neq i \) and \( \chi_{ii} = 1 - \chi_i^2 \sim 1 \).

- **Fitness statistics.** At each generation step \( t \) we can get information about the population averaged statistics of the fitness, in particular its mean and standard deviation, see Fig.(2.4). In a QLE phase we see no odd behaviour of this observables: after a short transient, they simply settle on two specific values, fluctuations aside. Such "asymptotic" value for the fitness mean is here positive, we comment on this below.

- **Clonal Structure.** The chart is drawn as follows. On the x-axis, time. Fix a specific \( t \) and consider the corresponding vertical line, from 0 to 1. Such line represent the composition

---

\(^3\) Why have we not used the symbol \( r \) as in Sec.(1.2.4)? The reason is that in general the outcrossing rate \( r^* \) is not the recombination rate \( r \). The former here is treated as a probability while the latter is a rate that can take any positive value, and taking into account the discreteness of the computer simulation as for mutations\(^4\) we should have \( r^* = 1 - e^{-r} \). However, as long as \( r \ll 1 \) they approximately coincide \( r^* \sim 1 - (1 - r) = r \).
Figure 2.2: Simulation QLES, Tab. (2.1). All-time evolution of the $L$ first order cumulants $\{\chi_i\}$. In a QLE phase they fluctuate around the random-state values $\chi_i \sim 0$. 

Figure 2.3: Simulation QLES, Tab. (2.1). All-time evolution of the $L$ second order cumulants $\{\chi_{ij}\}_{i=1}^{L}$. In a QLE phase they fluctuate around the random-state values $\chi_{ij} \sim 0 \forall i \neq j$ and $\chi_{ii} = 1 - \chi_i^2 \sim 1$. 

of the population at time $t$ by staking the frequencies of every genotype, from the largest (below) to the smallest (top). In the QLE phase will observe no clone emerging from the crowd, each genotype we be typically found in a single copy, which will result in a dust-like plot like in Fig. (2.5a). Just out of curiosity, we also show in Fig. (2.5b) the same chart for a different simulation that sets the system in a CC - phase as explained in Sec. (1.6.2).

* Genotype Snapshot. We have so far described all-time evolution of observables, the last two will instead be instantaneous. It is interesting to observe directly how a sample from the population looks like at a certain time, we show a possible visualization in Fig. (2.6): each vertical strip is genotype, from a random sample of 200 individuals out of $N$; each horizontal strip corresponds by consequence to one of the $L$ loci. Y/B stand for $\pm 1$. In
Figure 2.4: Simulation QLES, Tab.(2.1). All-time evolution of the fitness mean and standard deviation. In a QLE phase, fluctuations aside, they rapidly settle to their asymptotic values.

Figure 2.5: Evolution of the clonal structure of the population for the first 100 generations. Left: Simulation QLES, Tab.(2.1). In a QLE phase, no significant clones emerge, most of the genotypes are present as single copies. Right: Parameters of the simulation: $N = 1000, L = 25, T = 2000, \mu = 0.05, r = 0.05, \omega = 0.5, f_{ij} \sim \mathcal{N}(0, \sigma_e = 0.04)$. The system is in a CC phase, where few very fit genotypes compete against each other.

the QLE phase no pattern survives the erosion due to recombination, hence we will see an (approximately) random configuration.

**Fitness Distribution.** Lastly, a crucial observable is the instantaneous distribution of the fitness in the population. In other words, we fix a generation $t$ and evaluate $F(g)$ for all genotypes $g$ present in the population at that time, each counted with its multiplicity. The resulting distribution is shown in Fig.(2.7), from our QLES. Suppose now there were no selection but we had to evaluate a function like Eq.(1.3) for each genotype. The configuration of the genotypes would be random in absence of fitness benefits/penalties. Assuming no additive components in Eq.(1.3) and $f_{ij} \sim \mathcal{N}(0, \sigma_e), F(g)$ would be the sum of $L(L - 1)/2$ i.i.d. Gaussian numbers, hence $F(g) \sim \mathcal{N}(0, \sqrt{L(L - 1)/2} \cdot \sigma_e)$. This is only
approximately the case when turning on a selection mechanism, albeit weak enough not to undermine the global behaviour. For instance, since the fit genotypes are enhanced and the unfit are penalized, we expect the mean of the resulting distribution for the $F(g)$ to be shifted towards positive values, which explains the behaviour of the mean (greater than zero) in Fig. (2.4), too.

We have dedicated this short chapter to the description of FFPopSim, the package developed by F. Zanini and R. Neher for simulations in population genetics, based on the NS theory developed in Ch. (1). Relying on simulation tools in this field is often not choice, the reason
being the obvious difficulty in collecting long-term experimental data. With a few exceptions, one will be discussed in Sec. (5.2.4), there is often no way to observe evolution under controlled (lab) conditions, which makes it tough to validate by experiments a theoretical model. Hence we resort computational methods to simulate those data and test theoretical predictions on them. The question is: can we do that?

There is an elephant in the room of this chapter and time has come to deal with it. Why should we expect a simulated evolution based on a simple algorithm to yield surprising results? And if indeed surprising, why should they be of any biological interest, since a great amount of the biological complexity has been left out from the model?

To the first question, we readily answer by saying that, within the field of complex system, it is well-known that simple recurrent rules can produce extremely reach and complicated behaviours [Flake, 1998]. To the second question, an answer is less obvious. Let us now forget the details of our specific simulation tool and look at the picture from a broader perspective: suppose that we find an unexpected behaviour in a simulation of the evolutionary process (whatever the algorithm): how should we interpret it? In a paper of utmost interest, named The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities [Lehman et al., 2020] the authors gather together a selection of stories from scientists that, at some point of their research life, found themselves to deal with such a question. As a starting point, one should ask about the very meaning of a digital evolution, quoting from there:

(...) there exist abstract principles underlying biological evolution that are independent of the physical medium, and that these principles can be effectively implemented and studied within computers. As noted by Daniel Dennett, "evolution will occur whenever and wherever three conditions are met: replication, variation (mutation), and differential fitness (competition); no particular molecule (e.g. DNA or RNA) or substrate (e.g. specific physical embodiment) is required.

In other words, if we allow our simulation the philosophical status of evolution itself rather than mere attempt (more or less fortunate) of imitating natural evolution, then we have reason to expect much more than a reproduction of the experimental data and the relevance of the results should be assessed in this broader framework. The authors provide several examples of “unexpected behaviours” in digital evolution, and group them in four possible categories, that we will keep in mind for the results that will be presented later in this work:

1. misspecified fitness functions, in which digital evolution reveals the divergence between what an experimenter is asking of evolution and what they think they are asking; 2. unintended debugging, in which digital evolution reveals and exploits previously unknown software or hardware bugs; 3. exceeded experimenter expectations, in which digital evolution discovers solutions that go beyond what an experimenter thought evolution would produce; and 4. convergence with biology, in which digital evolution discovers solutions surprisingly convergent with those found in nature, despite vast divergence in medium and conditions.
In Ch. (1), under Quasi-Linkage Equilibrium conditions, we have paved the way for the inference from data of the epistatic fitness landscape by means of Eq. (1.28). Yet this last equation does not directly relate the \( \{ f_{ij} \} \) to statistics of the data, instead we there see in the rhs the couplings \( \{ J_{ij} \} \) of the Ising-like Eq. (1.17) that, by assumption, describes the one-genome distribution.

The link we are still missing is how to infer such coupling from raw data and it turns out that such a problem is of broad interest, much beyond the specific context of the problem described above. Recently in the scientific literature a number of different techniques have been developed to solve the so-called Inverse Ising Problem (IIP), the ensemble of these methods is commonly referred as Direct Coupling Analysis (DCA). Applications have been found in fields as diverse as computational biology, medicine, ecology, finance (…)

This chapter is dedicated to the Inverse Statistical Physics. In Sec. (3.1) we define the IIP and describe its connection with thermodynamics; in Sec. (3.2 - 3.3) a description will be found of few solution strategies for the IIP, namely the Maximum Likelihood (ML), the Mean Field approximation of the Maximum Likelihood (MF) and the Pseudolikelihood Maximization (PLM); in Sec. (3.4) we apply these methods to the our problem of the inference of epistasis from evolutionary data and discuss the results; finally in Sec. (3.5) we briefly mention some applications of this discussion beyond the population genetics concerns.

We will mainly follow [Nguyen et al., 2017] for the general discussion, other useful references being [Cocco et al., 2018], which deals with the broader Inverse Potts Problem, and [Zeng & Aurell, 2020a], more application-oriented, in particular to the problem of the inference of the fitness landscape.

### 3.1 AN INVERSE PARADIGM

#### 3.1.1 Statement of the IIP

As a first step, we set the framework. Let us consider an Ising model with \( L \) binary spin variables \( s_{i} = \pm 1, \) with \( i = 1, \ldots, L \). The Hamiltonian for an Ising system reads

\[
\mathcal{H}_{J,h}(s) = - \sum_{i} h_{i} s_{i} - \sum_{i<j} J_{ij} s_{i} s_{j},
\]

We use the notation \( s \) for a realization of the spin variable and \( s_{i} \) indicates the spin random variable. When it is not confusing, the bold notation will indicate the whole set of spin variables \( s = \{ s_{i} \} \). Analogously meaning for the bold notation for other quantities, e.g. \( J, h \).
where the $J$ are the pairwise couplings between the spin variables ($J_{ii} = 0 \forall i$) and the $h$ are the local magnetic fields; they are collectively referred as the parameters of the Ising problem. Under equilibrium conditions, the probability of a configuration $s$ is the celebrated Boltzmann distribution

$$p(s) = \frac{1}{Z} e^{-\mathcal{H}_{J,h}(s)} .$$

(3.2)

Note that the inverse temperature $\beta$ has been set to 1, this implies no loss of generality since in Eq.(3.2) only the products $\beta h_i$ and $\beta J_{ij}$ appear. The normalization $Z$ is the standard partition function

$$Z(J,h) = \sum_s e^{-\mathcal{H}_{J,h}(s)} .$$

(3.3)

The general expression for the expected value of a function $Q(\sigma)$ of the spin variables is

$$\langle Q(\sigma) \rangle = \sum_s p(s)Q(s) ;$$

(3.4)

In Ch.(1) we have already assigned the notations $\chi_i = \langle \sigma_i \rangle$ for the first order cumulant and $\chi_{ij} = \langle \sigma_i \sigma_j \rangle - \chi_i \chi_j$ for the second order cumulant. In this chapter it will be more convenient to use the moments instead, the first ones being the same $\chi_i$ as the first cumulants, the second ones defined as $\phi_{ij} = \langle \sigma_i \sigma_j \rangle = \chi_{ij} + \chi_i \chi_j$, see App.(A). In the present context, $\{\chi_i\}$ are the equilibrium magnetizations while $\{\phi_{ij}\}$ are the pair correlations. In the Forward Ising Problem the parameters $J,h$ of the Boltzmann distribution Eq.(3.2) are known and the job is to compute statistical observables e.g. $\chi_i, \phi_{ij}$. In an Inverse Ising Problem (IIP) the paradigm is reversed, let us emphasize it.

**Statement of the IIP**

Let $D = \{s^m\}$, $m = 1, \ldots , M$ be a set of samples of spin configurations for an Ising system whose parameters are unknown. The goal of the IIP is to infer such parameters from the available data $D$ i.e. from measurements of the observables. The equilibrium IIP is to do that when the spin configurations in $D$ are independent samples from the equilibrium Boltzmann distribution Eq.(3.2).\(^3\)

In this work we will only deal with the equilibrium reconstruction, therefore from now on such specification will be implied when referring to the IIP. We underline that even in the equilibrium case, the task is not at all straightforward: we are trying to reverse the map between micro and macroscopic laws as provided by the Statistical Mechanics machinery. No need to say, we do not expect to reconstruct the parameters exactly with a finite amount of data, however we can reasonably ask our predictions to be more accurate for larger dataset and exact in the limit of infinite data.

In Sec.(3.5) some applications beyond the strict interest of this Thesis are presented, many have

---

\(^2\) In the context of computational biology, such set $D$ is often encoded in the MSA i.e. the Multiple Sequence Alignment matrix, whose rows are samples and column are loci. See Ch.(6) for more on this.

\(^3\) For a non-equilibrium system, where the transition rates between states do not satisfy the detailed balance condition, the steady state cannot be described by the Boltzmann distribution with a known Hamiltonian. All the methods that we will develop in this chapter fail in the non-equilibrium case, were the IIP is a much tougher task, see [Nguyen et al., 2017; Zeng & Aurell, 2020a].
appeared in the recent years and many more are likely to emerge in the next ones. Indeed, the crucial factor for the settlement of an IIP is data availability: very roughly, every many-body system accessible for measurements at microscopic scales may be suitable for a IIP approach as soon as enough data are available.

3.1.2 Thermodynamics of the IIP

Since the very beginning, statistical physics has exploited the notion of thermodynamic potentials, extremely useful in characterizing the thermodynamic state, simplifying calculations and providing a bridge with Statistical Mechanics [Callen, 1985]. For the canonical ensemble, the exact point where such bridge appears is between the two sides of the following identity:

$$F(J,h) = - \log Z(J,h), \quad (3.5)$$

where the Helmholtz free energy $F(J,h)$ is recognized to be minus the logarithm of the partition function. In the forward problem, first and second order moments can be evaluated by means of simple derivatives of Eq. (3.5):

$$\chi_i = - \frac{\partial F}{\partial h_i}(J,h); \quad \phi_{ij} = - \frac{\partial F}{\partial J_{ij}}(J,h) = \frac{\partial^2 F}{\partial h_i \partial h_j}(J,h) - \frac{\partial^2 F}{\partial h_i \partial h_j}(J,h) \quad (3.6)$$

The IIP exchanges the roles of the parameters $J, h$ and the observables $\phi, \chi$: the latter are now fixed, the former to be computed. We seek accordingly a thermodynamic potential which is function of the observables and from which we can compute the parameters by means of simple derivatives, similarly to what is done in Eq. (3.6) for the forward problem. We get such potential by operating a Legendre transform of the Helmholtz free energy with respect to both couplings and fields:

$$S(\phi, \chi) = \min_{J,h} \left[ - \sum_i h_i \chi_i - \sum_{i<j} J_{ij} \phi_{ij} - F(J,h) \right], \quad (3.7)$$

which can be recognized to be the Shannon entropy function for the distribution Eq. (3.2), see App. (D). To see how the parameters of the Ising model follow from the latter, we perform the inverse transformation (i.e. again a Legendre transform, which is its own inverse) to write

$$F(J,h) = \min_{\phi, \chi} \left[ - \sum_i h_i \chi_i - \sum_{i<j} J_{ij} \phi_{ij} - S(\phi, \chi) \right]; \quad (3.8)$$

setting the derivatives of the term in the square brackets to zero one gets

$$J_{ij} = \frac{\partial S}{\partial \phi_{ij}}(\phi, \chi), \quad h_i = \frac{\partial S}{\partial \chi_i}(\phi, \chi). \quad (3.9)$$
However, from Eq. (3.6) we see that the derivatives of the Helmholtz free energy with respect to $J$ can be expressed in terms of those with respect to $h$, therefore for our purposes it is sufficient to operate a Legendre transform of the latter set. We define accordingly the Gibbs free energy as

$$G(J, \chi) = \max_h \left( \sum_i h_i \chi_i + F(J, h) \right)$$

(3.10)

i.e. $G(J, \chi)$ is defined to be minus the Legendre transform of $F(J, h)$ with respect to $h$, in view of the relation $\min f = -\max(-f)$. The magnetic fields can readily be computed with

$$h_i = \frac{\partial G}{\partial \chi_i}(J, \chi).$$

(3.11)

A great interest is in the second derivative of $G$ with respect to the magnetizations

$$\frac{\partial^2 G}{\partial \chi_i \partial \chi_j}(J, \chi) = \frac{\partial h_i}{\partial \chi_j}(J, \chi) = (\chi^{-1})_{ij},$$

(3.12)

where in $(a)$ we have used the inverse function theorem $\left[ \partial h / \partial \chi \right]_{ij} = \left[ (\partial \chi / \partial h)^{-1} \right]_{ij}$ and the linear response theory to relate the second order cumulant $(\chi)_{ij} = \chi_{ij} = \partial \chi_i / \partial h_i(J, h)$ to the susceptibility of the magnetization to a variation in the magnetic field. Eq. (3.12) turns out to be extremely useful: the rhs can be computed from data and, if $G$ is known, the corresponding system of equations can be solved yielding the couplings $J$. Differently from the Helmholtz free energy, a number of approximations exist for $G$, one of them will be discussed in Sec. (3.3.1).

Many of these methods are based on a variational principle; for the case of the Gibbs free energy it reads

$$G(J, \chi) = \max_h \left\{ \sum_i h_i \chi_i + \min_q \left\{ U[q] - S[q] \right\} \right\},$$

(3.13)

which comes from Eq. (3.10) and from the variational principle for the Helmholtz free energy

$$F(J, h) = \min_q \{ U[q] - S[q] \} = \min_q F[q]:$$

(3.14)

here $q$ is a probability distribution for the spin configurations, $U[q] = \langle \mathcal{H} \rangle_q$ and $S[q] = -\langle \log q \rangle_q$. Eq. (3.14) in turn stems from asking the Kullback-Leibler distance $D_{KL}(q||p)$ between the trial distribution $q$ and the Boltzmann distribution $p$ to be minimal, App. (D). The usefulness of this approach is lies in the fact that we may want to put constraints on $q$ i.e. to focus on a particular family of trial distributions $q$, hence obtaining an upper bound for $F(J, h)$.

As a final remark, we observe that when $q$ is taken from the set $\mathcal{G}$ of distributions for which $\langle \sigma_i \rangle_q = \chi_i$, the double extremum problem Eq. (3.13) is equivalent to the single conditional minimization

$$G(J, \chi) = \min_{q \in \mathcal{G}} \left\{ -\sum_{i < j} J_{ij} \langle \sigma_i \sigma_j \rangle_q - S[q] \right\}.$$  

(3.15)
3.2 Maximum Likelihood

We now go back to the central problem of the inference of the Ising parameters from data. From a theoretical point of view, a straightforward solution is provided by the Maximum Likelihood (ML) estimation, that we now describe and specify in the Ising case.

Let us consider the inference problem in a general framework: suppose a set of observations $x_1, \ldots, x_M$ drawn from a probability distribution $p(x_1, \ldots, x_M|\theta)$ with unknown parameter $\theta$. According to the Maximum Likelihood prescription, the best estimate $\theta^{ML}$ based on the available data is simply given by

$$\theta^{ML} = \arg \max_\theta p(x_1, \ldots, x_M|\theta).$$

(3.16)

This estimator is interesting in many respects: it converges in probability to the true parameter $\theta$ (consistency) and satisfies the Cramér - Rao lower bound for large samples i.e. there is no consistent estimator with smaller mean square error [Cover & Thomas, 2006; Rossi, 2018]. Eq.(3.16) can be understood in view of the notorious Bayes theorem: let $p(\theta)$ be the prior information on the parameter $\theta$, then the posterior distribution taking into account the additional information provided by the samples $x_1, \ldots, x_M$ is given by

$$p(\theta|x_1, \ldots, x_M) = \frac{p(x_1, \ldots, x_M|\theta)p(\theta)}{p(x_1, \ldots, x_M)}.$$  

(3.17)

If no prior knowledge is available, $p(\theta)$ is uniformly distributed over the space of possible parameters (each one is equally likely). As a result, in Eq.(3.17) the posterior distribution is proportional to the Likelihood function $p(x_1, \ldots, x_M|\theta)$ and it is maximized by Eq.(3.16). To avoid dealing with small numbers, it is common practice to maximize the logarithm of the likelihood, which has no consequences on the result: since the logarithm is a strictly monotonic function, the arg max is preserved. The log-likelihood per sample

$$\mathcal{L}_D(J, h) = \frac{1}{M} \log p(D|J, h)$$

(3.18)

is easily evaluated for the Ising parameters $J, h$ upon measurement of $M$ independent samples $D = \{s^m\}$ of the Boltzmann distribution Eq.(3.2):

$$\mathcal{L}_D(J, h) = \sum_i h_i \frac{1}{M} \sum_m s_i^m + \sum_{i<j} J_{ij} \frac{1}{M} \sum_m s_i^m s_j^m - \log Z(J, h)$$

$$= \sum_i h_i \langle \sigma_i \rangle^D + \sum_{i<j} J_{ij} \langle \sigma_i \sigma_j \rangle^D - \log Z(J, h),$$

(3.19)

where we $\langle Q \rangle^D = \frac{1}{M} \sum_m Q(s^m)$ is the sample average of the function $Q(s^m)$ of the spin variables $s^m$. We immediately observe that, up to a sign, the maximum log-likelihood is the entropy Eq.(3.7). It is worth stressing that nothing but the sample averages $\langle \sigma_i \rangle^D, \langle \sigma_i \sigma_j \rangle^D$ is needed
from data to determine Eq. (3.19) i.e. they are a sufficient statistics to determine the Ising parameters. It can be proved that maximizing the log-likelihood entails the minimization of the Kullback-Leibler distance $D_{KL}(p^D||p)$ between the Boltzmann distribution $p$ as in Eq. (3.2) and the empirical distribution $p^D = \frac{1}{M} \sum_m \delta_{s_m,s}$. Moreover, the log-likelihood is a strict concave function of the parameters and its maximum is unique.

According to the maximum likelihood receipt, the best estimate of the Ising parameters is

$$\{J^{ML}, h^{ML}\} = \text{arg max}_{J,h} \mathcal{L}_D(J,h).$$

(3.20)

A straightforward approach would be to implement a gradient-descent algorithm (Boltzmann machine learning): if $\eta$ is the learning rate parameter, one updates the parameters according to

$$h_i^{n+1} = h_i^n + \eta \frac{\partial \mathcal{L}_D}{\partial h_i}(J^n,h^n), \quad J_{ij}^{n+1} = J_{ij}^n + \eta \frac{\partial \mathcal{L}_D}{\partial J_{ij}}(J^n,h^n).$$

(3.21)

This procedure ends when the log-likelihood is maximized, explicitly:

$$0 = \frac{\partial \mathcal{L}_D}{\partial h_i}(J^n,h^n) = \langle \sigma_i \rangle^D - \langle \sigma_i \rangle,$$

(3.22)

$$0 = \frac{\partial \mathcal{L}_D}{\partial J_{ij}}(J^n,h^n) = \langle \sigma_i \sigma_j \rangle^D - \langle \sigma_i \sigma_j \rangle.$$

(3.23)

We conclude that the estimate in Eq. (3.20) is reached when the sample averages match those under the Boltzmann distribution.

From a theoretical point of view, no more effort would be needed, since the Boltzmann machine learning is the IIP solution we are looking for. There is a serious issue with Eq. (3.21), though: at each time step the averages $\langle \sigma_i \rangle, \langle \sigma_i \sigma_j \rangle$ must be computed, which means performing thermal averages over $2^N$ states. Evaluating the partition function $Z$ is an enormous computational burden, too. In practice, this is never feasible except for very small system sizes, which motivates the effort of developing approximate solutions to the IIP.

Before that, let us add a final remark on Eq. (3.17): what if we have some prior information on the parameters? For instance, one may want to use:

$$p_{J,h} = e^{-\lambda_1 \sum_i h_i^2 - \lambda_2 \sum_{i<j} J_{ij}^2} \Rightarrow \log p(J,h|D) = \mathcal{L}_D(J,h) - \lambda_1 \sum_i h_i^2 - \lambda_2 \sum_{i<j} J_{ij}^2,$$

(3.24)

which is called $L_2$ - regularization and assigns small probabilities to large fields and couplings. Several other choices are possible as regularization terms (and correspondent priors), they are useful in coping with overfitting and undersampling issues. Maximizing the posterior in Eq. (3.24) does not minimize anymore the Kullback-Leibler distance with the empirical distri-
bution, but it additionally avoids large couplings and fields. The values of $\lambda_1, \lambda_2$ are typically validated beforehand on an initial (smaller) dataset.

3.3 APPROXIMATE SOLUTIONS FOR THE IIP

The literature on the IIP has developed in recent times several approximation schemes, some of which exploit and borrow ideas from physics and statistics, others developed from scratch. Which one is the best to employ often depends on the specific problem e.g. on the topology of the network of couplings to be inferred (if known), on the size of the dataset (...) We here describe two such schemes and refer to [Nguyen et al., 2017] for an account of all the most relevant ones.

3.3.1 Mean Field Approximation

As is customary in statistical physics, a first straightforward approximation for the problem is the mean-field theory (MF). We assume that the Boltzmann distribution factorizes in the sites

$$p^{MF}(s) = \prod_i \frac{1 + \tilde{\chi}_i s_i}{2},$$

where different spins are independent variables and the effective magnetization $\tilde{\chi}_i$ results from both the local field $h_i$ and from the couplings $J_{ij}$ with all other spins. We now resume the variational expression of the Gibbs free energy for the IIP, Eq.(3.15): we recall that the set $\mathcal{G}$ of distributions $q$ over which the minimization is operated is that for which $\langle \sigma_i \rangle_q = \chi_i$ which in the MF case reads $\tilde{\chi}_i = \chi_i$. The minimization is trivial, since there is only one mean-field distribution Eq.(3.25) in $\mathcal{G}$, therefore we write:

$$\mathcal{G}^{MF}(\mathbf{J}, \chi) = - \sum_{i<j} J_{ij} \chi_i \chi_j + \sum_i \left[ \frac{1 + \chi_i}{2} \log \frac{1 + \chi_i}{2} + \frac{1 - \chi_i}{2} \log \frac{1 - \chi_i}{2} \right].$$

This is the approximation we needed in Eq.(3.12) to solve for the couplings $J$, indeed, for $i \neq j$,

$$J_{ij}^{ MF} = -(\chi^{-1})_{ij} ;$$

we remember that the element of the matrix $\chi$ are the second order cumulants $\chi_{ij}$. We can also get the fields $h_i$ from Eq.(3.11):

$$h_i^{ MF} = - \sum_{j \neq i} J_{ij}^{ MF} \chi_j + \frac{1}{2} \frac{1 + \chi_i}{1 - \chi_i} = - \sum_{j \neq i} J_{ij}^{ MF} \chi_j + \text{arctanh} \chi_i .$$

Eq.(3.27 - 3.28) are the mean-field guesses for the Ising parameters. The most appealing feature of this method is its computational simplicity, since parameters follow from simple combina-
tions of observable statistics \( \{ \chi_i, \chi_{ij} \} \). Nevertheless, the MF approximation Eq.(3.25) is an uncontrolled ansatz for the Boltzmann distribution, therefore we have no analytical control on the accuracy and we can but validate the method with numerical tests. Despite this, the mean-field scheme can be improved; as an example, the Thouless-Anderson-Palmer (TAP) reconstruction adds a “second order” term in \( J_{ij} \) to the Gibbs free energy. Remarkably, there exist also a systematic and controlled expansion (Plefka expansion) of the Gibbs free energy of the Sherrington-Kirkpatrick (SK) model\(^5\) which gives at the first order the MF result, at the second order the TAP formulas and allows to evaluate higher orders, too.

3.3.2 Pseudolikelihood Maximization

The second (and last) method that we discuss is the Pseudo-Likelihood Maximization (PLM). The starting point this time is not a thermodynamic potential, instead we directly exploit the structure of the Ising Hamiltonian \( \mathcal{H} \) Eq.(3.1) as follows. Let us consider a particular spin variable \( \sigma_i \) and distinguish the part \( \mathcal{H}_i \) of the Hamiltonian that depends on \( \sigma_i \) from the rest, that we collectively indicate with \( \mathcal{H}_\setminus i \):

\[
\mathcal{H}(s) = \mathcal{H}_i + \mathcal{H}_\setminus i = -h_is_i - \sum_{j \neq i} J_{ij} s_is_j + \mathcal{H}_i(s_i).
\]  

(3.29)

We can sum explicitly the terms related to \( \sigma_i \) in the partition function Eq.(3.3)

\[
\mathcal{Z}(J, h) = \sum_{s_i} 2 \cosh \left( h_i + \sum_j J_{ij} s_j \right) e^{-\mathcal{H}_\setminus i(s_i)},
\]  

(3.30)

and from Eq.(3.6) we get the first and second moments (one and two spin expectations) involving \( \sigma_i \), by simple derivatives with respect to the parameters:

\[
\langle \sigma_i \rangle = \left\langle \tanh \left( h_i + \sum_{j \neq i} J_{ij} s_j \right) \right\rangle,
\]  

(3.31)

\[
\langle \sigma_i \sigma_j \rangle = \left\langle \sigma_i \tanh \left( h_i + \sum_{k \neq i} J_{ik} s_k \right) \right\rangle.
\]  

(3.32)

The average in the rhs of the last equations is over the entire Boltzmann distribution Eq.(3.2) and these are still exact equations. The approximation is implemented in the next step: we

---

\(^5\)The celebrated SK-model [Sherrington & Kirkpatrick, 1975] assumes a simple Ising-like Hamiltonian Eq.(3.1) where all the fields are set to zero \( h_i = 0 \)\( \forall i \) while the couplings are drawn from a Gaussian distribution \( J_{ij} \sim N(J_0, \sigma_J)/\sqrt{L} \). It is a widely studied prototype of spin-glass system for which an equilibrium solution exist [Castellani & Cavagna, 2005].
substitute the (computationally) prohibitive averages in Eq. (3.31 - 3.32) with sample averages, labelled with the superscript \( D \).

\[
\langle \sigma_i \rangle^D = \left\langle \tanh \left( h_i^{PL} + \sum_{j \neq i} J_{ij}^{PL} s_j \right) \right\rangle^D , \tag{3.33}
\]

\[
\langle \sigma_i \sigma_j \rangle^D = \left\langle \sigma_i \tanh \left( h_i^{PL} + \sum_{k \neq i} J_{ik}^{PL} s_k \right) \right\rangle^D . \tag{3.34}
\]

The simplification is self-evident: Eq. (3.33 - 3.34) are a system of non-linear equations in the \( L \) variables \( h_i^{PL}, \{ J_{ij}^{PL} \}_{i \neq j} \) that can be solved with standard methods. The net effect of our assumption hence is to split the problem of estimating \( L^2 \) parameters (fields and couplings) in \( L \) separate problems, each of them involving only \( L \) parameters.

A different but equivalent interpretation of this approximation scheme is the following: let \( \sigma_i \) be a spin variable, we consider the conditional probability of \( s_i \) under the observation of the other variables \( s_{\bar{i}} \)

\[
p(s_i | s_{\bar{i}}) = \frac{1}{1 + e^{-2h_i (s_i + \sum_{j \neq i} J_{ij} s_j)}} = \frac{1}{2} \left[ 1 + s_i \tanh \left( h_i + \sum_{j \neq i} J_{ij} s_j \right) \right] \tag{3.35}
\]

and depends only on the field \( h_i \) and on the couplings \( \{ J_{ij} \}_{j \neq i} \). The last quantities also appear in the log-likelihood per sample \( \mathcal{L}_D^i \) for the last distribution of probability, which reads

\[
\mathcal{L}_D^i (J_{ij}, h_i) = \frac{1}{M} \sum_m \log \frac{1}{2} \left[ 1 + s_i^m \tanh \left( h_i + \sum_{j \neq i} J_{ij} s_j^m \right) \right] , \tag{3.36}
\]

cf Eq. (3.48), we used the notation \( J_{ik} = \{ J_{ij} \}_{j \neq i} \). Eq. (3.33 - 3.34) simply follow from setting the derivatives of \( \mathcal{L}_D^i \) with respect to \( h_i \) and \( \{ J_{ij} \}_{j \neq i} \) to zero, hence they are found to maximize the log-likelihood. Altogether, one might define the so-called pseudolikelihood as

\[
\mathcal{L}^{PL}(J, h) = \sum_i \mathcal{L}_D^i (J_{ij}, h_i) \tag{3.37}
\]

whose maximization yields the whole set of Ising parameters. Note that in general for the inferred couplings \( J_{ij} \neq J_{ji} \), even if the underlying model has symmetric couplings, due to sampling noise; in the latter case, a practical solution is to use the average \( \frac{1}{2} (J_{ij} + J_{ji}) \).

The PLM method is interesting in many respects. In the first place, it is consistent, the log-likelihood Eq. (3.37) has the same maximum as Eq. (3.19) in the limit of infinite data \( M \to \infty \), as it is evident from the fact that in the same limit the Eq. (3.33 - 3.34) coincide with Eq. (3.31 - 3.32). Moreover, the \( \mathcal{L}_D^i \) can be maximized independently allowing for a parallel implementation.

As a final remark, we underline that the computational complexity of this method scales polynomially both with the system size \( L \) and with the sample size \( M \), and it is usually slower than an MF scheme but much faster than the exact maximization of the likelihood.
### 3.3.3 MF vs PLM

We have not yet discussed the accuracy of the methods discussed in the previous sections. A standard procedure to test an approximate solution of the IIP (especially for uncontrolled ones) is to simulate data from an Ising model with known fields and couplings, then compare the results of the inference with the input values of the parameters. As mentioned above, the reconstruction errors depend on many factors (network topology of the couplings, ergodicity of the system...), we here focus on the dependence on the size \( M \) of the dataset \( D \) and on the coupling strength, by re-establishing and tuning the inverse temperature \( \beta \) in Eq. (3.2). The high temperature limit \( \beta \to 0 \) hence corresponds to that of low couplings/fields and vice versa. In Fig. (3.1) the performances of several inverse techniques are shown. The underlying data are generated from a Sherrington-Kirkpatrick model where \( J_{ij} \sim \mathcal{N}(0, \beta/\sqrt{L}) \) and a fully connected graph of interactions is assumed. For each couple \((\beta, M)\), \(10^4 L\) Monte Carlo steps with Metropolis transition rule are used to reach an equilibrium state and collect the samples. Let \( J_{ij}^0 \) be the input parameters of the simulation, the reconstruction error \( \gamma_J \) can be quantified as follows:

\[
\gamma_J = \sqrt{\frac{\sum_{i<j} (J_{ij}^0 - J_{ij})^2}{\sum_{i<j} (J_{ij}^0)^2}}.
\]

Equation (3.38)

Overall, we see that the PLM approximation yields a more accurate reconstruction of the model parameters with respect to the MF theory; the latter is indeed known to overestimate large couplings.

In the top panel, we see that all methods equally fail for \( \beta \to 0 \), since for too low coupling strengths the reconstruction is democratically hampered by sampling noise. On the other hand, for sufficiently high \( \beta \) the approximations on which our methods are based break down and the errors blow up, a possible cause being the ergodicity breaking for strong coupling. As for the dependence on \( M \), differently from the MF, in a PLM algorithm the reconstruction error can always be com-

---

Figure 3.1: Reconstruction of a fully connected Ising model. Both panels show the reconstruction error as in Eq. (3.38), in the upper one \( \gamma_J(\beta) \) for \( M = 15000 \), in the lower one \( \gamma_J(M) \) for different values of \( \beta \). \( L = 64 \). Results are shown for the following approximations: Mean Field (MF), TAP reconstruction (TAP), Bethe-Peierls method (BP), Sessak-Monasson method (SMA), Adaptive Cluster Expansion (ACE), Pseudolikelihood Maximization (PLH). From [2017].
pensated by a larger dataset, which leads to the polynomial behaviour shown in the lower panel for all the tested values of $\beta$.

What if the true couplings are not known? This is usually the case with real data. One possibility might be to try different methods and compare the likelihood of the resulting parameters, the better techniques resulting in higher likelihoods. Alternatively one can compare the statistics resulting from data generated from reconstructed parameters with those observed.

## 3.4 Inference in the NS Theory

We are now in the position to test Eq. (1.28), since the missing link between raw data and the couplings $J$ has been provided by the methods above. This is exactly what has been done in [Zeng & Aurell, 2020b], the content of which we now illustrate.

The testing strategy is very similar to the one employed in Sec. (3.3) and it is based on the following three steps: simulating evolutionary data by means of FFPopSim, inferring couplings by means of MF/PLM, finally comparing the epistatic parameters as they result from Eq. (1.28) with the input ones. Let us expand on these points.

1. **Simulating data.** We exploit the our simulation tool FFPopSim to generate evolutionary data, as explained in Ch. (2). Since we are interested here in testing the QLE regime, we set the initial parameters accordingly. A crucial choice is that of the fitness landscape, we here consider a Sherrington-Kirkpatrick fitness function, so we set $f_i = 0 \forall i$ and $f_{ij} \sim \mathcal{N}(0, \sigma_e)$. We will mainly focus on the dependence of our results on the parameters $r, \mu, \sigma_e$, a comprehensive summary of the initial parameters for the simulations is shown in Tab. (3.1), we will refer to this set as the Neher-Shraiman Test (NST).

<table>
<thead>
<tr>
<th>FFPopSim</th>
<th>NST</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>200</td>
<td>carrying capacity</td>
</tr>
<tr>
<td>L</td>
<td>25</td>
<td>n. of loci</td>
</tr>
<tr>
<td>T</td>
<td>2500</td>
<td>n. of generations</td>
</tr>
<tr>
<td>$\omega$</td>
<td>0.5</td>
<td>crossover rate</td>
</tr>
<tr>
<td>$r$</td>
<td>[0.0, 1.0]</td>
<td>outcrossing rate</td>
</tr>
<tr>
<td>$\mu$</td>
<td>[0.005, 0.1]</td>
<td>mutation rate</td>
</tr>
<tr>
<td>$\sigma_e$</td>
<td>[0.001, 0.02]</td>
<td>$f_{ij} \sim \mathcal{N}(0, \sigma_e)$</td>
</tr>
</tbody>
</table>

Table 3.1: Parameters on the NST. In light gray the parameters that are varied, the range indicated in the square brackets. SK fitness function. Random initial conditions.

2. **Inferring couplings.** Raw data from the previous point are a sequence of states of the population at each time $t$. From such data, we can compute empirical population averages to feed one or more of the inverse techniques discussed in this chapter, in particular MF (sometimes also referred as naive mean field nMF) and PLM. Due to random drift, the empirical quantities (averages over the population) fluctuate in time, in order to smooth
fluctuations out we compute the averages not only on the final state of the population but on the whole time series: if a single state is a matrix $N \times L$, the whole time series is a matrix $(N \cdot T) \times L$; as an example, $\langle s_i \rangle = \frac{1}{NT} \sum_{j=1}^{NT} s_i(j)$, where $j$ is the row index. The MF and PLM inference on data obtained from the whole time series are referred as all-time-MF and all-time-PLM, this specification will be always implied hereinafter.

3. Testing epistasis. The couplings resulting from the inverse methods enable us to reconstruct the epistasis by means of Eq. (1.28): $r$ is the known (input) recombination rate, $c_{ij}$ in [2020b] is computed as in Eq. (1.16) with $\rho = \omega = 0.5$. Let $f_{ij}^*$ be the inferred value of the epistatic fitness component between the loci $i, j$, similarly to Eq. (3.38), we quantify the reconstruction error $\varepsilon$ by

$$
\varepsilon = \sqrt{\frac{\sum_{i<j} (f_{ij}^* - f_{ij})^2}{\sum_{i<j} f_{ij}^2}}. \quad (3.39)
$$

Obviously, we want $\varepsilon$ to be as small as possible, a wrong functional dependence of $f_{ij}^*$ on the model parameters will yield $\varepsilon \gg 0$.

In Fig. (3.2) we see a typical outcome of the algorithm described just now, for the case when the inference procedure is reasonably accurate. The input and reconstructed epistatic fitness components are compared by means of a scatter plot; in the optimal case we would see the points along the diagonal line indicated, this is usually the case for the ML reconstruction, in the few cases when it is feasible. We also observe no relevant difference between MF and PLM in the final result of the inference procedure.

We can repeat the same steps for different values of the parameters and see if/how the performance of the NS inference of epistasis changes. This is done in Fig. (3.3), where the parameter space is explored in the directions $r - \mu, r - \sigma_e$. Here follows a list of conclusions that we can draw from these charts:

* $\mu$. The NS fitness reconstruction fails for low mutation rates. In fact, when mutation is insufficient, the structure of the population is essentially frozen and even if a defined
Reconstruction error $\epsilon$ as function of the parameters $r - \mu$, here $\sigma_e = 0.004$.

Reconstruction error $\epsilon$ as function of the parameters $\sigma_e - \mu$, here $\mu = 0.05$.

Figure 3.3: Phase diagrams for the reconstruction of the epistatic fitness components, from Eq. (1.28). NST simulations, as in Tab. (3.1). In the left column, alltime-MF is used to infer the couplings $J_{ij}$ from data, in the right column alltime-PLM. The magnitude of the reconstruction error Eq. (3.39) is encoded in the colours of the heat-map.

Fitness structure is present, for finite $N, T$ it is not reflected in data. On the other hand, we do not expect Eq. (1.28) to work for high $\mu$, too, since in deriving it we assumed $\mu \sim 0$. We will discuss this case in Ch. (4).

- **High $\sigma_e / r$.** The reconstruction error $\epsilon$ blows up for sufficiently low values of $r$, which can be easily understood in view of Eq. (1.28) i.e. $f^{*}_{ij} = J^{*}_{ij} \cdot r c_{ij} \sim 0$ which implies $\epsilon \sim 1$ regardless of the inference method employed. The case of too high values of $\sigma_e$ will be discussed in Ch. (5).

- **Low $\sigma_e / r$.** Results also are worse for sufficiently high recombination rates $r$ : this is due to the fact that the higher the reshuffling the smaller the couplings $J_{ij}$ inferred from data,
which become subject more and more to small-sample noise. For the same reason, we see worse results for sufficiently small $\sigma_e$.

Further plots and discussions can be found in [2020b], the ones presented above will be the starting point for our next chapters.

### 3.5 Datacentrism

Let us briefly sum up the content of our discussion in this chapter. The starting point has been the need to extract the couplings of a Ising-Boltzmann probability distribution from raw data, so to exploit Eq.(1.28). The Inverse Ising Problem (IIP) is indeed of greater interest, much beyond the context of the inference of a fitness landscape. After stating it properly in Sec.(3.1), we discussed an exact but often computationally unfeasible solution in Sec.(3.2) based on the maximization of the likelihood. This motivated us in Sec.(3.3) to look for approximate methods for the inference of Ising parameters, in particular Mean Field (MF) and Pseudolikelihood Maximization (PLM). As a first concrete application we have finally tested the NS predictions for the inference of the fitness landscape in a QLE state, Sec.(3.4)

There is no coincidence in the fact that most of the methods discussed here have arisen only in the last 20 years, together with an increasing interest in the IIP. All the frameworks in which it has emerged do share a common fundamental structure, which is the interaction of microscopic degrees of freedom that result in macroscopic observable behaviours.

One then may use the Occam’s razor$^6$: the simplest description of these degrees of freedom is clearly in terms of binary spin variables and which is the simplest non-trivial model that takes account of interactions between these spin? Clearly, an Ising model with pairwise couplings. This does not mean that such a description is always appropriate e.g. in many cases higher order interactions are not negligible, yet this has turned out to be a reasonably good approximation in several contexts.

Beside a suitable structure of the system and the Occam’s razor, the another crucial ingredient is data availability. In the recent years science has been flooded with data, new niches of complexity have become accessible to the scientific eye and a number of new questions have emerged in this datacentric world; the IIP is a prominent example: inspired, based on and shaped on data availability.

In order to be more concrete and to illustrate the power of these techniques, we conclude this chapter with a brief list of some of the systems in which the IIP and close relatives have been leveraged to infer information, apart from the one discussed in Sec.(3.4). As usual, for more info we refer to the reviews indicated at the beginning of this discussion and to earlier works there cited.

---

$^6$ There is no room here for a digression on the Occam’s razor (*lex parsimoniae*) but the topic is indeed of great interest, see for instance [https://plato.stanford.edu/entries/simplicity/](https://plato.stanford.edu/entries/simplicity/). The mantra is always the same: *Everything should be made as simple as possible, but not simpler* (often ascribed to A. Einstein, but the attribution is controversial).
**Protein Structure.** The goal is to get information about the 3D structure of a protein and the idea is to use co-evolutionary information. The input data are amino acid sequences of such protein as they appear in different species: although they may be different, the structure of the final protein must be the same, because it determines the physical and chemical properties as well as the interactions with other molecules in a cell. In an amino acid chain, each locus can be in one of the \( q = 21 \) possible states (20 amino acids and the gap state) and the statistical model with pairwise interactions is the Potts model with Hamiltonian:

\[
\mathcal{H} = -\sum_i h_i(s_i) - \sum_{i<j} J_{ij}(s_i, s_j) \tag{3.40}
\]

The goal of the inverse problem is to infer \( h_i(a), J_{ij}(a, b) \forall i, j, a, b \) where \( a, b \in 1, \ldots, q \), see also Ch.(6). The resulting coupling between the loci \( i, j \) is quantified by the Frobenius norm, also referred as Score \( S \):

\[
S_{ij} = \sum_{s_i, s_j} J_{ij}^2(s_i, s_j) \tag{3.41}
\]

which is interpreted as a measure of proximity: the higher \( S_{ij} \), the spatially closer loci \( i, j \) will be. The reason why we resort couplings of a Potts model instead of simple correlations to infer the so-called contact map is that the latter are transitive: if \( A \) is correlated with \( B \) and \( B \) is correlated with \( C \), we will find \( A \) correlated with \( C \), see Fig.(3.4). For a comprehensive discussion of this problem, we refer to [Cocco et al., 2018].

**Neural Firing Patterns.** The goal is to gain information about the interactions between a large number of neurons that leads to the emergence of a brain function. Such interactions are in the form of electrical pulses termed spikes: a neuron is emitting spikes at high rate is said to be active, at low rates it is silent. Now suppose \( N \) neurons and fix a \( \Delta t \) (typically in the order of ms): if the \( i \)-th neuron spikes in \( \Delta t \) we write \( s_i = +1 \), in the opposite case \( s_i = -1 \), so that we have a binary representation of the state of the
neural system. Repeating this observation $M$ times we have data to feed an inverse Ising method which will estimate the couplings $J_{ij}$ of the Ising distribution Eq.(3.2) between the neurons $i, j$. The Ising model for neuronal firing patterns has been proved to be a good description of the interacting system as long as $N\Delta t\nu \ll 1$, where $\nu$ is the average firing rate.

○ Financial markets. The prices of goods in the stock markets change in response to events of various nature, and are often correlated. The market is in general a highly chaotic and complex system, yet a first approach could be in terms of an Ising spin model as follows. Suppose $N$ different companies, each described by a binary spin variable which is $s_i = +1$ if at a particular time the prices for the shares of the $i$-th company goes up, $s_i = -1$ otherwise. The couplings $J_{ij}$ of the Ising model describe how the price changes in the shares $i$ affect those in shares $j$ and the analysis of the matrix $J$ could be useful in detecting clusters of companies in the market web.

Other applications include inference of ecological networks i.e. interaction between species in an environment, predictions of the effects of combinatorial antibiotic treatments, reconstruction of gene regulatory network and many more. A specific application will be discussed in Ch.(6), where we will analyze with inverse methods a topical dataset in these days i.e. that of the Sars-Cov-2.
4 | A GAUSSIAN CLOSURE FOR THE NS THEORY

Let us go back to the NS theory as encoded in Eq. (1.12, 1.15): we already know that these equations are not sufficient to draw quantitative predictions, a specific expression for $P(g)$ has to be provided. Once $P(g)$ is known, the solutions of the two equations above are in principle determined but hardly accessible from an analytical point of view; for instance, it is true that one can parametrize $P(g)$ (and its dynamics) in terms of its cumulants $\chi_{ijkl}$, but there are $2^L - 1$ of them, and in general they all appear in the RHS of Eq. (1.12, 1.15). The last ingredient that we need in order to tame these equations and make them accessible for a formal analysis is typically an additional assumption, which will hopefully open new methodological routes. In Sec. (1.3) this was done in the QLE framework $\sigma \ll r$ and an inference formula Eq. (1.28) was derived in the case where mutations are negligible $\mu \sim 0$. The methodological route aforementioned here was the perturbation expansion in the $\{J_{ij}\}$.

Considering the high number of parameters at stake, the QLE regime turns out to be a very narrow region of the parameter space, which remains mostly unexplored from a theoretical point of view. However, a relevant progress in this direction has recently appeared in [Mauri et al., 2020], where the authors suggest a new Gaussian Ansatz for $P(g)$. Albeit apparently similar to the QLE assumption of a Ising-like $P(g)$, the logic of the Gaussian Ansatz is slightly different, as we shall point out.

In Sec. (4.1), we will introduce the Gaussian Ansatz in detail; subsequently, we will exploit it to express the dynamical equations for first and second cumulants in a closed form and perturbatively derive a new formula for the inference of epistasis from data. Following [Zeng, Mauri, et al., 2020], in Sec. (4.2) we will test in silico our theoretical predictions. Finally, in Sec. (4.3) future directions of investigation are presented.

4.1 A GAUSSIAN ANSATZ

4.1.1 The Logic Of The Gaussian Ansatz

In order to understand where exactly the Gaussian Ansatz comes into play, let us move one step further and substitute in the dynamical equations for first and second order cumulants Eq. (1.12, 1.15) the explicit form of the fitness landscape Eq. (1.3), up to pairwise terms:

$$ F(g) = \sum_i f_i s_i + \sum_{i<j} f_{ij} s_i s_j , $$

(4.1)
where we set \( f_{ii} = 0 \) \( \forall i \) and \( f_{ij} = f_{ji} \) \( \forall i, j \). The resulting expressions are

\[
\dot{\chi}_i = s_i [F(g) - \langle F \rangle] - 2\mu \chi_i \\
= \sum_j f_j (s_is_j) + \sum_{j<k} f_{jk} (s_is_js_k) - \sum_j f_j \chi_i \chi_j - \sum_{j<k} f_{jk} \chi_i \chi_j - 2\mu \chi_i 
\]

(4.2)

\[
\dot{\chi}_{ij} = \langle (s_i - \chi_i)(s_j - \chi_j) \rangle [F(g) - \langle F \rangle] - (4\mu + rc_{ij}) \chi_{ij} \\
= \langle s_is_j [F(g) - \langle F \rangle] \rangle - \chi_i \chi_j + 2\chi_i \mu - \chi_j (\chi_i + 2\chi_i \mu) - (4\mu + rc_{ij}) \chi_{ij} \\
= \sum_k f_k (s_is_js_k) + \sum_{k<l} f_{kl} (s_is_js_ksl) - \sum_k f_k \chi_k (s_is_j) - \sum_{k<l} f_{kl} (s_is_j) (s_ksl) + \\
- \chi_i (\chi_j + 2\chi_i \mu) - \chi_j (\chi_i + 2\chi_i \mu) - (4\mu + rc_{ij}) \chi_{ij} 
\]

(4.3)

for the latter we have used Eq.\((1.12)\) and left implicit \( \dot{\chi}_i \).

Our goal is to express the expectations in the \( \text{RHS} \) of Eq.\((4.2, 4.3)\), in terms of the cumulants of the distribution \( P(g) \) (of all orders, in principle). We already know how to relate the 2–points expectation \( \langle s_is_j \rangle \) to the cumulants i.e. by definition of \( \chi_{ij} = \langle s_is_j \rangle - \chi_i \chi_j \). Evidently, the crucial issue is how to evaluate \( \langle s_is_js_k \rangle \) and \( \langle s_is_js_ksl \rangle \) which are respectively the 3,4–points expectation. We can slightly reduce the burden of this task by observing that, since for our Ising-alleles \( s_i^2 = 1 \), we only need to know \( \langle s_is_js_k \rangle_{i\neq j \neq k} \) \( \langle s_is_js_ksl \rangle_{i\neq j \neq k \neq l} \) no two same indices. We are now forced to choose a specific form of \( P(g) \), here comes the Gaussian Ansatz.

Let us consider a population where the mutation rate is high enough with respect to the fitness strength that no two individuals are present with the same genotype (clones). In such a system, where \( \mu \gg \sigma \), we expect correlations of order higher than two to be negligible, therefore, for the purpose of estimating 3,4–points expectations, it is appropriate to model the distribution of probability as:

\[
P(g,t) = \frac{1}{Z} \exp \left[ -\frac{1}{2} \sum_{i,j} (s_i - \chi_i)(\chi^{-1})_{ij}(s_j - \chi_j) \right], \tag{4.4}
\]

where \( Z \) is a normalization and \( \chi \) is the covariance matrix i.e. \( \chi_{ij} = \langle s_is_j \rangle - \langle s_i \rangle \langle s_j \rangle \). In words, we model \( P(g) \) as a multivariate gaussian distribution, whose cumulants (connected correlation functions) of order \( > 2 \) are exactly zero, see App.\((A)\). Let us stress that for Eq.\((4.4)\) to be valid, we are forced to allow \( s_i \in \mathbb{R} \).

Metaphorically, we are describing a population as a cloud of similar sequences distributed around the value \( \{ \chi_1, \ldots, \chi_L \} \), where \( L \) as usual is the number of loci [Mauri et al., 2020]. A first crystal-clear advantage of the Gaussian ansatz is that it will allow us to write only \( L(L+1)/2 \) dynamical equations for \( \{ \chi_i \}_i \) and \( \{ \chi_{ij} \}_{i \neq j} \) instead of \( 2^L \) for each possible \( g \).

Using Eq.\((4.4)\) and with some effort, one is able to express the 3,4–points correlations in terms of the first and second order cumulants, a proof is provided in App.\((A)\).

\[
\langle s_is_js_k \rangle_{i\neq j \neq k} = \chi_i \chi_j \chi_k + \chi_i \chi_{jk} + \chi_{ij} \chi_k + \chi_{ik} \chi_j 
\]

(4.5)
\( \langle s_j s_k s_l \rangle_{i \neq j \neq k \neq l} = \chi_i \chi_j \chi_k + \chi_i \chi_j \chi_l + \chi_i \chi_k \chi_l + \chi_j \chi_k \chi_l + \chi_i \chi_j \chi_k + \chi_i \chi_j \chi_l \)

(4.6)

Let us notice that in the rhs of Eq. (4.2, 4.3), apart from parameters of the model, only \( \{ \chi_i \}, \{ \chi_{ij} \} \)
will be found; this is the reason why the Gaussian assumption above is termed closure (GC).

The key contribution of the Gaussian Ansatz is exactly and nothing more of the content of Eq.(4.5, 4.6), and now we can exploit it to carry out the calculation of the dynamical equations.

4.1.2 Closed Equations for Cumulants Dynamics

\[ \dot{\chi}_i \quad \text{Dynamics of the first cumulants, from Eq. (4.2).} \quad \forall i \in 1, \ldots, L \]

\[ \dot{\chi}_i = \sum_j f_j \langle s_j \rangle_i + \sum_{j < k} f_{jk} \langle s_j s_k \rangle - \sum_j f_j \chi_i \chi_j - \sum_{j < k} f_{jk} \chi_i \chi_j - 2 \mu \chi_i \]

\[(a) \sum_j f_j \chi_{ij} + \sum_{j < k} f_{jk} (\chi_i \chi_k + \chi_j \chi_k) + \chi_i \chi_j + \chi_i \chi_k \]

\[= \sum_j f_j \chi_{ij} + \sum_{j < k} f_{jk} (\chi_i \chi_k + \chi_j \chi_k) - 2 \mu \chi_i \]

\[= \sum_{j < k} f_{jk} \chi_i (\chi_k + \chi_j) - 2 \mu \chi_i \]

\[= \sum_j f_j \chi_{ij} - \sum_{j < k} f_{jk} \chi_i \chi_j + \sum_{j < k} f_{jk} (\chi_i \chi_k + \chi_j \chi_k) - 2 \mu \chi_i \]

\[(b) \sum_j f_j \chi_{ij} - \sum_{j < k} f_{jk} \chi_i \chi_j + \sum_{j < k} f_j \chi_i \chi_k - 2 \mu \chi_i \]

\[= \sum_{j < k} f_{jk} \chi_i (\chi_k + \chi_j) - 2 \mu \chi_i \]

\[= \sum_j f_j \chi_{ij} - 2 \sum_{j < k} f_{jk} \chi_i \chi_j + \sum_{j < k} f_{jk} \chi_i \chi_k - 2 \mu \chi_i \]

\[(d) \sum_{j < k} f_{jk} \chi_i (\chi_k + \chi_j) - 2 \mu \chi_i \]

(4.7)

In (a) we expanded \( \langle s_j s_k \rangle \) and, after distinguishing the case where \( i \neq j \neq k \), we exploited Eq.(4.5); in (b) we used \( \chi_{ii} = \langle s_i^2 \rangle - \langle s_i \rangle^2 = 1 - \chi_i^2 \); in (c) we added and subtracted a sum; in (d) we used \( f_{ii} = 0 \ \forall i \). This first result deserves to be emphasized, \( \forall i \)

\[ \dot{\chi}_i = \sum_j f_j \chi_i + \sum_{j < k} f_{jk} \chi_k - 2 f_{ij} \chi_i - 2 \mu \chi_i . \]

(4.8)
\[ \dot{\chi}_{ij} \quad \text{Dynamics of the second cumulants, from Eq. (4.3)} \quad \forall i, j \in 1, \ldots, L \land i \neq j \]

\[
\dot{\chi}_{ij} = \sum_k f_k \langle s_is_j s_k \rangle + \sum_{k < l} f_{kl} \langle s_is_j s_k s_l \rangle - \langle s_is_j \rangle \left( \sum_k f_k \chi_k + \sum_{k < l} f_{kl} \langle s_is_k \rangle \right) + \\
- \chi_i \sum_k \chi_{ik} \left( \hat{f}_k - 2f_{ik} \chi_i \right) - \chi_j \sum_k \chi_{jk} \left( \hat{f}_k - 2f_{jk} \chi_j \right) - (4\mu + rc_{ij})\chi_{ij} \\
\] (4.9)

where we have substituted the result Eq. (4.8) and the definition \( \hat{f}_k = f_k + \sum_j f_{jk} \chi_j \). For the sake of clarity, let us analyze separately the terms highlighted in blue (B), red (R) and violet (V) and cyan (C). In order to substitute Eq. (4.5, 4.6) we again have to deconstruct the sums distinguishing cases where some of the indices are equal.

\[
V = (\chi_{ij} + \chi_{ji}) \left( \sum_k f_k \chi_k + \sum_{k < l} f_{kl} (\chi_{kl} - \chi_{lk}) \right) \\
= (\chi_{ij} + \chi_{ji}) \left( \sum_k f_k \chi_k + f_j \chi_i + f_i \chi_j - \sum_{k \neq i} f_{ik} (\chi_{ik} + \chi_{ki}) \right) + \\
+ \sum_{k \neq i,j} \left[ f_{ik} (\chi_{ik} + \chi_{ki}) + f_{jk} (\chi_{jk} + \chi_{kj}) + f_{ij} (\chi_{ij} + \chi_{ji}) \right] \\
B = \sum_{k \neq i,j} f_k \langle s_i s_j s_k \rangle + f_i \chi_j + f_i \chi_i \\
= \sum_{k \neq i,j} f_k (\chi_i \chi_j \chi_k + \chi_i \chi_j \chi_k + \chi_j \chi_i \chi_k + \chi_k \chi_i \chi_j) + f_i \chi_j + f_i \chi_i \\
R = \sum_{k < l} \sum_{k \neq i,j} f_{kl} \langle s_i s_j s_k s_l \rangle + \sum_{k \neq i,j} \left[ f_{ik} \langle s_i s_k \rangle + f_{jk} \langle s_j s_k \rangle \right] + f_{ij} \\
= \sum_{k < l} \sum_{k \neq i,j} f_{kl} (\chi_i \chi_j \chi_k \chi_l + \chi_i \chi_j \chi_k \chi_l + \chi_i \chi_j \chi_k \chi_l + \chi_i \chi_j \chi_k \chi_l + \chi_j \chi_i \chi_k \chi_l + \chi_j \chi_i \chi_k \chi_l + \chi_k \chi_j \chi_i \chi_l + \chi_k \chi_i \chi_j \chi_l) + \\
+ \chi_k \chi_j \chi_i \chi_l + \chi_i \chi_k \chi_l + \chi_k \chi_i \chi_l + \chi_k \chi_i \chi_l + \chi_k \chi_i \chi_l + \chi_k \chi_i \chi_l + \chi_k \chi_i \chi_l + \chi_k \chi_i \chi_l) + \\
+ f_{jk} (\chi_j \chi_i) + f_{ij} (\chi_{ij} + \chi_{ji}) \right] + f_{ij} \\
C = \chi_i \sum_k \chi_{jk} (f_k + \sum_l f_{kl} \chi_l - 2f_{jk} \chi_j) \\
= \chi_i \sum_k \chi_{jk} (f_k + \sum_l f_{kl} \chi_l - 2f_{jk} \chi_j) + \chi_i \chi_j \chi_i \chi_j \chi i + \chi_i \chi_i \chi_i \chi_i \chi i + \chi_i \chi_i \chi_i \chi_i \chi i \\
+ \chi_i (1 - \chi_i^2) (f_j + \sum_l f_{jl} \chi_l) \\
\]

In the last line we have used \( \chi_{ii} = 1 - \chi_i^2 \), \( f_{ii} = 0 \ \forall i \) and the definition of \( \hat{f}_i \). Moreover, let us note that there is a term in Eq. (4.9) which is nothing but (C) after exchanging \( i \leftrightarrow j \).
Summing the all the terms in Eq.(4.9) and simplifying everything possible we get:

\[
\dot{\chi}_{ij} = - (4\mu + r c_{ij}) \chi_{ij} - 2 f_j \chi_i \chi_{ij} - 2 f_j \chi_j \chi_{ij} + f_{ij} (1 - \chi_{ij}^2 - \chi_{ij}^2 - 2 \chi_i \chi_j \chi_{ij}) + \\
+ \sum_{k \neq i,j} f_k (\chi_{ik} + \chi_{jk} - \chi_{ij} \chi_{ik} + \chi_{ik} \chi_{jk} - \chi_{ij} \chi_{jk} - \chi_{ij} \chi_{k}) + \\
+ \sum_{k \neq i,j} f_k (\chi_{ik} + \chi_{jk} - \chi_{ij} \chi_{ik} + \chi_{ik} \chi_{jk} - \chi_{ij} \chi_{jk} - \chi_{ij} \chi_{k}) + \\
(\chi_{ij}^2 - \chi_{ij} - \chi_{ij}^2) \sum_k f_k \chi_i + (\chi_{ij}^2 - \chi_{ij}) \sum_l f_l \chi_l + \\
+ \sum_{k < l \neq i,j} f_k (\chi_{ik} \chi_{jl} + \chi_{jk} \chi_{il} + \chi_{ik} \chi_{jk} + \chi_{il} \chi_{jk} + \chi_{ik} \chi_{jl}) + \\
- \sum_{k \neq i,j} \chi_{ik} \chi_{jk} \sum_l \chi_{il} f_{kl} - \sum_{k \neq i,j} \chi_{ij} \chi_{kl} \sum_l \chi_{il} f_{kl},
\]

(4.10)

The latter is already the final stage of our calculation and it will be the starting point of our future calculations; nevertheless, for the sake of elegance, it is possible to rewind the sums (i.e. reversing the “decomposition” where we treat separately cases in which some indices are equal). The outcome is our second important result for this chapter: for \( i \neq j \),

\[
\dot{\chi}_{ij} = - (4\mu + r c_{ij}) \chi_{ij} - 2 \chi_j (f_i \chi_i + f_j \chi_j) + 2 f_{ij} \chi_i \chi_j (\chi_{ij} + 2 \chi_{ij}) + \\
- 2 \chi_{ij} \sum_k \left[ f_k (\chi_{ik} + \chi_{jk} \chi_{ik}) + f_{jk} (\chi_{ik} + \chi_{jk} \chi_{ik}) \right] + \sum_{k,l} f_{kl} \chi_{ik} \chi_{jl}.
\]

(4.11)

4.1.3 New Inference Formula for Epistasis

A primary interest with Eq.(4.8, 4.10) is to understand their stationary solutions, which entails solving simultaneously \( O(1^2) \) equations. However, since here we are modestly interested in the regime with high mutation and/or recombination rate, we limit ourselves to the case \((4\mu + r c_{ij}) \to \infty\). We will follow the approach reported in our recent contribution, based on a perturbative expansion of the correlations. [Zeng, Mauri, et al., 2020]

As a first step, we will investigate the further subcase where there is a purely epistatic fitness landscape and \( \chi_i = 0 \ \forall i \), therefore the results of this subsection will not hold when \( f_i \neq 0 \) or when something else causes the first order cumulants to substantially deviate from zero.

Since the first cumulants all vanish, we are only interested in Eq.(4.10) so let us rewrite it in the simpler form of current interest:

\[
\dot{\chi}_{ij} = - (4\mu + r c_{ij}) \chi_{ij} + f_{ij} (1 - \chi_{ij}^2) + \sum_{k \neq i,j} \left[ f_k (\chi_{ik} - \chi_{ij} \chi_{jk}) + f_{jk} (\chi_{ik} - \chi_{ij} \chi_{jk}) \right] + \\
+ \sum_{k < l \neq i,j} f_{kl} (\chi_{ik} \chi_{jl} + \chi_{jk} \chi_{il}).
\]

(4.12)
Let us define \( \epsilon = 1/(4\mu + r_{ij}) \to 0^+ \), to be used as small parameter for the expansion.\(^1\) The trick is now to assume:

\[
\chi_{ij} = \chi_{ij}^{(0)} + \epsilon \chi_{ij}^{(1)} + \epsilon^2 \chi_{ij}^{(2)} + \epsilon^3 \chi_{ij}^{(3)} + \mathcal{O}(\epsilon^4)
\]

and impose the stationary condition \( \dot{\chi}_{ij} = 0 \) from Eq.\((4.12)\), order by order in \( \epsilon \). The first terms are easy to work out:

\[
\mathcal{O}(\epsilon^{-1}) : \chi_{ij}^{(0)} = 0
\]

\[
\mathcal{O}(1) : \chi_{ij}^{(1)} = f_{ij}
\]

\[
\mathcal{O}(\epsilon) : \chi_{ij}^{(2)} = 2 \sum_{k \neq i,j} f_{ik} f_{jk}
\]

\[
\mathcal{O}(\epsilon^2) : \chi_{ij}^{(3)} = \sum_{k < l} \sum_{k \neq i,j} f_{kl} (f_{ik} f_{jl} + f_{jk} f_{il}) + \sum_{k \neq i,j} f_{ik} \left( 2 \sum_l f_{jl} f_{kl} - f_{ij} f_{ik} \right) + \sum_{k \neq i,j} f_{jk} \left( 2 \sum_l f_{il} f_{kl} - f_{ij} f_{jk} \right) - f_{ij}^3
\]

Therefore, up to the first order in \( \epsilon \), we can write:

\[
\chi_{ij} = \frac{f_{ij}}{4\mu + r_{ij}}
\]

\((4.14)\)

This is to be compared with Eq.\((1.28)\). In other words, thanks to the Gaussian Ansatz and leveraging a perturbative approach, we have derived a new route to infer the epistasis between loci \( i,j \) starting from correlations between them, Eq.\((4.14)\). We also note that the \( n \)-th order term is of relative size \( L \sigma(f) \) compared to the \((n-1)-\)th; we may hence expect the first order to be accurate as far as \( L \sigma(f) < 1 \).

Now that we understand the argument, let us drop the hypothesis of \( \chi_i = 0 \ \forall i \) and consider the full Eq.\((4.10)\). The perturbative evaluation is now much more demanding but assuming Eq.\((4.13)\) and following the same steps as above we get up to the first order in \( \epsilon \) :

\[
\chi_{ij} = \frac{f_{ij}}{4\mu + r_{ij}} (1 - \chi_i^2)(1 - \chi_j^2).
\]

\((4.15)\)

As it should be, we recover Eq.\((4.14)\) when setting \( \chi_i = 0 \ \forall i \). Turning around this into an inference formula we get:

\[
\chi_{ij} = \frac{f_{ij}}{4\mu + r_{ij}} \frac{4\mu + r_{ij}}{(1 - \chi_i^2)(1 - \chi_j^2)}
\]

\((4.16)\)

\(^1\) In principle, we should write \( \epsilon_{ij} \) but since we set an high crossover rate, we will have \( c_{ij} \sim 1/2 \ \forall i,j \), see Sec.\((1.2.5)\), so that we can safely assume it is indeed constant and forget about this unimportant complication.
where the star in \(f^*_{ij}\) means that these are the inferred fitness parameters. Eq.(4.16) has a self-evident advantage with respect to Eq.(1.28): there is no more need for any DCA inference, since epistasis is reconstructed directly from the population averages \(\{\chi_i\}, \{\chi_{ij}\}\). Besides the theoretical delight, this result is important also for practical reasons, since the inference procedure, even in its most streamlined versions (e.g. MF), can be very expensive in terms of computational time.

In any case, we do not know so far the performance of Eq.(4.16), we obviously hope it will work at least as well as the Eq.(1.28). Therefore, we shall delay the joy for the result to the end of the chapter, once we have tested it in the next Sec.(4.2).

### 4.2 Gaussian Closure Under Examination

Testing Eq.(4.16) versus Eq.(1.28) is precisely the goal of [2020], which will guide us throughout this section.

#### 4.2.1 Experimental Layout

We will follow a similar approach as the one presented in Sec.(3.4) based on [Zeng & Aurell, 2020b]; it is briefly summerized here together with the description of the setup.

- **Simulation Generalities.** Our simulation tool FFPopSim can be exploited to simulate a population of (on average) \(N\) individuals, each of which is a gene chain made of \(L\) loci \(s_i = \pm 1\). We let this population evolve for \(T\) generations driven by the mechanisms discussed in Sec.(1.2): mutations at rate \(\mu\), recombination at rate \(r\) and natural selection encoded in the fitness function

\[
F(g) = \sum_i f_i s_i + \sum_{i<j} f_{ij} s_i s_j.
\]

- **Simulation Settings.** Since our Gaussian Closure is expected to work also for non-zero \(\{\chi_i\}\), differently from [2020b], we assume non-zero additive components \(\{f_i\}\) of \(F(g)\). Specifically, both additive and epistatic fitness parameters will be Gaussian distributed with zero means and standard deviations \(\sigma_a, \sigma_e\), respectively: \(f_i \sim \mathcal{N}(0, \sigma_a)\), \(f_{ij} \sim \mathcal{N}(0, \sigma_e)\). In the range of parameters tested in [Zeng, Mauri, et al., 2020], there is no evidence of a dependence of the fitness effects on the specific realizations of the Gaussian distributed parameters, therefore we will safely consider \(\sigma_a, \sigma_e\) as effective coarse-grained descriptors of additive and epistatic fitness. We will run simulations for different parameters \(\mu, r, \sigma_e\), fixing all the others. The simulation ranges/values for each parameter are shown in Tab.(4.1). \(c_{ij}\) as in Eq.(1.16).

- **Probing Gaussian Closure: Algorithm.** We will use all-time data to smooth out fluctuations due to the finite size of the population, see Sec(3.4). From them, it is possible to get population-averaged quantities \(\{\chi_i\}, \{\chi_{ij}\}\) and enforce the two approaches under
<table>
<thead>
<tr>
<th>FFPopSim</th>
<th>Values</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structure</td>
<td>N</td>
<td>200</td>
</tr>
<tr>
<td></td>
<td>L</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>10,000</td>
</tr>
<tr>
<td></td>
<td>ρ</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>σ_{ij}</td>
<td>0.05</td>
</tr>
<tr>
<td>Drivers</td>
<td>r</td>
<td>[0.0, 1.0]</td>
</tr>
<tr>
<td></td>
<td>( \mu )</td>
<td>[0.05, 0.5]</td>
</tr>
<tr>
<td></td>
<td>( \sigma_e )</td>
<td>[0.004, 0.04]</td>
</tr>
</tbody>
</table>

Table 4.1: Simulation settings for FFPopSim as employed in [2020]. Light Gray for parameters that are varied in order to test the reconstructions routes GC vs KNS.

Figure 4.1: Scatter Plots for set MR. The red stars result from the inference based on the Gaussian Ansatz, Eq.(4.10); blue dots for KNS reconstruction Eq.(1.28) with MF inferred couplings. Epistatic fitness strength \( \sigma_e = 0.004 \), other parameters as shown in Tab.(4.1).
examination: either inferring epistatic fitness components directly from correlations and means exploiting Eq.\( (4.16) \) or implementing DCA to reconstruct \( \{ f_{ij} \} \), then finally using Eq.\( (1.28) \). About the latter, we will employ the only Mean Field method \( MF \) as a DCA technique; results for PLM are found to be similar (data not shown). Finally, we will compare the results of the reconstructed \( \{ f_{ij}^* \} \) (by GC and KNS) with the true values \( \{ f_{ij} \} \) in input and, in so doing, test the relative performance of the two procedures. The error in the reconstruction can be quantified by the root mean square error introduced in Eq.\( (3.39) \):

\[
\epsilon = \sqrt{\frac{\sum_{ij} (f_{ij}^* - f_{ij})^2}{\sum_{ij} f_{ij}^2}},
\]

which we want to be as small as possible.

○ **Scanning the parameter space.** Our simulations will be divided in two sets. The first one, named MR, aims to explore the parameter space in the \( \mu, r \) directions for a fixed epistatic strength \( \sigma_e = 0.004 \). In Fig.\( (4.1) \) we show some scatter plots as resulting from the algorithm aforementioned; each column has the same \( r \), increasing from left to right, and each row has the same mutation rate, increasing from bottom to top. For each inference technique, the core information of each scatter plot can be captured by \( \epsilon \in \mathbb{R}^+ \), Eq.\( (3.39) \). In Fig.\( (4.2) \) there is an heat map for \( \epsilon \), as evaluated while scanning the parameter space in the directions \( \mu \in [0.05, 0.5] \) and \( r \in [0.0, 1.0] \).

The second set of simulations, we shall call it ER, is intended to ascertain the performances of the two reconstructions as a function of \( \sigma_e, r, \) for a fixed \( \mu = 0.2 \). In Fig.\( (4.3) \) each column has the same \( r \), increasing from left to right, and each row has the same epistatic...
strength, increasing from bottom to top. In Fig. (4.4), finally, the heat map for $\varepsilon$ by tuning $c_\varepsilon \in [0.05, 0.5]$ and $r \in [0.0, 1.0]$. 

4.2.2 Results & Comments

Inference of epistasis based on the Gaussian Ansatz outperforms the KNS reconstruction. We will now supply more the details and (most importantly) the boundary of this statement, which stems from the observation of Fig. (4.1-4.4) straightaway. Let us have a closer look to them one by one.

* MR, Fig. (4.1, 4.2). As we know from Sec. (1.3, 3.4), in the KNS theory the mutation rate cannot be zero, otherwise the QLE would be but a long-lived transient towards fixation.
Figure 4.4: Heat Maps for the ER set. The colour represents the reconstruction error \( \epsilon \) given in Eq.\((3.39)\). Left: KNS reconstruction Eq.\((1.28)\) with nMF inferred couplings. Right: inference based on the Gaussian Ansatz, Eq.\((4.16)\). Mutation rate \( \mu = 0.2 \), other parameters as in Tab.\((4.1)\).

Nevertheless, Eq.\((1.28)\) assumes \( \mu = 0 \) therefore we do not expect it to work for sufficiently high mutation rates. Moreover, since QLE requires ah high recombination rate, we do not expect it to work in a region of the phase space where \( r \sim 0 \) either. Both these expectations are confirmed by Fig.\((4.1, 4.2a)\). Reconstruction based on Eq.\((4.16)\) on the other has a better performance everywhere throughout the MR set, in particular for high mutation rates. For extremely high recombination and mutation rates (top-right) the noise due to the strength of the reshuffling of the population is likely to worsen the accuracy of the allele statistics (means and correlations) for a finite-time simulation, which ultimately results in a sparser scatter plot and a slightly higher reconstruction error \( \epsilon \).

* ER, Fig.\((4.3, 4.4)\). Since we have fixed an high value of the mutation rate \( \mu = 0.2 \), we do not expect the KNS to work anywhere. Indeed, it does not. Inference based on the Gaussian Ansatz on the other hand has excellent performances except in a region of high epistasis and high recombination. We noted in Sec.\((4.1.3)\) that the perturbative expansion is meant to be accurate whenever \( L \sigma_e < 1 \); accordingly in Fig.\((4.4b)\) we see the error increasing for increasing \( \sigma_r \). However, it is impossible to sweep it under the carpet: there must be a different and deeper explanation for the behaviour in Fig.\((4.3b, 4.3c)\), which shows some clear patterns (i.e. the two symmetric clouds of reconstructed points) inexplicable in terms of a simple increased amount of noise. As we shall see, with this remark, we have just opened up the Pandora’s box, we will devote the entire Ch.\((5)\) to dig deeper and deeper into this conundrum.

Finally, let us also stress that for the Gaussian Ansatz to hold, the \( \{f_i\} \) have to be small, too. Increasing the magnitude of the additive fitness, the evolutionary process is more and more in danger of encouraging the emergence of clones, against the hypothesis of a monoclonal population at the very beginning of this chapter.
4.3 DIRECTIONS

Is it needed to assume a specific form for the probability distribution \( P(g) \) of the form Eq.\((1.17)\)? The result of this chapter seems to suggest that no, it is not. Indeed, rewinding to the very beginning, we realize that to get Eq.\((4.8, 4.11)\) only two assumptions are needed: the Stosszahlansatz Eq.\((1.8)\) for the two-genome distributions, and a standard Gaussian closure of the dynamical equations for the moments. Explicit assumptions on the underlying probability distributions are not needed: even in the region where we have used it \((4\mu + r_{ij}) \to \infty\), the Gaussian hypothesis it is not meant to be an effective description of all the statistical properties of the population, it is just appropriate for the purpose of estimating the 3, 4–points correlations. This last remark is an ending point for this chapter, but several improvements are left for the future. For instance, since one of the strongest limitation to our results is the hypothesis of a mono-clonal population, the next natural step would be that of generalising the present approach to consider Gaussian mixture model for a multi-cluster population i.e. one in which there is the coexistence of several clades, far apart in the genotype space, each of them described as a Gaussian "cloud". Another one is to generalize these results to multi-allelic loci, following the philosophy of [Gao et al., 2019]. Both these directions involve a growing amount of complexity, which will eventually result in an increased biological realism. For our part, as they say, we hope to return to these questions in a future contribution.
Evolution is highly complex. This overwhelming complexity might be daunting for a theoretical physicist dealing with the problem of its formal description, but it is the reason of the stunning variety of life that we observe. A first step in approaching the modelling of the evolutionary process is to set the stage i.e. the parameter space, the "amount of complexity" we want to take into account. Given a set of parameters and a further evolutionary rule that governs the process, we are able to study and/or simulate the evolving system e.g. the time evolution of interesting observables. Then, we are interested on the stability of our conclusions under variation of one or more initial parameters, at least in their asymptotic behaviour. This last task might be more challenging since typically there is no single theoretical tool (an approach, an equation, a method...) capable of explaining the behaviour of the system everywhere in the parameter space, and one has to humbly make additional hypotheses. Indeed, introducing new hypotheses is nothing but restricting the region under investigation to a narrower size: obviously, the narrower the region under investigation, the higher the chance to succeed (but the less the relevance of the result and the glory that stems from it). Much like a mosaic, designed by inlaying bits of coloured glass and stone of various size, we aim at understanding and tiling the parameter space with piece-wise theoretical explanations.

Both the domains of the QLE and of the Gaussian Closure are but tiny fractions of the parameter space and a number of unexplored effects and behaviours may be there right behind the corner as soon as we step out of them.

Here in Sec.(5.1) we introduce, describe and characterize the behaviour of a population of genomes in a regime where the epistatic components of the fitness landscape are not small and dominate the dynamics. We shall call it NRC-phase, since its first clear signature is for each locus the Non-Random Coexistence of both the alleles in the population. In Sec.(5.2) will also review some recent results from the literature hunting for further relevant insights and discuss a putative experimental observation of the NRC-phase, previously reported in literature. Finally, in Sec.(5.3) we will sum up what we do (not) know.

5.1 HALLMARKS OF THE NRC-PHASE

5.1.1 Localization in the Parameter Space

Qualitatively speaking, we will explore a regime characterized by an high mutation rate $\mu$ (unlike QLE) and high epistasis $\sigma_e$ (unlike QLE and Gaussian Ansatz); these two requirement will hold
throughout the chapter and are at the heart of the NRC-phase. If the mechanisms of our model had independent effect on our observables, the simplest and most rational approach would be to study in the first place the case where everything but mutations and epistasis is turned off, then switch on one by one any other "additive" complication to the minimal model. Unfortunately, this almost never the case with complex systems, certainly not in our case, as simulations show; therefore we will place ourselves in a convenient point and later try to understand if/how our conclusions are stable under perturbation in some directions.

Apart from the aforementioned strong mutations and epistasis, we will start from describing the behaviour of genomes that are allowed to recombine \( r \neq 0 \) and that try to maximise a purely epistatic fitness function

\[
F(g) = \sum_{i<j} f_{ij} s_i s_j ,
\]

i.e. a Sherrington-Kirkpatrick (SK) fitness function, where \( f = \{ f_{ij} \}_{ij} \) is a symmetric matrix of Gaussian distributed random numbers \( f_{ij} \sim N(0, \sigma_e) \), \( f_{ii} = 0 \) \( \forall i \), \( f_{ij} = f_{ji} \) \( \forall i, j \), see Sec.(3.4).

We trivially observe that such a fitness function is invariant under the transformation \( g \rightarrow -g \), hence we expect these two to be selected with equal probability.

Finally, let us note that out of the domain of the QLE and the Gaussian Ansatz, it would be surprising to see either Eq.(1.28) or Eq.(4.16) working: indeed, they both fail, as it will be clear soon, therefore, for the time being, we leave hope to carry out this task and will not discuss reconstruction any later.

### 5.1.2 7 Observations

A possible set of parameters is the summarized in Tab.(5.1): we shall call it RS (Reference Simulation) and refer to it for future comparisons. It is not at all a random choice of values, they are instead tuned so that the system dynamics lies at the edge between the well-known QLE phase and the new NRC phase, to be explored. Our line of attack will then be to describing

<table>
<thead>
<tr>
<th>FFPopSim</th>
<th>RS</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>500</td>
<td>carrying capacity</td>
</tr>
<tr>
<td>L</td>
<td>25</td>
<td>n. of loci</td>
</tr>
<tr>
<td>T</td>
<td>10.000</td>
<td>n. of generations</td>
</tr>
<tr>
<td>Drivers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>r</td>
<td>0.5</td>
<td>outcrossing rate</td>
</tr>
<tr>
<td>( \rho )</td>
<td>0.5</td>
<td>crossover rate</td>
</tr>
<tr>
<td>( \mu )</td>
<td>0.5</td>
<td>mutation rate</td>
</tr>
<tr>
<td>( \sigma_r )</td>
<td>0.024</td>
<td>( f_{ij} \sim N(0, \sigma_e) )</td>
</tr>
</tbody>
</table>

Table 5.1: Parameters of the RS. High mutation and recombination regime, random crossovers. SK fitness function. Quenched \( f_{ij} \sim N(0, \sigma_e) \), \( f_{ii} = 0 \) \( \forall i \), \( f_{ij} = f_{ji} \) \( \forall i, j \). Random initial configuration.
the latter in contrast to the former, exploiting the computational tools already introduced in Ch.(2).

(a) Evolution of all the $L$ first order cumulants $\{\chi_i\}$.

(b) Evolution of $\chi_5$, that follows the orange-like pattern in Fig.(5.1a).

(c) Evolution of $\chi_7$, that follows the red-like pattern in Fig.(5.1a).

Figure 5.1: RS, Tab.(5.1). Evolution for the first order cumulants $\{\chi_i\}$ at the edge of instability QLE$\leftrightarrow$NRC. While $\chi_i \sim 0$ in QLE, we see $|\chi_i| \sim \alpha \neq 0$ in a NRC phase. Two different groups of alleles emerge, following two symmetric trajectories.
1. $\chi_i$: First Order Cumulants. In Fig.(5.1a) we show the all-time dynamics of all the magnetizations $\chi_i = \langle s_i \rangle$, cf. Fig.(2.2). We observe the first cumulants intermittently hovering between the expected behaviour in a QLE phase, i.e. $\chi_i \sim 0$, and a new phase where qualitatively $|\chi_i| \sim a \neq 0$ i.e. where for each locus both the alleles are found in the population in a non-random fashion. The value of $a$ is observed to primarily depend on $\mu, \sigma_e$. This observation for the $\{\chi_i\}$ is at the very core of this NRC phase, hence the name Non-Random Coexistence.

When increasing more and more the strength of epistasis $\sigma_e$ the intermittence disappears and the dynamic of the system is always NRC-like, we will comment more on this soon. Obviously, being $\mu \gg 0$ we will never observe the fixation of any allele, the constant influx of mutations will always tend to destroy patterns established by selection, this balance will eventually lead to an equilibrium i.e. to a specific value of $a$.

We underline it is not at all trivial the existence of a transition in both directions QLE $\leftrightarrow$ NRC: one cannot assume in general that the same mechanism is (somehow) responsible for both the directions.

Finally, let us note that there is an additional element of regularity hiding in Fig.(5.1a): the time trajectory of all the $\chi_i$ in the latter are either of the shape shown in Fig.(5.1b) or the specular one in Fig.(5.1c) but never a mixture of these two. The already noted symmetry $g \rightarrow -g$ of the fitness function is evidently playing a role.

2. $\chi_{ij}$: Second Order Cumulants. In Fig.(5.2) the all-time dynamics of the allele correlations $\{\chi_{ij}\}_{j=1}^L$ is displayed, analogous plots are observed $\forall i$, cf Fig.(2.3). We notice the same pattern of intermittence as for the first order cumulants: while in QLE $\chi_{ii} = 1 - \chi_i^2 \sim 1 \land \chi_{ij} \sim 0$, in the NRC regime one qualitatively has $\chi_{ii} < 1 \land |\chi_{ij}| \sim \beta \neq 0$, where $\beta$ again depends primarily on $\mu, \sigma_e$. In this case such behaviour is almost overshadowed by the fluctuations.

3. Clonal Structure. One straightforward explanation springs to mind when examining the plots just shown in Observations 1. and 2.: Clonal Competition (CC) phase. Suppose there were in the population two clones, one larger than the other, whose relative sizes were determined by the fitness of their genotypes. This could match for instance what we have observed for the $\chi_i$ in the NRC-phase and in particular the long-time coexistence of the two groups of alleles. In fact, we are working in a very high $\sigma_e$ regime and it is well known that increasing epistasis with respect to recombination (in absence of mutations) enhances the probability of the onset of a CC-phase [R. Neher et al., 2013; R. A. Neher & Shraiman, 2009]. The mutation mechanism nevertheless may affect this picture and to test these hypotheses we resort the Clonal-Structure Plot introduced in Sec.(2.2), as shown in Fig.(5.3), cf Fig.(2.5a).

The result is indeed surprising: we can rule out a CC-phase, since in correspondence to a NRC region, most of the population is made by single-clone genotypes (dust-like region); however at least one non-singular clone emerges from the dust: its (fluctuating) relative size depends on $\mu, \sigma_e$ and this behaviour survives until the NRC-phase melts in the QLE.
randomness. What is even more intriguing, the observer fraction of individuals within the largest clone is not at all sufficient to explain the aforementioned parameter $\alpha$ so the we can rule out the hypothesis of a completely random population apart from a single relevant clone that biases the results.

4. **Fitness Statistics.** What drives the system in/out from a NRC-phase? A hint comes from our last *all-time* plot: population-wide fitness statistics (mean, st.dev.), see Fig.(5.4), cf Fig.(2.4). We consistently note that the NRC phase is characterized by an higher fitness mean with respect to the QLE expectation, therefore selection will encourage the former at the expense of the latter. Mutations, on the other hand, represent a counterforce since they destroy existing genotypes and create new ones: in a regime where both of them are important, their balance is crucial. We may guess that, at the edge between the two phases, the accidental appearance of very fit clones is responsible for the transition $\text{QLE} \rightarrow \text{NRC}$, their accidental disappearance due to mutations is responsible for the opposite one. We will extensively build on these two intuitions in Sec.(5.2.1 - 5.2.2).

5. **Genotype Snapshot.** The first of our *instantaneous* plots adds no information to what we already know, but gives a much clearer picture. In order to understand the behaviour in a NRC-phase, we fix for instance $T = 8500$, the resulting population snapshot is shown in Fig.(5.5), cf Fig.(2.6). The difference from what we had observed in the QLE phase is self-evident: strong patterns of regularity emerge, accordingly with the allele means fluctuating around non-zero values.

6. **Fitness Distribution.** In Fig.(2.7) we have found that in QLE the instantaneous fitness distribution for a population evolving in a Sherrington-Kirkpatrick fitness landscape with
Figure 5.3: RS, Tab. (5.1). Clonal structure throughout evolution. For each $t$, a vertical line shows the relative size of the clones present in the population, from the largest (bottom); zoom to the first 10%. The NRC phase is marked by the emergence of non-trivial clones, yet too small for the population to be regarded as in a CC-phase.

Figure 5.4: RS, Tab. (5.1). Evolution of fitness mean and st.dev. The QLE ↔ NRC pattern is clearly visible and entails a jump in both mean and st.dev. to higher values, suggesting the emergence of very fit genotypes.

Figure 5.5: RS, Tab. (5.1). Population snapshot at $t = 8500$, 200 samples. $Y/B$ for $\chi_i = \pm 1$. Strong patterns emerge in a NRC phase, as suggested by the dynamics of the first order cumulants.
Figure 5.6: RS, Tab. (5.1). Fitness distribution at \( t = 8500 \), zoom on the high fitness tail. The asymmetry of the distribution is due to the mechanism of selection, that penalizes unfit individuals. The presence of spikes in the high fitness region is peculiar of the NRC phase: its (dis)appearance follows the same pattern of instability QLE ↔ NRC as in Fig. (5.1a).

Mutations and recombination has approximately the shape of a Gaussian \( \sim \mathcal{N}(0, \sqrt{\sigma_e}) \). This observable in the NRC phase is of particular interest, considered what has been said above on the fitness statistics, a typical result is the one in Fig. (5.6). Two behaviours strike our attention:

* Asymmetry. Ignoring for the moment the high fitness tail, let us focus on the main body of the distribution, which is clearly not symmetric but biased towards positive values. This is mostly due to the balance between mutation and selection: while the first favours randomness, the second tends to encourage the emergence of individual whose fitness is higher than the average, penalizing unfit individuals. Even in absence of recombination, there is a number of phenomena that are governed by this apparently simple balance, see for instance [Desai & Fisher, 2007; Walczak et al., 2012].

* Fit Spikes. Interestingly, the NRC phase is marked by the rise of a group of very-fit individuals, see zoom in Fig. (5.6). At least in this regime, their (dis)appearance is always an hallmark of the transition out from / into the NRC phase. This in turn suggests an explanation for the jump observed in the fitness statistics, since fluctuations in the number of individuals in the fittest region have strong (delayed) effects of the fitness mean [R. Neher & Shraiman, 2012]. Finally, as one might expect, simulations show that the size of the fit peaks grows by increasing the fitness strength (data not shown).

7. Quenched disorder As a final observation, we report the dependence of the behaviour of the simulation from the specific realization of the \( \{ f_{ij} \} \), in the sense that, at the edge
of instability, the dynamics of the system strongly depends on the details of the fitness landscape. The designation of quenched disorder comes from the analogy with the spin-glasses [Castellani & Cavagna, 2005]. This tells us that $\sigma_e$ is but a hyper-parameter and it not suitable if one wants to capture all the relevant information about the epistatic fitness landscape. This last point is true at the edge of instability between QLE-NRC but not for weak epistasis (always QLE) or extreme epistasis (always NRC), as we shall see in Sec.(5.1.3).

### 5.1.3 What If(s)

This section is dedicated to the report of some further preliminary observations on the NRC phase when stepping out of the familiar framework explored above. It is meant to be neither an exhaustive or a sufficient mapping of all the ways of interest in which one can exploit FFPopSim to get some insights on the NRC regime; instead it must be regarded as a humble attempt of addressing some questions that may arise when looking at the results previously illustrated. For this reason, our choice has been to put it in the form of answers to questions like "What if ... ?". Some of them will turn out to be a completely new direction to pursue, further efforts will be required in future works.

- $f_{ij} > 0 \forall i \neq j$ (analogously for $f_{ij} < 0 \forall i, j$). Let us recall $F(g) = \sum_{ij} f_{ij} s_i s_j$, $f_{ij} \sim N(0, \sigma_e)$. In this setting the system is frustrated, a frustration that vanishes if the $f_{ij}$ all have the same sign. In the latter case, the selection will favour either the state $\bar{g} : \{s_i = +1 \forall i\}$ or $-\bar{g}$, and it is only a matter of chance whether the dynamics will lead the system toward one or the other. For instance, in Fig.(5.1-5.2) all the means/correlations will behave in the same way, following the same branch, symmetry will disappear. This trivial prediction is indeed confirmed by simulations, see Fig.(5.7)

![Figure 5.7: Evolution of first order cumulants. Parameters of the simulation as for RS, Tab.(5.1), except for $\sigma_e = 0.0087$ and $f_{ij} \geq 0 \forall i, j$. The symmetry observed in Fig.(5.1) is broken.](image-url)
○ **No recombination,** $r = 0$. Recombination, along with mutations, contributes to reshuffle the genetic pool, hence increasing the probability of the appearance of new genotypes in the population. Since we have only $N$ walkers in a space of genotypes which has $2^L$ possible states ($N \ll 2^L$), if we suspect the transition QLR$\rightarrow$NRC is due to the population finding some very fit sequences and climbing the corresponding fitness hill, the typical time scales of mutations and recombinations will play a crucial role. We cannot turn off mutations since, as we stated above, they are by definition strong in a NRC phase. Nevertheless, we can assess the simulations in absence of recombination, $r = 0$. Using RS Tab.(5.1) with $r = 0$ destroys any signature of the NRC, in according to our expectations. One way to compensate this effect is to increase the fitness strength $σ_e$, in fact for a sufficiently high epistasis, we see again a behaviour of allele means very similar to that observed in Fig.(5.1), although some trajectories are much more unstable, see e.g. Fig.(5.8). In general, we observe much stronger fluctuations, which forces us to choose an high value of the system expected size $N$ in order to get a clearer picture. Transition probabilities appear to be affected by recombinations as well. But the most important remark is the following: our simulations suggest that when $r = 0$, even for very high values of fitness strength, observations 3., 4., 6. in Sec.(5.1.2) do not hold true anymore. In other words, the mechanisms that sustains the birth/death of fit individuals in the high fit tail of the fitness distribution seem to strongly depend of the recombination mechanism. These remarks support the hypothesis that the drivers (or consequences) of the transitions between the two phases here are fundamentally different from those in the case with recombination; in absence of further investigations, one cannot regard the latter as a mere "additive" complication to the model.

![Figure 5.8: Evolution of first order cumulants. Parameters of the simulation as for RS, Tab.(5.1) except for $N = 10.000, T = 2.000, σ_e = 0.09, r = 0$. In the case with no recombination, the behaviour of the system shows substantial differences with respect to the case $r \neq 0$.](image)

○ **Additive fitness switched on,** $f_i \neq 0$. Additive fitness drives each allele to fixate; more specifically $s_i \xrightarrow{add.} \text{sgn}(f_i)$. Differently from epistasis, additive fitness acts **independently**
in each locus, and in presence of epistasis these two variants of natural selection can compete and give rise to a complex equilibrium. We here focus on the narrower question: is the NRC due to the specific role of epistasis? To test it, we can switch on additive components of fitness, for instance we can use $RS$ with in addition $f_i \sim \mathcal{N}(0, \sigma_a)$, with $\sigma_a = 2.0 \gg \sigma_e$. The outcome is: no, even in this case the observations in Sec.(5.1.2) are confirmed, evidently as long as the overall fitness is sufficiently high, it does not matter whether it is additive or epistatic. In the case where both of them are present, a more suitable selection parameter is $\sigma^2 = \sigma^2_a + \sigma^2_e$.

- **$N - \sigma_e$ tuning.** As we know from Sec.(1.4), the size of fluctuations depends on $\sqrt{N}$. The higher the fluctuations in the higher the probability of finding the system in an atypical state. In the transition QLE $\rightarrow$ NRC is triggered by the system hitting a small subset of the possible states, then we expect larger fluctuations to enhance the corresponding chances. Accordingly, for a fixed $\sigma_e$, we will see QLE for sufficiently low $N$ and NRC for sufficiently high $N$. On the other hand, for fixed $N$, we already know that higher epistasis will favour a NRC phase. In Fig.(5.9) we qualitatively test these expectations for different values of $N-\sigma_e$. With regard to the intermediate region in $\sigma_e$, we see explicitly the effect of the dependence of the system fate on the specific realization of the $\{f_{ij}\}$ (some simulations may seem to be QLE or NRC for $T < \infty$ even if in a instability region); on the contrary, the extrema do not show such behaviour.

### 5.2 UNDERSTANDING THE NRC-PHASE: INSIGHTS

Now that we have characterized the NRC phase by several observations, time has come to sketch correspondences between them, exploiting both intuition and the existing literature.

#### 5.2.1 QLE $\rightarrow$ NRC : Intuition

A number of previous contributions have investigated the balance between mutations and selection; however, because of the complex entanglement of the phenomena at stake, most of the researchers have worked under several strong assumptions, as the only hope of drawing quantitative predictions. But with doing this, one inevitably loses generality of the results and the art lies into finding an equilibrium.

As an example, let us briefly consider [Desai & Fisher, 2007]; the authors study the mechanism that leads to the fixation of a new advantageous mutation in the population, which may be the phenomenon that drives our system into the NRC phase. They conceptually isolate three different regimes: 1. strong-selection, weak mutation, small populations (*successional-mutations regime*); 2. clonal interference in large population 3. strong-selection, strong-mutation (*concurrent mutations regime*), which is the relevant case for us. Their approach is based on the estimation of the typical timescales for the appearance, establishment (i.e. survival to random drift) and
Figure 5.9: Qualitative scan of the parameter space in the $N, \sigma_e$ directions, all other settings as in RS, Tab.(5.1). Each point is obtained as follows: we fix a pair $(N,\sigma_e)$ and run two simulations for $T = 10,000$ generations. If at any point we observe a transition QLE$\leftrightarrow$NRC, we mark the point as red, instability. If the system throughout the simulations appears to always be in a QLE/NRC phase, we colour the point in blue / green, respectively. In order to classify the two phases automatically, we choose the fitness mean, cf Fig.(5.4), as an observable and set a threshold. Due to the finiteness of the simulation time and to the dependence on the quenched disorder, the instability region is likely to be broader than what shown here.

fixation of new mutations, depending on the parameters of the model. Unfortunately, their investigation is conducted in a framework that is essentially different from the present one: they assume no epistasis, no recombination, only beneficial mutations; we cannot expect to match their predictions with our data, and we will not even try.

Nevertheless, we can pursue a similar line of thought based on the probabilities of appearance and emergence in order to develop a plausible mechanism that could drive the QLE $\rightarrow$ NRC transition in the framework of our interest. Let us consider a system like that in RS, two processes should be relevant:

1. Appearance of a genotype with high fitness in the population. Let’s call $P_A(g|\mu,r,N,\{f_{ij}\})$ the probability for a sufficiently fit genotype $g$ to appear. We have only $N \ll 2^L$ individuals, there should be some $t_A(\mu,r,N,\{f_{ij}\})$ typical waiting time. How to estimate the $t_A$ is an open question.

2. Establishment of the clone with the genotype $g$, i.e. the clone is large enough so that it manages to survive to the random drift, mutations and recombination. Let’s call $P_E(g|\mu,r,N,F(g))$ the corresponding probability.
One may think to a threshold value $F^*(\mu, r, N)$ above which a clone is likely to emerge. This framework could suggest an explanation for the behaviour observed Fig. (5.9): let us define $F_{\text{max}} = \max_g F(g)$, if $\sigma_e$ is low enough, $F_{\text{max}} < F^*$ and no genotype is able to emerge (QLE phase). When $\sigma_e$ is such that $F_{\text{max}} > F^*$ then it is possible for one or more clones to emerge (QLE/NRC instability), here the quenched disorder plays a fundamental role. Finally, for very high values of $\sigma_e$, $F_{\text{max}} \gg F^*$, several clones emerge and disruptive forces (or whatever) do not manage to sweep them all out (NRC phase).

There is obvious need to fill the gap between intuition and quantitative predictions, before moving one step forward and testing these hypotheses. For the moment, we can but limit ourselves to point the way and hope to come back on it soon.

### 5.2.2 NRC $\rightarrow$ QLE: Muller’s Ratchet

We have observed that the NRC phase is marked by the presence of very fit spikes in the fitness distribution of the population. Whether the disappearance of these fit spikes causes or is caused by the transition NRC $\rightarrow$ QLE, we cannot say; however we can exploit the observation of the erosion of the fittest classes to draw correspondences with previous results in literature. In fact, there is a well-known phenomenon that could be useful, famous enough to have its own name: Muller’s Ratchet, [Muller, 1964].

We still face the same problem as in the previous section, many works on this topic are based on approaches that are too different from ours to match quantitative predictions, but some insights can be drawn from them. We now summarize briefly the content of [R. Neher & Shraiman, 2012], among the most recent efforts that are relevant for this section:

√ A click of Muller’s ratchet is the loss of the most fit class of individuals. The rate of the ratchet is given by the inverse of the mean time between successive clicks of the ratchet.

√ The authors investigate a model where the population of $N$ individuals is grouped into discrete classes, each characterized by the number $k$ of deleterious mutations; There is no recombination $r = 0$, deleterious mutations happen at rate $u$, each mutation causes a fixed fitness loss $s \ll 1$. A master equation

$$\frac{d}{dt} n_k = s(\bar{k} - k) n_k - un_k + un_{k-1} + \sqrt{\bar{n}_k \eta_k}$$

drives the stochastic evolution of $n_k = \text{number of individuals in the } k\text{-th class, } \eta_k$ is the noise term.

√ When $\lambda = u/s \gg 1$ the top fit class of individuals contains only a few individuals and is susceptible to an accidental extinction. In general, the magnitude of the fluctuations in the number of individuals of the most-fit class is governed by the combination $Ns$.  

Figure 5.10: Deleterious mutation–selection balance. The population is distributed among classes of individuals carrying \( k \) deleterious mutations. Classes with few mutations grow due to selection (red arrows), but lose individuals through mutations (green arrows), while classes with many mutations are selected against but replenished by mutations. From [2012].

Taking into account the fluctuations of most-fit class, those (delayed) of the mean fitness and by exploiting a stochastic path integral approach, it is possible to estimate the mean time between clicks given by:

\[
T_{\text{click}} = \frac{2.5 \xi(\lambda)}{\alpha(\lambda) s \sqrt{Nse^{-\lambda}}} e^{N\alpha(\lambda)e^{-\lambda}},
\]

where \( \alpha(\lambda) \sim O(1) \) and \( \xi(\lambda) \sim \log \lambda \).

In other words, since \( T_{\text{click}} \) is the escape time from the NRC phase, as the argument goes, we expect in a rough approximation \( T_{\text{click}} \sim \exp(N) \).

### 5.2.3 Escape Times

As we have already pointed out, there is no \textit{a priori} reason to believe that the two transitions QLE \( \leftrightarrow \) NRC are governed by the same mechanism: indeed, we will see they are not. We have described in the previous two sections ideas that differ substantially in methods, but look at basically the same observable: the escape time. Let us call \( t_{QLE}, t_{NRC} \) the escape times from the QLE and NRC phases, respectively, and let us question our simulations about them. The algorithm is the following:

1. **Simulation.** If we tune our parameters so to set the system in a region of strong instability QLE \( \leftrightarrow \) NRC and run simulations for \( T \) as large as possible, we will observe the system undergoing multiple transitions in both directions. Let us call such a simulation ET, for instance we can set Tab.(5.2)
### Table 5.2: Parameters of the ET. The total number of generations simulated is 150 times higher with respect to RS.

<table>
<thead>
<tr>
<th>FFPopSim</th>
<th>ET</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>575</td>
<td>carrying capacity</td>
</tr>
<tr>
<td>L</td>
<td>25</td>
<td>n. of loci</td>
</tr>
<tr>
<td>T</td>
<td>150,000 × 10</td>
<td>n. of generations</td>
</tr>
<tr>
<td>r</td>
<td>0.5</td>
<td>outcrossing rate</td>
</tr>
<tr>
<td>ρ</td>
<td>0.5</td>
<td>crossover rate</td>
</tr>
<tr>
<td>µ</td>
<td>0.5</td>
<td>mutation rate</td>
</tr>
<tr>
<td>σₑ</td>
<td>0.029</td>
<td>( f_{ij} \sim N(0, \sigma_e) )</td>
</tr>
</tbody>
</table>

2. **Classification.** The mean fitness of the population is a suitable observable in order to set a threshold and automatically classify the two phases (above: NRC; below: QLE), hence to evaluate the number of generations spent in one phase since the last transition (escape time). As an output of this last step, we get a list of escape times from the QLE phase and analogously for the NRC phase.

3. **Distribution.** We can draw histograms out of those two lists and observe the distributions of the escape times \( t_{\text{QLE}} \), \( t_{\text{NRC}} \), see Fig. (5.11). In both cases, we observe a distribution that in a first approximation can be described as exponential, so we try to fit it with

\[
y(T) = \gamma_i e^{-a_i T}, \quad i = \text{QLE, NRC}.
\]  

In particular we are interested in \( a_{\text{QLE}} \), \( a_{\text{NRC}} \), that set the rates of the transitions, and use \( t_{\text{QLE}} \sim 1/a_{\text{QLE}} \), \( t_{\text{NRC}} \sim 1/a_{\text{NRC}} \) as qualitative measures of the expected escape time from the two phases. One caveat is necessary: we do not expect our histograms to be accurate at low magnitudes of escape time, since the algorithm that leads us from the raw simulation data to these figures has been sullied by elements of arbitrariness (e.g. the value of the threshold to classify the phases). Furthermore, the data systematically show fatter tails than those captured by the exponential fit, one more source of errors. Several of these problems may be fixed (or, at least, better controlled), nothing but time and patience is required.

4. **N Dependence.** Finally, in order to check the dependence of these last quantities on the size of the population, as suggested in the previous section, we repeat the steps 1.-3. for different values on \( N \) and plot the behaviour of \( t_{\text{QLE}}(N) \), \( t_{\text{NRC}}(N) \) in Fig. (5.12).\(^1\)

The output of this procedure seems to be unequivocal, the two directions of the transitions QLE\(\leftrightarrow\)NRC are governed by two different mechanisms, that scale with \( N \) differently. The (rough estimate of the) escape time from the QLE phase appears to be almost insensitive

\(\text{The range of } N \text{ values that can be tested is both upper and lower bounded by the computational resources: whenever the transitions are rarer, it is necessary to simulate more generations in order to collect a sufficient statistics for the histograms.}\)
Figure 5.11: Distribution of Escape Times from the QLE phase (yellow) and from the NRC phase (gray). Parameters of the simulations as in Tab. (5.2), two example system sizes \( N = 575, 675 \). For each phase, we attempt a linear fit in the semi-log scale, in particular we get \( a_{\text{QLE}}, a_{\text{NRC}} \) in Eq. (5.1). The exponential curves in linear scale resulting from the fitting procedures are also shown.

Figure 5.12: Behaviour of \( t_{\text{QLE}} \sim 1/a_{\text{QLE}}, t_{\text{NRC}} \sim 1/a_{\text{NRC}} \) as a function of the carrying capacity \( N \). The plot is a strong evidence of the fact that the transitions QLE→NRC and NRC→QLE are due to different mechanisms. While \( t_{\text{QLE}} \) seems to be almost insensitive to \( N \), \( t_{\text{NRC}} \) is compatible with a behaviour \( \sim \exp(N) \), as confirmed by the coefficient of determination \( R^2 \sim 1 \).

to the size of the system, while that from the NRC phase is compatible with an exponential behaviour \( \sim \exp(N) \). This latter result is consistent with our previous discussion on the
Muller’s Ratchet and can be used as a starting point for a model of the phenomenon for instance in the framework of an escape time over a potential barrier, akin the *Arrhenius formula* 

\[ T_{\text{esc}} \sim \gamma \exp\left[\frac{(U(b) - U(a))}{D}\right] \]

where \( U(b) - U(a) \) is the height of the potential barrier and \( \gamma, D \) constants, see [Gardiner, 2004], Sec.5.2.7.

### 5.2.4 Speculation: A NRC-phase for E.Coli?

If one had to name only one experiment in evolutionary biology, than no doubt about the answer: it would be the celebrated R. Lenski’s *E.coli Long Term Evolution Experiment (LTEE)*. It is an ongoing study that has been tracking genetic evolution in 12 initially identical populations of asexual *Escherichia coli*, in the same medium, since 24 February 1988 (currently more than 70,000 generations observed).²

The paramount interest in this experiment lies in the fact that it is the largest dataset available that puts evolution under the spotlight of the experiment: bacteria grow, mutate, evolve (recombination is negligible for *E.coli*). The complexity of the evolutionary process emerges clearly from the ensemble of observed behaviours across populations: clades arise and diverge, small mutations leading to genetic catastrophes, intricate interactions with the environment. In a recent contribution, [Good et al., 2017] have worked out the latest analysis of the LTEE, enquiring the stochastic dynamical process that governs how mutations arise and spread through the populations. There is no room here to report it all, we shall instead focus on what the authors call *quasi-stable coexistence* therein.

To measure the dynamics of the molecular evolution, they proceed as follows:

1. In each of the 12 LTEE populations (named Ara+1 to Ara+6 and Ara-1 to Ara-6) samples are taken and sequenced at 500-generation intervals across 60,000 generations of evolution;

2. Mutations that have reached \( \sim 10\% \) frequency in at least two sampled time points are identified and the frequency of the derived allele is tracked through the rest of the time-course.

The result of this procedure is showed in Fig. (5.13). The reader that has made the effort of reading the chapter this far should leap back when observing this figure, especially Ara-6, Ara-1, Ara+5, but caution is mandatory. Let us focus e.g. on Ara-6. Current models of both “periodic selections” (where individual driver mutations fix in a sequence of discrete selective sweeps), or clonal-competition predict that, sooner or later, mutations should either fixate in the population or go extinct. But in Fig.(5.13) we observe clearly mutations segregating into (at least) two intermediate-frequency clades that coexist for long periods, hence the name *quasi-stable coexistence*; specifically, \( 9/12 \) populations have clades that coexist for more than 10,000 generations.

---

generations, the relative abundance of the two clades varies from population to population. What causes the transition to the quasi-stable coexistence phase? What are the mechanisms that sustain it? What is it affected by?

There seem to be no answer to these questions so far, the authors suggest that a crucial role could be played by negative frequency-dependent selection (removal of deleterious alleles that depends on the current fraction of such alleles) or coupling between ecologically divergent phenotypes and fitness gain (which entails interaction at the phenotype layer between individuals and environment, projected in the genotype space through the unknown genotype-phenotype map).

Now let us look back at what we have described in this chapter and in Ch.(1). Is it not the coupling environment-selection after all what we encode when using a selection based on a fitness function? Is it not true that our NRC phase has a similar behaviour for the allele frequencies? Could such behaviour of the E.coli be induced or sustained by the same fundamental mechanisms that induce or sustain the NRC phase in our simulations with \( r = 0 \) (yet, unknown)? We do not expect the NRC phase to grasp as a whole the complexity of the phenomena of which Fig.(5.13) is a footprint, but simply (in the best of possible worlds) its fundamental nature, in a much rougher fashion.

Is it an explanation? It is not. Is it an hypothesis? Not even that, there are too few (if any) theoretical and experimental evidences to give it such a dignity. Up to now, it is speculation, a conjecture, a proto-hypothesis, could it become hypothesis, could it be a mere coincidence: future works, as usual, will give verdict.

Figure 5.13: Allele frequency trajectories \( \nu_i \in [0, 1] \) of all de novo mutations detected in the 12 LTEE populations. Note that in the framework of this chapter \( \nu_i = (1 - \chi_i) / 2 \). From [2017].
5.3 SPECTRUM

Let us briefly recap the state of the art. In Ch.(4) we had found that for high enough mutation rate and selection strength, reconstruction fails. This failure is due the system undergoing a transition to what we have named NRC (Non-Random Coexistence) phase, where the QLE symmetry is broken, e.g. $|\chi_i| \sim \alpha \neq 0$. Several observations have been pointed out for the case where there is recombination and selection is based on a SK-fitness function, Sec.(5.1.2); some preliminary variations on the theme have been explored in Sec.(5.1.3). Looking for useful insights, we have exploited some results from literature in an effort of developing intuition, Sec.(5.2.1 - 5.2.2) and some further analysis has been carried out in Sec.(5.2.3). Finally, we have speculated on the biological relevance of our results, based on a recently published experimental study in Sec.(5.2.4).

Although we have learned a lot about it, the NRC phase remains obscure in many respects, which is reflected in the number of open questions that we have been pointed out throughout the chapter. In particular, there is little understanding on which are the drivers of the transitions QLE$\leftrightarrow$NRC and which mechanisms sustain the NRC phase. No scientifically relevant prediction can be built on these results until a deeper theoretical understanding is provided, we strongly believe that future efforts should focus in this direction.

Finally, recalling the discussion in Sec.(2.3), what if all of this is nothing but the result of a bug in the code? If not a bug, what if these results are code-dependent, in the sense that they are due to the specific choices of the algorithms implemented? These questions are perfectly legitimate, after all we are using as a simulation tool an external software, FFPopSim which, albeit widely used in the last years, could have issues simply not yet spotted or reported in literature (at least, to our knowledge). In absence of further evidences, the whole spectrum of possibilities should be given the same dignity, from the outstanding biological relevance to the possibility of an hidden code bug.
COVID-19 would not need any introduction. The pandemic disease is caused by the Severe Acute Respiratory Syndrome CoronaVirus 2 (SARS-CoV-2), a single-stranded RNA beta-coronavirus of the family Coronaviridae. According to the World Health Organization (WHO) and up to September 2020, more than 26,000,000 infections and over 800,000 deaths. Directly or indirectly, it has swept away any notion of “normality” in us all, undermined the foundation of all social structures and human activities, called into question the status quo and forced everyone at any scale to look for new answers, to react.

Like any other group of human beings, science has been shaken, as could be expected for such a closely networked world in a time of restricted mobility, of borders and only-virtual interactions. Most of the worldwide famous scientific journals have granted open access to Covid-related publications (a clear hallmark of an apocalypse) and loosened the meshes of the peer-review process to speed up the production and sharing of new results: the quest for treatments and a vaccine for the SARS-CoV-2 undoubtedly is a priority and time is a crucial variable. This has led in few months to an astonishing quantity of published papers (more than 37,000 on September 2020), so many that even a browser has been created with the only purpose of bringing order in this crowd: COLLABOVID, https://www.collabovid.org/.

Such huge number of papers can be understood in light of the fact that, aside from the already-insiders (epidemiologists, virologists...) a number of labs and scientists, that maybe were addressing different questions but had in their hand useful tools, tried to contribute to the cause within their abilities, in a worldwide effort that has involved biologists, data scientists, mathematicians and, of course, physicists. Although one of the consequences is an increased background noise, this convergence could be crucial to reduce the time-scale of the quest for a response (treatment and/or vaccine), which so far seems to be the only possible way out.

As far as we are concerned, in our toolbox we have a technique, the Direct Coupling Analysis (DCA), that is suitable for extracting information from biological data [Cocco et al., 2018; Gao et al., 2019] and has already been used for predicting the effects of combinatorial antibiotic treatments [Nguyen et al., 2017], albeit in the case of bacteria. We dedicate this chapter to our contribution to the Covid cause, which is aimed at detecting epistasis in the SARS-CoV-2 genome, starting from sequences available at the time of our research. Upon refining raw data, we will discuss the reasonableness of running DCA and hence implement PLM; from the resulting strongest couplings, we will try to highlight those due to epistasis and finally discuss their potential usefulness. The content of the whole chapter is a summary of our recent work [Zeng, Dichio, et al., 2020] and we refer to it (and references therein) for a more extensive treatment.

1 https://www.who.int/, Situation Reports are uploaded daily.
6.1 DATA JOURNEY FROM BIO LABS TO DCA SCORES

As data analysts in an institute for theoretical physics, we are not able to follow the flow of the information from its very source i.e. cells infected by the SARS-CoV-2. Our starting point is instead a large database of genomic sequences collected all over the world by bio labs that soon after the outbreak of the epidemic started sequencing the virus, in a massive unprecedented effort. The rapid sharing and availability of these data has been made possible by exploiting some pre-existing tools, developed in the last years for studying other epidemics e.g. influenza, Ebola, Zika, measles...

Among the most remarkable, the GISAID initiative [Shu & McCauley, 2017], https://www.gisaid.org/, provides (upon registration) a free-access repository of SARS-CoV-2 sequences, which has been ceaselessly feeded with new data at a growing rate, at the same pace as the spread of the virus. In September 2020 more than 95,000 sequences are available. For our analysis, we have used four dataset, we shall call them DS-1.4, 2.268 sequences downloaded on the 1/4/2020; DS-8.4, 3,490 sequences, 8/4/2020; DS-2.5, 10,587 sequences, 2/5/2020; DS-8.8, 51,676 sequences, 8/8/2020. For our discussion, let us focus on the last one, DS-8.8.

In order to understand which kind of data we download from GISAID, let us consider as an example the original sequence of the SARS-CoV-2 [Wu et al., 2020], that we will later use as reference. It is referred as Wuhan-Hu-1 and in FASTA format this is how it looks like:

```
>hCoV-19/Wuhan/Hu-1/2019|EPI_ISL_402125|2019-12-31
ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCT
```

where the first line is an heading and the second is a string of L characters, where L~30,000 for the SARS-CoV-2. Each character is a standard IUPAC code, the most frequent being A,C,G,T (nucleobases), N (any base), - (gap).

If M is the number of sequences in a dataset, the first task is to build an MSA matrix out of them (rows = sequences, columns = loci). This is not at all a trivial task: we cannot simply stack genotypes because in general they have different lengths. Developing efficient criteria and methods to transform these data in a set of sequences with the same length is a challenging task in itself; fortunately, we can again exploit pre-existing tools, in particular MAFFT, [Katoh et al., 2017] https://mafft.cbrc.jp/alignment/software/. As a result, we get M sequences with L bases each, again in FASTA format.

6.1.1 MSA Refinement

In a preliminary step for our analysis, we cast the sequences from the last step in a M × L machine-readable MSA, discarding the headings and stacking genotypes. For simplicity, we

---

2 See https://en.wikipedia.org/wiki/Nucleic_acid_notation. Besides those already mentioned, other symbols may be found, for instance W = A or T, B = C or T or G... This is due to the uncertainty of the sequencing output.
also convert any character which is different from -NACGT i.e. KFY... into N (a fraction around $10^{-6}$ of the total number of entries), so that we are left with $q = 6$ states. Formally,

$$\text{MSA} = \{ s_i^m \mid i = 1, \ldots, L; m = 1, \ldots, M \}.$$  

In principle, now that we have at our disposal the MSA, we can initialize a DCA algorithm and look for relevant couplings among loci. However, considering the size our matrix, running DCA at this stage would entail a tremendous computational effort, and would not be even wise: typically, only a very small fraction of all the loci in a genome shows substantial variability i.e. that emerges from the noise of sequencing errors. We can take this into account by filtering the loci, therefore we prune the ensemble of loci with the following criteria: a locus is discarded if either the same character is found in more than the $96.5\%$ of the sequences or the sum of the percentages of ACGT is less than the $20\%$. Such a filter drastically reduces the number of loci under investigation to a number that for the DS-8.8, we find to be 689.

One last refinement is useful. A common problem with such data is that they are never random samples from the global population [MacLean et al., 2020]: groups of sequences are biased e.g. by the country they occur in, by contact tracing... In order to mitigate these effects, we can adopt a standard re-weighting procedure. Let $x \in [0, 1]$ be a preassigned threshold, we define two sequences $s^{(a)}, s^{(b)}$ to be similar if $\text{dist}(s^{(a)}, s^{(b)}) \geq x$, where $\text{dist}$ is the Hamming distance between the two sequences, normalized over $L$. For each sequence $s^{(a)}$, we can define the weight $w_a$ from:

$$w_a^{-1} = \left| \{ b \in \{ 1, \ldots, M \} : \text{dist}(s^{(a)}, s^{(b)}) \geq x \} \right|,$$

i.e. $w_a$ is the inverse of the number of sequences in the MSA which are similar to $\sigma^a$. Attaching these weights to sequences will penalize large groups of similar sequences and smooth out their impact on the inference procedure. Different choices for $x$ are possible, as discussed in [Morcos et al., 2011], the final results are weakly dependent on the threshold if $x \in [0.7, 0.9]$, while setting $x = 1$ entails considering only unique sequences in the dataset: we will adopt this latter choice and implement the filter for each dataset.

### 6.1.2 PLM Inference

In this section we will feed our refined MSAs to a DCA algorithm to try to infer relevant couplings between loci. But for this to be meaningful, as we know from Ch.(3), the data must be random samples from an Ising-Potts distribution Eq.(6.2): why should this be the case for the SARS-CoV-2?

$$P(\sigma) = \frac{1}{Z} e^{-\sum_i h_i(\sigma_i) + \sum_{ij} J_{ij}(\sigma_i, \sigma_j)}$$  

3 In order to quantify this statement, one can introduce the so called effective number of sequences $B_{\text{eff}} = \sum w_a$, which depends on $x$. For a discussion, see [Zeng, Dichio, et al., 2020], where also the dependence on the number of states $q$ is considered. As an example, for the dataset DS-8.8, setting $q = 6, x = 1.0$, we have $B_{\text{eff}} = 19708$. 

Researchers have long known that coronaviruses in general exhibit a large amount of recombination [Lai & Cavanagh, 1997]; based on this evidence, we assume that the distribution of genotypes in the viral population of the SARS-CoV-2 is in the state of Quasi-Linkage Equilibrium, therefore, as we know from Sec. (1.3), the distribution underlying the samples in our MSA is assumed to be an Ising-Potts distribution Eq. (6.2).

The consequences of this assumption go much beyond the possibility of running DCA: according to the KNS-theory, inferring the couplings implies, up to some factors, inference of the fitness landscape; this, in turn, paves the way for the application of our results to the development of drugs/vaccines, as we will discuss in Sec. (6.2). Let us remark that unfortunately a number of phenomena in real data can contaminate the randomness of our samples, we will take this into account (at least partially) in Sec. (6.1.3).

Now that we are convinced of the meaning of running DCA on our dataset, we can choose an algorithm and carry out the task. We choose the Pseudolikelihood Maximization PLM, introduced in Sec. (3.3), as implemented in [Gao et al., 2019]. Differently from the previous chapters, we will run DCA inference for a Potts model with multiple states \( q > 2 \) but the generalization of the procedure from a formal point of view is straightforward. For simplicity, we preliminary transform \(-N\text{ACGT} \rightarrow \{0,1,2,3,4,5\}\), the probability of \( s_i \) at the locus \( i \) conditioned on the sequence \( s_{\setminus i} \) without the \( i \)-th locus is then:

\[
P(s_i | s_{\setminus i}) = \frac{\exp \left( h_i (s_i) + \sum_{j \neq i} J_{ij} (s_i, s_j) \right)}{\sum_u \exp \left( h_i (u) + \sum_{j \neq i} J_{ij} (u, s_j) \right)},
\]

where \( u = \{0,1,2,3,4,5\} \) are the possible states of locus \( i \), cf Eq. (3.35). According to the PLM recipe, the inferred parameters are found by maximizing the corresponding log-likelihood function:

\[
\mathcal{L}_D^i (J_{si}, h_i) = h_i (s_i) + \sum_{j \neq i} J_{ij} (s_i, s_j) - \left( \log \sum_u \exp \left( h_i (u) + \sum_{j \neq i} J_{ij} (u, s_j) \right) \right).
\]

All other remarks on the PLM hold true as introduced in Sec. (3.2 - 3.3); in particular, we have used an asymmetric version of the PLM with \( L^2 \)-regularization parameter \( \lambda_2 = 0.1 \). As a result, for each pair of loci \( i \neq j \) we get a \( q \times q \) matrix \( J_{ij} \) and we score the strength of the coupling between \( i, j \) by the Frobenius norm of \( I_{ij} \), Eq. (3.41).

Intuitively we expect two loci be bound by a strong coupling if every time there is a mutation in one of them, the other mutates as well. The more this co-mutation is systematic, the tighter their bond.

Let us consider DS-8.8. An ordered list of the strongest coupled pair of loci can be prepared, and a number of useful information can be extracted from it, by comparison with the reference sequence, see Tab. (6.1): for instance, for each identified locus, we can trace back to which coding region it belongs to (i.e. which protein it codes for), which is the corresponding mutation.
A number of contributions have appeared recently that analyze the genomic structure of the SARS-CoV-2 and its molecular biology: from popularizing articles [https://www.nytimes.com/interactive/2020/04/03/science/coronavirus-genome-bad-news-wrapped-in-protein.html] to academic papers like [Gordon et al., 2020], from which Fig. (6.1) is taken.

At this stage Tab. (6.1) should be read with caution: if two loci were biologically bounded by a strong coupling then we would expect to find them in that table, but the converse is.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Extremum i locus - protein</th>
<th>Extremum i mutation - type</th>
<th>Extremum j locus - protein</th>
<th>Extremum j mutation - type</th>
<th>PLM score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1059 - nsp2</td>
<td>C</td>
<td>T - non.</td>
<td>25563 - ORF3a</td>
<td>G</td>
</tr>
<tr>
<td>2</td>
<td>28882 - N</td>
<td>G</td>
<td>A - syn.</td>
<td>28883 - N</td>
<td>G</td>
</tr>
<tr>
<td>3</td>
<td>28881 - N</td>
<td>G</td>
<td>A - non.</td>
<td>28882 - N</td>
<td>G</td>
</tr>
<tr>
<td>4</td>
<td>28881 - N</td>
<td>G</td>
<td>A - non.</td>
<td>28883 - N</td>
<td>G</td>
</tr>
<tr>
<td>5</td>
<td>8782 - nsp4</td>
<td>C</td>
<td>T - syn.</td>
<td>28144 - ORF8</td>
<td>T</td>
</tr>
<tr>
<td>6</td>
<td>14805 - nsp12</td>
<td>C</td>
<td>T - non.</td>
<td>26144-ORF3a</td>
<td>G</td>
</tr>
<tr>
<td>7</td>
<td>3037 - nsp3</td>
<td>T</td>
<td>C - syn.</td>
<td>14408 - nsp12</td>
<td>T</td>
</tr>
<tr>
<td>8</td>
<td>18877 - nsp14</td>
<td>C</td>
<td>T - syn.</td>
<td>25563 - ORF3a</td>
<td>G</td>
</tr>
<tr>
<td>9</td>
<td>3037 - nsp3</td>
<td>T</td>
<td>C - syn.</td>
<td>23403 - S</td>
<td>G</td>
</tr>
<tr>
<td>10</td>
<td>14408 - nsp12</td>
<td>T</td>
<td>C - syn.</td>
<td>23403 - S</td>
<td>G</td>
</tr>
<tr>
<td>11</td>
<td>1059 - nsp2</td>
<td>C</td>
<td>T - non.</td>
<td>18877 - nsp14</td>
<td>C</td>
</tr>
<tr>
<td>12</td>
<td>17858 - nsp13</td>
<td>A</td>
<td>G - non.</td>
<td>18060 - nsp14</td>
<td>C</td>
</tr>
<tr>
<td>13</td>
<td>17747 - nsp13</td>
<td>C</td>
<td>T - non.</td>
<td>17858 - nsp13</td>
<td>A</td>
</tr>
<tr>
<td>14</td>
<td>17747 - nsp13</td>
<td>C</td>
<td>T - non.</td>
<td>18060 - nsp14</td>
<td>C</td>
</tr>
<tr>
<td>15</td>
<td>11083 - nsp6</td>
<td>G</td>
<td>T - non.</td>
<td>26144 - ORF3a</td>
<td>G</td>
</tr>
<tr>
<td>16</td>
<td>20268 - nsp15</td>
<td>A</td>
<td>G - syn.</td>
<td>25563 - ORF3a</td>
<td>G</td>
</tr>
<tr>
<td>17</td>
<td>11083 - nsp6</td>
<td>G</td>
<td>T - non.</td>
<td>14805 - nsp12</td>
<td>C</td>
</tr>
<tr>
<td>18</td>
<td>11083 - nsp6</td>
<td>G</td>
<td>T - non.</td>
<td>28144 - ORF8</td>
<td>T</td>
</tr>
<tr>
<td>19</td>
<td>8782 - nsp4</td>
<td>C</td>
<td>T - syn.</td>
<td>11083 - nsp6</td>
<td>G</td>
</tr>
</tbody>
</table>

Table 6.1: Top-200 significant couplings between loci of the SARS-CoV-2 for the DS-8.8. Links that involve noncoding sites in the 5' or 3' region of the Wuhan-Hu-1 have been discarded. For each link and for each of its extrem we provide the following information: 1. locus: its localization in the reference sequence Wuhan-Hu-1; 2. protein: which protein it codes for; 3. mutation: which are the major/minor alleles of the corresponding mutation; 4. type: whether the it is a synonymous (syn.) or non-synonymous (non.) mutation. As an example, "1059-nsp2, C|T -non." stands for a non-synonymous mutation at the locus 1059 of the genomic sequence of the SARS-CoV-2, in the coding region for the protein nsp2; the major allele in that locus is cytosine (C), thymine (T) the minor.

visualize these top links is the circular plot shown in Fig. (6.2). Analogous tables and plots can be generated for the data-sets DS-1.4, DS-8.4, DS-2.5. The reason why we have so far processed the four dataset in parallel is to test whether or not our predictions of on the top-scored links are stable over time: in [Zeng, Dichio, et al., 2020] (and references therein) we prove that not only this is true, but also that this conclusion dramatically changes if one uses correlations in place of couplings: no reliable interaction can be retrieved in the latter case, which further motivates our DCA-approach to the analysis.
Figure 6.1: Genomic organization of SARS-CoV-2. The genomic sequence consists of $\sim 30$ Kb RNA strand, it encodes for both structural and non-structural proteins. The colour intensity is proportional to the protein sequence similarity with SARS-CoV homologues. From [2020].

not true, not all (if any of) the links that appear as coupled after running DCA are necessarily bound together by biology. The most important issue has to do with the assumed randomness of the samples and before drawing any conclusion we must deal with it: in the next section we will precisely discuss phylogeny.

6.1.3 Randomization Sieve

We refer to [Horta & Weigt, 2020] as a main reference for this section, where a exhaustive explanation can be found of the methods that will be employed here. As a first step, let us explain what phylogeny is and why we must take care of it.

A phylogenetic tree $L$ is a branching diagram showing the evolutionary interrelations of a group of organisms derived from a common ancestor. An example is shown in Fig. (6.3): sequences observable today are the leaves of the tree (ABCDE), the common ancestor is the root. Branching points correspond to events that can differentiate two sequences e.g. emergence of a new mutation, geographical separation of two groups of individuals etc. On distinct branches, sequences are assumed to evolve independetly.

In other words, a phylogenetic tree is a reconstruction of the past history of the data that we observe now and, since these data are always finite, the resulting trees have to be regarded as hypotheses, probabilistic statements. Such graphs have soon appeared for the SARS-CoV-2: for a up to date version, we refer to the remarkable visualizations developed by the team of Nextstrain [Hadfield et al., 2018], https://nextstrain.org/. How to draw these graphs is a interesting question in itself, but orthogonal to our direction of investigation: we only mention one class of such techniques, the so called distance-based methods, whose input information is a matrix of pairwise genetic (Hamming) distances.

The very existence of phylogenetic trees is in sharp contrast with our assumption on the ran-
Figure 6.2: Top-100 significant pairwise epistasis from the D5-8.8 between loci in coding regions of the SARS-CoV-2. Blue lines indicate for top 50, red lines for short distance links (equal or less than 3 base pairs), grey lines top 50-200. Red lines show short-distance links (less or equal to 3 bases) blue lines show links of longer distance. The coloured links are the same pairs as listed in Table (6.1). See [2020] and references therein.

**Dom samples**, in fact the resulting statistics is biased\(^5\) and the DCA inference will will lead to fields and couplings that are due to phylogeny instead of biological coupling. Specifically, the DCA recipe fails to be accurate at its very beginning, when assuming the factorization of the likelihood function \(L\) in a product probabilities of the different samples \(\{s^a\}\), Eq.(3.37). This is not true if the latter are not i.i.d. and in general when considering explicitly the dependence of the likelihood function on the phylogenetic tree \(T\) one has:

\[
\mathcal{L}(\{h_{ij}\},T) \neq \mathcal{L}_{iid}(\{h_{ij}\}) = \prod_a P(s^a).
\]

\(^5\) For instance, let us look again at Fig.(6.3) and suppose that the branch labelled as C has become extinct by chance before any sequence had been collected. In A,B and in their descendants (until further mutations) we would always observe the yellow-green mutations appearing together, just because they have appeared in the same branch of the tree. This, however, would imply a strong correlation between them that cannot be ascribed to any underlying biological mechanism!
Our strategy to take these effects into account aims at pruning the couplings listed in Tab. (6.1) for DS-8.8 to those that are due to biological epistasis and not generated by phylogenetic effects. We do this by comparing the results that we have got so far with those obtained by randomized background, according to the two following tests:

- **Profile randomization.** Each of the columns of the MSA is reshuffled so that all kinds of correlations are destroyed while only the single site statistics (frequencies) are conserved. Let MSA-PRO be the resulting multiple alignment matrix. We expect none of the couplings in Tab. (6.1) to survive, this is a check test.

- **Phylogeny randomization.** Columns and rows of the original MSA are changed simultaneously so that inter-sequence distances are held fixed, this amounts to preserve the phylogenetic tree that underlies the data. Let MSA-PHY be the resulting multiple alignment matrix. We expect only the couplings due to phylogeny in Tab. (6.1) to survive the reshuffling.

In [Zeng, Dichio, et al., 2020], each randomization strategy is run 50 times, each time we consider the top couplings obtained and compare with Tab. (6.1). As expected, from none of the 50 MSA-PROs is possible to significantly recover any of the couplings listed in Tab. (6.1). Many of them on the contrary show up with significant scores when running DCA in 24 out of 50 of the MSA-PHYs: we will discard these couplings for the subsequent analysis. Interestingly, at least eight of the couplings in Tab. (6.1) stand out and survive the phylogeny test. They are listed in Tab. (6.2), together with additional information on which amino acid mutations they are associated to.

### 6.2 Results: Spotlight on Strong Couplings

The final outcome of our analysis is the content of Tab. (6.2). Let us recap how we got there: we started from whole-genome sequences of the SARS-CoV-2 from the GISAID repository (DS-1.4,
Table 6.2: Top-200 significant epistatic couplings between loci of the SARS-CoV-2 for the DS-8.8. Links that involve noncoding sites in the 5' or 3' region of the Wuhan-Hu-1 have been discarded. Links that are due to phylogenetic effects have been discarded, too. For each link and for each of its extrema we provide the following information: 1. locus: its localization in the reference sequence Wuhan-Hu-1; 2. protein: which protein it codes for; 3. mutation: which are the major/minor alleles of the corresponding mutation; 4. type: whether it is a synonymous (syn.) or non-synonymous (non.) mutation 5. amino acid mutation: which amino acid mutation is induced by the base change. As an example, "1059-nsp2, C|T-non., T85I" stands for a non-synonymous mutation at the locus 1059 of the genomic sequence of the SARS-CoV-2 in the coding region for the protein nsp2, with major allele cytosine (C) and minor thymine (T), that causes a mutation of the 85th amino acid of that protein from Threonine (T) to Isoleucine (I).

<table>
<thead>
<tr>
<th>Rank</th>
<th>Extremum i mutation - type</th>
<th>Extremum i locus - protein</th>
<th>Amino acid mutation</th>
<th>Extremum i mutation - type</th>
<th>Extremum i locus - protein</th>
<th>Amino acid mutation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1059 - nsp2</td>
<td>C</td>
<td>T - non.</td>
<td>T85I</td>
<td>G</td>
<td>T - non.</td>
</tr>
<tr>
<td>5</td>
<td>8782 - nsp4</td>
<td>C</td>
<td>T - syn.</td>
<td>S76S</td>
<td>T</td>
<td>C - non.</td>
</tr>
<tr>
<td>9</td>
<td>14805 - nsp12</td>
<td>C</td>
<td>T - non.</td>
<td>T442I</td>
<td>G</td>
<td>T - non.</td>
</tr>
<tr>
<td>21</td>
<td>1059 - nsp2</td>
<td>C</td>
<td>T - non.</td>
<td>T85I</td>
<td>C</td>
<td>T - syn.</td>
</tr>
<tr>
<td>26</td>
<td>17858 - nsp13</td>
<td>A</td>
<td>G - non.</td>
<td>T541C</td>
<td>C</td>
<td>T - syn.</td>
</tr>
<tr>
<td>27</td>
<td>17747 - nsp13</td>
<td>C</td>
<td>T - non.</td>
<td>P504L</td>
<td>A</td>
<td>G - non.</td>
</tr>
<tr>
<td>36</td>
<td>17747 - nsp13</td>
<td>C</td>
<td>T - non.</td>
<td>P504L</td>
<td>C</td>
<td>T - syn.</td>
</tr>
<tr>
<td>47</td>
<td>11083 - nsp6</td>
<td>G</td>
<td>T - non.</td>
<td>L37F</td>
<td>26144 - ORF3a</td>
<td>26144 - ORF3a</td>
</tr>
</tbody>
</table>

DS-8.4, DS-2.5, DS-8.8): after preprocessing the data, we run DCA to infer the strongest couplings. By comparing the latter with those inferred from randomized backgrounds (profile and phylogeny) we tried to isolate those that can be reasonably ascribed to epistasis. Our conclusions are stable in the time window under observation, which is not the case if bare correlations are used in the place of couplings, see Fig.(4) in [2020].

A word of caution is needed since it may be possible in principle that some phylogenetic effects have escaped our randomization sieve and are still contaminating our results in subtler ways. New data available and possibly different ways of accounting for phylogeny will shed some light on this issue.

A discussion on the role of the proteins highlighted in Tab.(6.2) for the molecular biology of the SARS-CoV-2 is by far outside the domain of this Thesis. Therefore our tracking of the data journey will not go beyond, here is the limes of what we, as physicists, can confidently state: we can but pass the torch to biologists for the next, crucial step. It is important nevertheless to place this work into the context of the existing literature and to understand how these data can be useful.

Several previous contributes have indeed turned the spotlight on the loci of the SARS-CoV-2 genome that show a non trivial variability (i.e. not due to sequencing errors), even at very early stages of the epidemics, with few data available. As an example, [Pachetti et al., 2020] isolated and characterized 8 such loci from a dataset of only 220 sequences and used these information to argue that European, North American and Asian strains of the virus might coexist, each of them characterized by a different mutation pattern. Our focus instead has been on biologically coupled variability and for this to be true for two loci: 1. they both have to show a relevant
amount of mutations; 2. their mutation pattern should be coupled; 3. this coupling should be due to epistasis and not to phylogeny or other factors.

Epistatic interactions are important if they can be modulated by pharmacological intervention both for the development of new drugs or for the combinatorial treatments with existing drugs: this is the next, crucial step aforementioned. For instance, suppose there are drugs that act on targets around both loci, modulating the fitness of the respective variants. Epistasis then points to the possibility that using both drugs simultaneously may have a more than additive effect. We can give some more concreteness to this point by scanning [Gordon et al., 2020], where authors identify 332 high-confidence protein-protein interactions, yielding 66 candidate druggable human proteins or host factors targeted by either existing FDA-approved or investigational drugs. Five out of eight pairs in Tab.(6.2) involve either synonymous mutations or coupling in the same gene. Focusing on the most interesting pairs of far-apart non-synonymous mutations (1059, 25563) and (11083, 26144) and (14805, 26144) we find that the second locus lies in the coding region for the ORF3a, for which no potential drugs are listed in [2020]; on the contrary, one or more already approved and practical drugs exist for nsp2 and nsp6, at one extremum of the first and last pair above. Our results could be useful in indicating future direction of investigation e.g. rapidly suggesting and/or assessing new combinatorial treatments (considered the combinatorial explosion of possibilities); if successful, they would lead eventually to a relevant optimization of resources and above all the most important: time.
A | MOMENTS VS CUMULANTS

This appendix is aimed at introducing and discussing the meaning of moments and cumulants of a probability distribution. We shall also discuss the specific case of the Gaussian distribution. Most of the following material is drawn from [Gardiner, 2004].

Let \( X = (X_1, \ldots, X_n) \) be a vector of random variables with pdf \( p(x) \), then the characteristic function is

\[
\phi(q) = \int p(x)e^{iqx} \, dx.
\]  

(A.1)

with \( \phi(0) = 1, |\phi(q)| \leq 1. \phi(q) \) completely is a characterization of the probability distribution \( p(x) \) i.e. it completely determines its behaviour and properties. If the raw moments \( \langle \prod_i X_i^{m_i} \rangle \) exist, then

\[
\langle \prod_i X_i^{m_i} \rangle = \left[ \prod_i \left( -i \frac{\partial}{\partial q_i} \right)^{m_i} \phi(q) \right]_{q=0}.
\]  

(A.2)

Moreover, it is possible to define the cumulant generating function (the reason for this name will be clear soon) as

\[
\psi(q) = \log \phi(q).
\]  

(A.3)

Suppose that it is possible to expand \( \phi(q) \), hence \( \psi(q) \) in power series about the origin (i.e. all derivatives exist); then we can formally write:

\[
\phi(q) = \sum_{r=1}^{\infty} \frac{i^r}{r!} \sum_{\{m\}} \langle X_1^{m_1} \cdots X_n^{m_n} \rangle \frac{q_1^{m_1} \cdots q_n^{m_n}}{m_1! \cdots m_n!} \delta \left( r, \sum_{i=1}^n m_i \right),
\]  

(A.4)

\[
\psi(q) = \sum_{r=1}^{\infty} \frac{i^r}{r!} \sum_{\{m\}} \langle\langle X_1^{m_1} \cdots X_n^{m_n} \rangle\rangle \frac{q_1^{m_1} \cdots q_n^{m_n}}{m_1! \cdots m_n!} \delta \left( r, \sum_{i=1}^n m_i \right),
\]  

(A.5)

where \( \langle X_1^{m_1} \cdots X_n^{m_n} \rangle \) are defined in Eq.(A.2) and \( \langle\langle X_1^{m_1} \cdots X_n^{m_n} \rangle\rangle \) are the cumulants of the distribution \( p(x) \). Both these two set of quantities stem from the same characteristic function \( \phi(q) \) and are alternative parametrizations of the probability distribution. Expanding and comparing Eq.(A.4, A.5) one finds:

\[
\langle\langle X_i \rangle\rangle = \langle X_i \rangle
\]

\[
\langle\langle X_i X_j \rangle\rangle = \langle X_i X_j \rangle - \langle X_i \rangle \langle X_j \rangle
\]

i.e. the first two cumulants are the means and covariances (second central moments). No general formula exists for the explicit expression of the cumulants but a diagrammatic method is suggested in [2004]. There is no room here for an in dept discussion on the importance of
moments and cumulants and on why to bother at all with introducing two same sets of quantities that encode basically the same information; we refer to [Gnedenko, 1998] for an extensive treatment from the probability-theory point of view. Nevertheless, as far as we are concerned in this work, a crucial difference deserve our attention.

Higher order cumulants contain information of decreasing significance. On the contrary, we have \( \langle X^{2n} \rangle \geq \langle X^n \rangle^2 \) and thus all moments contain information about lower moments [Gardiner, 2004]. Interestingly but not surprisingly, when drawing parallels between statistical physics and the path-integral formulation of field theories, one finds that the moments \( \langle X_1 \ldots X_n \rangle \) are the \( n \)-points correlation functions while the \( \langle \langle X_1 \ldots X_n \rangle \rangle \) are the \( n \)-points connected correlation functions i.e. the sum over all \( n \)-points 1PI Feynman diagrams, the key ingredients of such theories. [Peskin & Schroeder, 1995; Zinn-Justin, 2010].

### A.1 Example: Gaussian Case

Let us specify the above discussion to the case of a multivariate Gaussian probability distribution:

\[
p(x) = \frac{1}{Z} e^{-\frac{1}{2}(x-\mu)^T C^{-1}(x-\mu)},
\]

where \( \mu \) is the mean and \( C \) is the covariance matrix, \( Z = [(2\pi)^n \det(C)]^{-\frac{1}{2}} \) is the normalization. The characteristic function of such distribution is

\[
\phi(q) = e^{i q^T \mu - \frac{1}{2} q^T C q}
\]

and from this we deduce that all cumulants of order \( \geq 2 \) vanish. Therefore, we expect all raw moments to be expressed in terms only of first and second order cumulants. Finally, because of they are of interest for this work, let us evaluate the first four raw moments of this distribution using Eq.(A.2) (different indices).

\[
\langle X_i \rangle = -i \frac{\partial}{\partial q_i} \phi(q) \bigg|_{q=0} = \mu_i \\
\langle X_i X_j \rangle_{i \neq j} = (-i)^2 \frac{\partial}{\partial q_i} \frac{\partial}{\partial q_j} \phi(q) \bigg|_{q=0} = C_{ij} + \mu_i \mu_j
\]
\[(X_i X_j X_k)_{i \neq j \neq k} = (-i)^3 \frac{\partial}{\partial q_k} \frac{\partial}{\partial q_j} \frac{\partial}{\partial q_i} \phi(q) \bigg|_{q=0} = i \left[ -C_{jk} (i \mu_j - \sum_n C_{jm} q_n) - C_{ik} (i \mu_i - \sum_n C_{in} q_n) \\
- C_{ij} (i \mu_k - \sum_n C_{kn} q_n) + (i \mu_k - \sum_n C_{kn} q_n) \times (i \mu_j - \sum_n C_{jn} q_n) \right] e^{iq^T \mu - \frac{1}{2} q^T C q} \bigg|_{q=0} = \mu_i C_{jk} + \mu_j C_{ik} + \mu_k C_{ij} + i \mu_i \mu_j \mu_k \times \times (i \mu_j - \sum_n C_{jn} q_n) \times (i \mu_k - \sum_n C_{kn} q_n) \times (i \mu_i - \sum_n C_{in} q_n) \right] e^{iq^T \mu - \frac{1}{2} q^T C q} \bigg|_{q=0} \]

\[\langle X_i X_j X_l X_k \rangle_{i \neq j \neq k \neq l} = (-i)^4 \frac{\partial}{\partial q_l} \frac{\partial}{\partial q_k} \frac{\partial}{\partial q_j} \frac{\partial}{\partial q_i} \phi(q) \bigg|_{q=0} = \{ C_{ij} C_{kl} + C_{ik} C_{jl} + C_{il} C_{jk} (i \mu_j - \sum_n C_{jn} q_n) (i \mu_i - \sum_n C_{in} q_n) + \\
+ C_{jl} (i \mu_k - \sum_n C_{kn} q_n) (i \mu_i - \sum_n C_{in} q_n) + C_{il} (i \mu_j - \sum_n C_{jn} q_n) \times (i \mu_k - \sum_n C_{kn} q_n) + \right] e^{iq^T \mu - \frac{1}{2} q^T C q} \bigg|_{q=0} \]

\[= C_{ij} C_{kl} + C_{ik} C_{jl} + C_{il} C_{jk} + C_{ij} \mu_k \mu_l + C_{ik} \mu_j \mu_l + C_{il} \mu_j \mu_k + C_{jk} \mu_i \mu_l + \mu_i \mu_j \mu_k + C_{il} \mu_j \mu_k \mu_l = 1 \text{ EXAMPLE: GAUSSIAN CASE} | 99 \]
We dedicate this appendix to a brief introduction to the simplest models of genetic drift, named after the seminal works of R.A. Fisher [Fisher, 1930] and S. Wright [Wright, 1931]: we will use the shorthand FW-model. For more information, see for instance [Blythe & McKane, 2007; Messer, 2016; R. A. Neher & Walczak, 2018].

Let us step back from the model described in Ch.(1) and forget about recombination and mutations. As a starting point, we also assume that selection is absent and stochastic effects alone drive the evolution: we say these are neutral models.

In a FW model, we have an haploid population of constant size $N$. Evolution is encoded in a series of discrete, non-overlapping generations. Each time step $t \rightarrow t+1$ entails a replacement of the entire population, with a process akin to the sampling with replacement of coloured balls from an urn, this is the so called beanbag population genetics, see Fig.(B.1).

**Figure B.1:** Beanbag population genetics. The population at $t+1$ is constructed from that at $t$ (size $N$) by (i) selecting a gene from $t$ at random; (ii) copying this gene; (iii) placing the copy in the next generation; (iv) returning the original to the parent population. This algorithm is repeated until until the population at generation $t+1$ has size $N$. From [2007].

We can easily quantify the last statement as follows: consider a biallelic ($\pm 1$) population and pick a locus $\alpha \in 1, \ldots, L$, suppose that at time $t$ an allele (+1, say) for the locus $\alpha$ is present in $j$ out of $N$ individuals, then the probability that at time $t+1$ it will be found $i$ times is

$$p_{ij} = \binom{N}{i} \left( \frac{j}{N} \right)^i \left( 1 - \frac{j}{N} \right)^{N-i}, \quad i, j \in \{1, \ldots, N\}. \quad (B.1)$$

In words, the probability of drawing an allele from the "urn" of the genetic pool depends at each "drawing" from the current frequency of that allele in the population. The natural generalization of Eq.(B.1) to the case where the population has $L$ loci is the multinomial distribution:

$$p_{ij} = N! \prod_{\alpha=1}^{L} \frac{1}{i_{\alpha}!} \left( \frac{j_{\alpha}}{N} \right)^{i_{\alpha}}, \quad i_{\alpha}, j_{\alpha} \in \{1, \ldots, N\} \forall \alpha, \quad (B.2)$$

where $i_{\alpha}, j_{\alpha}$ express the number of instances of the allele +1 at locus $\alpha \in \{1, \ldots L\}$ present in the population after and before the generation step, respectively; in a FW model, $\sum_{\alpha} i_{\alpha} =$
\[ \sum a_i = N. \]

The state of the population at time \( t + 1 \) depends exclusively on that of the population at time \( t \), therefore this is a Markov process [Gardiner, 2004]. We can exploit this observation to set up the Markov chain formalism: let us consider again one gene, we call \( P(i, t|0_0, 0) \) (or \( P(i, t) \), simply) the probability of finding \( i \) instances of the allele +1 at time \( t \), given that there were \( i_0 \) initially. We define

\[
P(t) = \begin{bmatrix} P(1, t) \\ P(2, t) \\ \vdots \end{bmatrix} \quad \text{and} \quad P = \begin{bmatrix} p_{11} & p_{12} & \cdots \\ p_{21} & p_{22} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} : \text{(B.3)}
\]

\( P(t) \) is a probability vector, \( P \) is the stochastic transition matrix, \( p_{ij} \) are defined by Eq.(B.1). The following normalizations hold: \( \sum_i P(i, t) = 1, \sum_i p_{ij} = 1 \forall i \). Note that

\[
P(i, t + 1) = \sum_j p_{ij} P(j, t). \quad \text{(B.4)}
\]

A generation step can be now written as:

\[
P(t) = P(t-1) = P^2 P(t-2) = \cdots = P^t P(0) \quad \text{(B.5)}
\]

and number of conclusions can be drawn from this set-up. As a first step, suppose there is a stationary state, i.e. \( \exists P^* | PP^* = P^* \); we would like to have an explicit expression for \( P^* \), eigenvector of \( P \) with eigenvalue \( \lambda = 1 \). We might guess that the condition for the stability is that the an allele fixes or, by symmetry, gets extinct. This would imply:

\[
P^* = \begin{bmatrix} 1 - \xi \\ 0 \\ \vdots \\ 0 \\ \xi \end{bmatrix}, \quad \text{(B.6)}
\]

where \( \xi \) is the probability that the allele +1 fixates. Indeed it can be easily verified that such a probability vector is a fixed point of the map Eq.(B.5). We still need to learn about the value of \( \xi \), though. It is illuminating to find out that the expected number of alleles +1 does not change from generation to generation, as it is showed by the following line:

\[
\langle i(t+1) \rangle = \sum_i i P(i, t+1) = \sum_i i \sum_j p_{ij} P(j, t) \overset{(a)}{=} \sum_j j P(j, t) = \langle i(t) \rangle, \quad \text{(B.7)}
\]
where in (a) we have used $\sum_{i} ip_{ij} = N(j/N) = j$. Therefore, if $i_{0}$ is the initial number of +1 alleles in the population and since in the limit $t \to \infty$ we have as the only possible states either fixation with probability $\xi$ or extinction with probability $1 - \xi$, we can write

$$\lim_{t \to \infty} \langle i(t) \rangle = N\xi + 0(1 - \xi) = i_{0} \iff \xi = \frac{i_{0}}{N}.$$  \hspace{1cm} (B.8)

In words, the FW model predicts that, sooner or later, the alleles either fixate or go extinct and the corresponding probabilities only depend on their initial frequency in the population.

The Markov process formalism provide a precise framework to handle genetic drift in a population and to implement also other mechanisms that take into account e.g. mutations, selection. Yet the calculations soon become cumbersome to work out and in order to deal with these difficulties a number approximations have been proposed. In particular M. Kimura [Kimura, 1955] showed that in large populations the discrete FW process can be approximated by a continuous time, continuous space diffusion process. We will not describe the Kimura’s theory in detail, but as a conclusive point, let us show and discuss its core equation.

Let $x = i/N$ be the frequency of the allele +1, $P(x, t)$ the probability of finding $x$ at time $t$. Let us suppose that, together with genetic drift, there is also a selective advantage for the allele +1 i.e. the individuals with that allele have on average $1 + s$ offspring, those with $-1$ only on average just 1; let also $\sigma^{2}$ be the variance in the number of offspring. Finally, assume also that fluctuations are uncorrelated across individuals and generations (non-heritable). Then in a diffusion approximation the distribution of the variant frequency evolves according to the following Fokker-Planck Equation (Kolmogorov forward equation) [Gardiner, 2004]:

$$\frac{\partial P(x, t)}{\partial t} = \left[-s \frac{\partial}{\partial x} x(1-x) + \frac{\sigma^{2}}{2N} \frac{\partial^{2}}{\partial x^{2}} x(1-x)\right]P(x, t).$$  \hspace{1cm} (B.9)

The term with the first order derivative comes from the selection mechanism with strength $s$; the second one is the drift term, which results from the variance $\sigma^{2}$ in offspring number. As it should be, the latter is damped when increasing the population size and vanishes when $N \to \infty$. In a population of size $N$, the frequency $x$ of a variant therefore has a diffusion constant $\sigma^{2}/Nx(1-x)$. Starting from Eq.(B.9), it is possible to evaluate several relevant statistics e.g. the allele-frequency distributions, fixation probabilities, expected time for fixation/extinction (...) Nevertheless, such an approximation is not feasible when, for instance, fluctuations are correlated over generations and their effect is crucial for the fate of the population.
In this appendix we briefly summarize the structure of a basic version of the code that has been used throughout the Thesis for the evolutionary simulations. The code has been written in Python 2.7 and it is summarized in Fig.(C.1).

The first section of the code is handled by the FFPopSim package, imported as a module in Python 2.7. The initialization consists in setting the parameters of the simulation e.g. Tab.(2.1), the fitness parameters are generated separately with the desired distribution. The population is then evolved one generation at a time. During the process there is the possibility of extracting information and plot some instantaneous observables; at the end of the process, we can also get plots for all-time evolution of $\chi_i, \chi_{ij}, \langle F \rangle, \ldots$. Some of them have been discussed in Ch.(2).

If we are also interested in the inference of the fitness landscape, then we have to use the evolutionary information on $\chi_i, \chi_{ij}$. In the case of the NS inference, we first have to infer the couplings $J_{ij}$ of the Ising distribution with a DCA algorithm; this is done by exploiting a simple MF algorithm (written separately and imported) and/or a more sophisticated PLM method, for which we have used the code developed in [Gao et al., 2019] (the same has been used in Ch.(6)).

Finally, we can reconstruct the epistatic fitness landscape by exploiting one of Eq.(1.28, 4.16) for the NS inference Ch.(1) or the GC inference Ch.(4), respectively. Scatter plots and heat maps for the reconstruction error can be drawn, as explained in Sec.(3.4).
Where does the Boltzmann distribution come from? What is the frequently invoked principle of maximum entropy? Which kind of entropy does it deal with?

These fundamental questions are the target of the present appendix and the answers have root in the information theory: we can but refer to [Cover & Thomas, 2006; Mezard & Montanari, 2009] for extensive general treatments of this topic. We here summerize the key ideas following [Nguyen et al., 2017] and apply our results to the context of the Inverse Ising Problem.

We first address the question of the origin of the Boltzmann distribution. When dealing with probability theory and combinatorics, a useful analogy often is that of balls in a box/urn. So let us consider \( M \) indistinguishable balls and suppose we want to place them in the \( R \) compartments of a box. We then fix the occupancy \( n_r, \ r \in 1 \ldots R \) (i.e. the number of balls) of each compartment, clearly \( \sum_{r=1}^{R} n_r = M \); in the language of statistical physics we would say that we are fixing a macrostate. Assuming a large \( M \) we can write \( n_r = M q_r \) and the number of configurations compatible with the previous constraints is

\[
W = \frac{M!}{\prod_{r=1}^{R} n_r!} \sim \frac{e^{M M^M}}{\prod_{r=1}^{R} e^{n_r n_r^r}} = \frac{1}{\prod_{r=1}^{R} q_r q_r^r}, \tag{D.1}
\]

where in (a) we have used the Stirling’s approximation \( n_r! \sim e^{n_r n_r^r} \). The logarithm of the number of microstates consistent with a given macrostate is proportional to the Gibbs entropy:

\[
\frac{1}{M} \log W \sim - \sum_{r=1}^{R} q_r \log q_r. \tag{D.2}
\]

Let us now consider \( \log W \) in Eq.(D.2) as a function of \( q_r \). In the statistical physics interpretation, the argument goes as follows: each of the \( R \) compartments of the box is a microstate of the system, with energy \( E_r \). The \( M \) balls in the compartments can be viewed as a set of copies of the system, named ensemble of replicas; each of the replicas can exchange energy with the others but the ensemble is isolated and has a fixed total energy \( ME \). Assuming that each state of the such ensemble of replicas (i.e. each set of \( \{q_r\} \) for which \( \sum_r n_rE_r = ME \) has the same probability, the statistics of \( q_r \) will be dominated by a sharp maximum of \( W \), hence \( \log W \), as a function of \( q_r \), subject to the constraint that \( \sum_r q_r = 1 \) and \( \sum_r q_r E_r = E \). Following [Jaynes, 1957], we implement these constraints using the Lagrange multipliers. This is done by setting to zero the derivative of

\[
- \sum_{r=1}^{R} q_r \log q_r + \eta [\sum_{r=1}^{R} q_r - 1] + \lambda [\sum_{r=1}^{R} q_r E_r - E] \tag{D.3}
\]
with respect to \( q_s \), yielding the *Boltzmann distribution*

\[
q_s = e^{-\eta} e^{\lambda E_s} = \frac{1}{Z} e^{\lambda E_s};
\]  \hspace{1cm} (D.4)

in the last expression we have defined \( \eta = 1 - \log Z \) and \( \lambda \) is determined by the condition \( E = \frac{\partial}{\partial \lambda} \log Z \).

It is worth noting that the rhs Eq.(D.2) has a familiar form in information theory, where the fundamental quantity is the celebrated *Shannon entropy*: let \( X \) be a discrete random variable with alphabet \( \mathcal{X} \) and probability distribution \( p(x) = \Pr\{X = x\}, x \in \mathcal{X} \), then

\[
S(X) = - \sum_{x \in \mathcal{X}} p(x) \log p(x)
\]  \hspace{1cm} (D.5)

is the Shannon entropy of \( p \). This quantity is a measure of the *uncertainty* of the random variable \( X \) and has a number of interesting properties [Cover & Thomas, 2006]. In fact, the entire statistical mechanics could to be regarded as an application of the Shannon information theory: one may interpret the thermodynamic entropy as the amount of Shannon information needed to define the detailed microscopic state of the system.

In deriving Eq.(D.4) we have operated a constrained maximization of the entropy Eq.(D.2), where the constraint was a fixed total "energy". A similar argument can be used for the *maximum entropy estimation* of the shape of an unknown probability distribution, subject to one or more generic constraints. Suppose we want to determine the probability distribution \( p_r \) compatible with the constraints \( \sum_r p_r = 1 \) and \( \sum_r p_r E_r = E \). Consider \( M \) independent casts of a fair die with \( R \) faces and let \( n_r \in 1, \ldots, M \) be the number of time that the outcome \( r \in 1, \ldots, R \) is observed. The probability of a specific set of outcomes \( \{n_r\} \) is

\[
\frac{M!}{\prod_{r=1}^R n_r!} \prod_{r=1}^R \left( \frac{1}{R} \right)^{n_r},
\]  \hspace{1cm} (D.6)

and its logarithm is \(-M \sum_r q_r \log q_r - M \log R\), where \( q_r = n_r / M \) i.e. the Shannon entropy minus a constant. The idea now is that the best estimate for the probabilities \( p_r \) corresponds to the set of \( \{n_r\} \) that are most likely and at the same time verify the desired constraint. Therefore maximizing the Shannon entropy subject to the constraints \( \sum_r p_r = 1 \) and \( \sum_r p_r E_r = E \) is the max-entropy estimate of \( p \), that takes into account the information of the constraints (and nothing more). If the total "energy" \( \sum_r n_r E_r \) is fixed, we have already seen that such distribution is Eq.(D.4). If the die is unfair, so that the outcome \( r \) has probability \( q^0_r \) one maximises instead

\[
\log \left( \frac{M!}{\prod_{r=1}^R n_r!} \prod_{r=1}^R \left( \frac{q^0_r}{n_r} \right)^{n_r} \right) = -M \sum_r q_r \log \frac{q_r}{q^0_r}.
\]  \hspace{1cm} (D.7)
This last quantity, too, has a generalization in the broader context of the information theory.
Using the same notation as for Eq. (D.5), we define the Kullback-Leibler distance between the distributions $p_1(x)$ and $p_2(x)$ as
\[
D_{KL}(p_1 \| p_2) = \sum_{x \in \mathcal{X}} p_1(x) \log \frac{p_1(x)}{p_2(x)} ;
\] (D.8)
up to a sign, this is exactly Eq. (D.7). The $D_{KL}$ is a measure of the inefficiency of the distribution $p_2$ when the true distribution is $p_1$, it is non-negative and it equals zero if and only if $p_1 = p_2$.

As an application of this discussion most relevant for the context of this Thesis, let us consider the following problem. Suppose we have a system with $N$ binary variables $s_i$, the distribution $p(s)$ of which is not known. Suppose also we have a set of observation from which we extract the first and second moments $\chi_i = \langle s_i \rangle$, $\phi_{ij} = \langle s_i s_j \rangle$. According to the max-entropy receipt, the best guess for $p(s)$ given $\chi, \phi$ is computed by setting to zero the derivatives with respect to $p$ of
\[
-\sum_s p(s) \log p(s) + \eta \left( \sum_s p(s) - 1 \right) + \sum_i h_i \left( \sum_s p(s) s_i - \chi_i \right) + \sum_{i<j} J_{ij} \left( \sum_s p(s) s_i s_j - \chi_{ij} \right) ,
\] (D.9)
where we have used the Lagrange multipliers $\eta, h, J$. We get the distribution
\[
p(s) = e^{-1+\eta e^{\sum_i h_i s_i + \sum_{i<j} J_{ij} s_i s_j}} = \frac{1}{Z} e^{\sum_i h_i s_i + \sum_{i<j} J_{ij} s_i s_j} .
\] (D.10)
where $Z$ is fixed by the normalization and $h, J$ are chosen so to reproduce the observed moments $\chi, \phi$. We immediately recognize in the previous equation the Boltzmann distribution Eq. (3.2) with the Ising Hamiltonian Eq. (3.1). The Shannon entropy of this distribution is readily computed to be
\[
S = -\sum_i h_i \chi_i - \sum_{i<j} J_{ij} \phi_{ij} + \log Z .
\] (D.11)
Simple and elegant as it is, the argument used above has at least one major conceptual difficulty: considered that by hypothesis the underlying distribution is unknown, why should we a priori disregard data beyond the first two orders? Indeed there is no lack of criticism around the max-entropy principle, especially when applied to infer distribution of real (e.g. biological) data [Aurell, 2016].

Possible contexts in which the approach discussed above is suitable are for instance 1. when there is lack of data, in which case higher order statistics are poorly determined; 2. when the data are generated by an equilibrium model with at most pairwise interactions; 3. when models with higher order interactions $J_{ijk\ldots}$ can be reasonably approximated by simple Ising models; 4. when the true distribution is so complicated that there is no other choice than trying to employ an “effective” Ising model which can be useful in deriving bounds e.g. for energy and entropy. For more on these, see [Nguyen et al., 2017] and references therein.
Glossary

**allele** One of the possible alternative forms of a gene, often distinguished from other alleles by phenotypic effects. 4

**archaea** From the Greek word *archaios*, meaning “ancient” or “primitive”. Prokaryotic organisms (i.e. no nucleus) typically found inhabiting and thriving in extreme environmental conditions e.g. exceedingly salty or scorching. 3

**bottleneck** A drastic reduction in population size and consequent loss of genetic diversity, followed by an increase in population size. The rebuilt population has a gene pool with reduced diversity caused by genetic drift. 5

**central dogma** Originally stated by F. Crick, it says that the genetic information flow progresses from DNA to RNA (transcription) to proteins (translation). Exceptions are known. 4

**chromosome** In prokaryotes, a DNA molecule containing the organism’s genome; in eukaryotes, a DNA molecule complexed with RNA and proteins to form a threadlike structure containing genetic information arranged in a linear sequence. 4

**crossing-over** The exchange of genetic material during sexual reproduction between two homologous chromosomes. It happens during meiosis. 5

**diploid** *(2n)* A cell or organism in which each chromosome exists in pairs; having two of each chromosome. 4

**DNA** *Deoxyribonucleic Acid*. A macro-molecule usually consisting of nucleotide polymers comprising antiparallel chains in which the sugar residues are deoxyribose and which are held together by hydrogen bonds between base pairs. 4

**eukaryote** Organism having true nuclei and membranous organelles and whose cells divide by mitosis and meiosis. 3

**evolution** In biology, the change in inherited traits over successive generations in populations of organisms. 3

**fitness** Expected reproductive success of an organism. For modelling purposes, this is often equated the average number of offspring in the subsequent generation (absolute fitness). 4

**frequency-dependent selection** The fitness of a genotype depends of the frequency of that genotype in the population. It is *positive if* fitness increases with frequency, *negative* in the opposite case. 24

**gene** A DNA sequence coding for a single polypeptide. It is the basic physical and functional unit of heredity. From a functional point of view, any discrete locus of heritable,
genomic sequence which affect an organism’s traits by being expressed as a functional product or by regulation of gene expression.

**genetic drift** Sampling fluctuations of genotype or allele frequencies, most often observed in small populations.

**genotype** The allelic or genetic constitution of an organism; often, the allelic composition of one or a limited number of genes under investigation.

**haploid** (n) A cell or an organism having one member of each pair of homologous chromosomes.

**heredity** (also inheritance). Transmission of traits from one generation to another. The study of heredity in biology is genetics.

**hybridization** A situation when two strains of one species that have been evolving in isolation for some time come in contact again.

**lineage** Temporal series of populations, organisms, cells, or genes connected by a continuous line of descent from ancestor to descendant.

**meiosis** The process of cell division in sexually-reproducing organisms during which the diploid number of chromosomes is reduced to the haploid number.

**mutation** Any process that produces an alteration in DNA or chromosome structure; in genes, the source of new alleles. Among the most common: insertion, duplication, deletion, translocation, inversion, point mutation. They are silent if they do not alter the polypeptide chain, missense if they cause a substitution of a different amino acid in the resulting protein, nonsense if they result in a premature stop codon.

**natural selection** Differential reproduction among members of a species owing to variable fitness conferred by genotypic differences.

**nucleobases** (also nitrogenous bases). Nitrogen-containing biological compounds, the most common being adenine (A), cytosine (C), guanine (G), thymine (T), and uracil (U). The bases A, T, C, G are found in the DNA, in the RNA T is replaced by U.

**phenotype** Ensemble of the observable characteristics or traits of an organism.

**population** A group of organisms of a species that interbreed and live in the same place at the same time. They are capable of interbreeding or reproduction.

**population genetics** Study of the genetic composition of populations, including distributions and changes in genotype and phenotype frequency in response to the processes of natural selection, genetic drift, mutation and gene flow.

**prokaryote** Organism lacking nuclear membranes and true chromosomes. Bacteria and blue–green algae are examples of prokaryotic organisms.

**quantitative genetics** Study of the genetic basis underlying phenotypic variation among individuals, with a focus primarily on traits that take a continuous range of values e.g. height, weight, longevity.
**recombination** The process that leads to the formation of new allele combinations on chromosomes. In eukaryotes, genetic recombination during meiosis can lead to a novel set of genetic information that can be passed on from the parents to the offspring. In bacteria recombination happens by transformation (ability to take up DNA from the surroundings), transduction (transfer of genetic material by the intermediary of viruses), and conjugation (direct transfer of DNA from a donor to a recipient).

**transcription** Transfer of genetic information from DNA by the synthesis of a complementary RNA molecule using a DNA template.

**translation** The derivation of the amino acid sequence of a polypeptide from the base sequence of an mRNA molecule in association with a ribosome and tRNAs.


Wright, S. (1931). Evolution in mendelian populations. Genetics, 16(2), 97. https://www.genetics.org/content/16/2/97


ACKNOWLEDGEMENTS

L’ultima pagina, in italiano. Non si può ringraziare in poche righe la moltitudine persone che popolano uno spazio di ricordi lungo due anni, in poche righe non si disegna il mondo. Non sarà una lista, non sarà abbastanza, i grazie non sono mai abbastanza.

L’ultimo punto di questa pagina, sarà il punto d’arrivo di questo percorso, l’ultimo inchino a teatro, l’ultimo metro di terra arata. Ho sempre immaginato i momenti finali come un banchetto a cui tutti sono invitati e c’è una tavola lunghissima e l’atmosfera agrodolce dei momenti, per definizione, irripetibili, che non tornano. Propongo allora un brindisi.

Un brindisi ai miei relatori. Erik, per esser stato molto più che un insegnante: un maestro; Hongli, la cui passione e dedizione alla ricerca è oltre ogni umana pretesa. Ad entrambi, per l’opportunità che ho avuto, per esser io stato sì studente ma collega.

Un brindisi agli attori protagonisti, agli attori non protagonisti, ai figuranti, a tutti coloro che hanno animato le sequenze, i dialoghi; a coloro che, di nascosto, hanno diretto la regia, suggerito la sceneggiatura. Per tutte le volte che, zoppicanti e stonati, ci siamo ritrovati l’uno davanti all’altro, a farci compagnia, a scegliere d’essersi insieme, a lasciarsi andare. Non m’era più in cuore la ruota delle stagioni e il gocciare del tempo inesorabile.

Un brindisi alla mia famiglia, per tutte le volte che non ci sono stato, per tutte le volte che invece c’ero ed ero felice senza motivo, come a casa, come da bambino, com’è sempre stato. Da vicino, di un’opera si distinguono i dettagli, da lontano se ne coglie il senso.

Un brindisi a Trieste, ad i posti che ci hanno ospitato, nello spazio di un caffè, in una notte, in tutti questi anni. Per tutti i ricordi che abbiamo nascosto fra queste vie; per esser stata teatro e palco di ogni momento, di ogni granello d’esperienza, di tutte le volte che si è riso senza pensieri, di tutte le volte che la notte si è stancata troppo presto. Per tutte le volte che il mare ci ha fatto sentire minuscoli, ma senza confini.

Un brindisi a Stoccolma, la sua bellezza vasta, luminosa, silenziosa. Per quando è scomparso il confine fra i giorni, e ci siamo sentiti vivi; per quando il mondo fuori di lì è sembrato distante, per non esserci mai sentiti nel posto sbagliato.


Felicità raggiunta, si cammina
per te sul fil di lama.
Agli occhi sei barlume che vacilla,
al piede, teso ghiaccio che s’incrina;
e dunque non ti tocchi chi più t’ama.