Minimal models of evolution: germline fitness effects of cancer mutations and stochastic tunneling under strong recombination

I n a u g u r a l - D i s s e r t a t i o n

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität zu Köln

vorgelegt von

Andrej Fischer

aus Staßfurt

Köln, 2011

Berichterstatter: Prof. Dr. Alexander Altland
Prof. Dr. Thomas Wiehe

Tag der letzten mündlichen Prüfung: 5.12.2011

Abstract

In a time where data on the genetic make-up of organisms is available in abundance, the theory of evolution is of immediate importance to answer key questions of biology: How can one explain the variation seen in the DNA of different organisms and species? What are the effects of changes in the DNA on the function of cells? What are the driving mechanisms of diseases with a genetic component such as cancer? Minimal mathematical models of evolution provide a basis for the interpretation of DNA data. The explanations they offer are concrete and testable, their assumptions and limitations explicit. The application and further development of minimal evolution models is the main theme of this work. In the first part, the functional effects of mutations found in cancer cells are analyzed from the perspective of germline evolution. This is the process that produced the DNA of organisms as we see it today. Mutations have an effect on the fitness of healthy cells. This impact can be estimated from the variation seen in the sequences of protein domains. It is found that this evolutionarily informed conservation score has utility to identify cancer driver genes, especially if they are tumor suppressor genes. The relevance of this fitness scale for cancer mutations is demonstrated on a data set of mutations in protein kinase genes. This analysis is followed by an application of Hidden Markov Models (HMM) to the detection of signals of positive selection in cancer mutation data. Cancer as an evolutionary process of cells is markedly different from the process of germline evolution. Cancer-specific selection can be seen in genes, whose activity or lack thereof is essential for the progress of cancer. These cancer genes exhibit an increased rate of amino acid changing mutations, beyond the level expected by chance. The identification of these genes is a statistical task for which HMM are shown to be most suitable. Finally, an extended mathematical model of evolution is analyzed which describes the adaptation of a sexually reproducing population to a global fitness maximum via compensatory mutations. In a two-locus/two-allele model, the compound effects of mutation, selection, genetic drift, recombination and sign epistasis lead to the interesting situation of adaption via the crossing of a fitness valley in genotype space. This bottleneck can be overcome by rare large fluctuations in the allele frequencies overcoming the effect of recombinatorial reshuffling. The relevant time scales are derived for a parameter regime that includes large recombination.

Zusammenfassung

In einer Zeit, in der Daten über den genetischen Aufbau von Organismen im Überfluss verfügbar sind, spielt die Evolutionstheorie eine zentrale Rolle in der Beantwortung von Schlüsselfragen der Biologie: Wie erklärt sich die Variation, die man in der DNA von verschiedenen Organismen und Spezies findet? Welche Effekte haben Veränderungen in der DNA auf die Funktion von Zellen? Was sind die treibenden Kräfte bei Erkrankungen mit genetischer Komponente, wie etwa bei Krebs? Mathematische Evolutionsmodelle bilden eine Grundlage zur Interpretation von DNA Daten. Unter expliziten Voraussetzungen liefern sie konkrete und prüfbare Vorhersagen. Die Anwendung und Weiterentwicklung minimaler Evolutionsmodelle ist das Leitmotiv dieser Arbeit. Zuerst werden die funktionalen Effekte von Mutationen in Krebszellen analysiert. Dies geschieht aus der Perspektive von Keimzellevolution, die die DNA hervorbrachte, die wir heute in allen Zellen finden können. Mit Hilfe von öffentlich zugänglichen Sequenz-Daten über Protein-Domänen kann abgeschätzt werden, wie groß der evolutionäre Fitnesseffekt von Mutationen ist. Mithilfe dieser Möglichkeit Mutationen zu bewerten können dann Gene identifiziert werden, die für die Krebs-Evolution entscheidend sind. Die Relevanz dieses Ansatzes wird an einem Datensatz von Krebsmutationen in Protein-Kinase Genen exemplarisch dargestellt. Darauf folgt eine Anwendung der Methode der Hidden Markov Modelle (HMM) um Signale von positiver Selektion in Krebs-Daten zu finden. Krebs- und Keimzellevolution sind prinzipiell verschiedene Prozesse. Krebs-spezifische Selektion kann aber über eine erhöhte Rate von nicht-synonymen Mutationen in Genen nachgewiesen werden, die für das Fortschreiten der Krebsentwicklung aus- oder auch eingeschaltet sein müssen. Die Identifikation dieser Krebsgene kann mittels HMM effektiv durchgeführt werden. Im letzten Teil der Arbeit wird ein erweitertes mathematisches Evolutionsmodell analysiert, das Adaption zu einem Zustand maximaler Fitness durch kompensatorische Mutationen bei sexueller Fortpflanzung beschreibt. Dieses Modell beschreibt die Evolution von zwei Loci mit jeweils zwei Allelen. Die gemeinsamen Effekte von Mutation, Selektion, Rekombination und Epistase führen hier zu der Situation, dass ein “Fitness-Tal” durchquert werden muss um den Genotyp höchster Fitness zu erreichen. Adaption geschieht durch seltene zufällige Fluktuationen der Allelfrequenzen, in denen der Effekt des Durchmischens durch Rekombination überwunden wird. Die relevanten Zeitskalen für diesen Prozess werden für einen weiten Parameterbereich hergeleitet, der auch starke Rekombination beinhaltet.

Contents

List of Figures

List of Tables

Mathematical notation

Symbol	Name	Definition
$\Gamma(x)$	Gamma function	$\Gamma(x) = \int_0^\infty \mathrm{d}t\; t^{x-1}e^{-t};\ \mathrm{Re}(x) > 0$
$\Gamma(x,a)$	Incomplete Gamma function	$\Gamma(x,a) = \int_a^\infty \mathrm{d}t\; t^{x-1}e^{-t};\ \mathrm{Re}(x) > 0$
$\mathrm{Beta}\,(\boldsymbol{\alpha})$	Beta function	$\mathrm{Beta}\,(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^n \Gamma(\alpha_i)}{\Gamma(A)};\ \boldsymbol{\alpha} \in \mathbb{R}^n_{\geq 0}$
Δ^{n-1}	$(n-1)$-Simplex	$\Delta^{n-1} = \left\{ \boldsymbol{\theta} \in \mathbb{R}^n \;\middle|\; \sum_{i=1}^n \theta_i = 1,\ \forall i = 1,\ldots,n:\ 0 \leq \theta_i \leq 1 \right\}$
$\mathrm{Dir}\,(\boldsymbol{\theta} \mid \boldsymbol{\alpha})$	Dirichlet distribution	$\mathrm{Dir}\,(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) = \frac{1}{\mathrm{Beta}(\boldsymbol{\alpha})} \prod_{i=1}^n \theta_i^{\alpha_i - 1};\ \boldsymbol{\alpha} \in \mathbb{R}^n_{\geq 0},\ \boldsymbol{\theta} \in \Delta^{n-1}$
$P(\boldsymbol{k} \mid \boldsymbol{\theta})$	Multinomial sampling distribution	$P(\boldsymbol{k} \mid \boldsymbol{\theta}) = \frac{(\sum_{i=1}^n k_i)!}{\prod_{i=1}^n k_i!} \prod_{i=1}^n \theta_i^{k_i};\ \boldsymbol{k} \in \mathbb{N}^n,\ \boldsymbol{\theta} \in \Delta^{n-1}$
$\mathbb{E}^{\pm}$	Shift operator	$\mathbb{E}^{\pm} f(n) = f(n \pm 1);\ f : \mathbb{N} \to \mathbb{R},\ n \mapsto f(n)$
$I_\alpha(x)$	Modified Bessel function (first kind)	$I_\alpha(x) = \sum_{m=0}^\infty \frac{1}{m!\,\Gamma(m+\alpha+1)} \left(\frac{x}{2}\right)^{2m+\alpha};\ \alpha, x \in \mathbb{R}$
H_n	Harmonic number	$H_n = \sum_{k=1}^n \frac{1}{k};\ n \in \mathbb{N}$
$\psi(x)$	Digamma function	$\psi(x) = \frac{\mathrm{d}}{\mathrm{d}x} \ln \Gamma(x);\ x \in \mathbb{R}$
$D_{\mathrm{KL}}\,(p \,\|\, q)$	Kullback-Leibler divergence	$D_{\mathrm{KL}}\,(p \,\|\, q) = \int_{-\infty}^{\infty} \mathrm{d}x\; p(x) \ln \frac{p(x)}{q(x)}$

Chapter 1

Synopsis

"Nothing in biology makes sense,
except in the light of evolution."

Theodosius Dobzhansky, 1973

The title of Dobzhansky's 1973 article [1] may be provocative and by now overused. But nowadays, more than ever before, the theory of evolution has become an essential part of biology. The first complete draft of the human genome[1] in 2001 [2] marked the beginning of a new era. Ten years later, new technological advances made it possible to sequence thousands of humans across the world[2] and to assemble the genomes of over 180 more eukaryote species[3]. This huge amount of sequence data sheds new light on the genetic factors for common diseases with the help of so called "genome-wide association studies" [3, 4]. Currently, a large international sequencing project aims to find the genetic causes of cancer [5].
In order to make sense of this plethora of data, one needs to understand the mechanisms behind the process that produced these DNA sequences and their statistical characteristics: evolution. Mathematical models of evolution try to capture the influences of forces such as selection, mutation and genetic drift (chance) on the distribution of genes in the gene pool. These effects thereby become quantifiable and ultimately measurable. This provides a starting point to explain genetic variation.
The broad theme of this work is the application and analysis of minimal models

[1] Please see page 16 for a glossary of genetics related terms.
[2] See e.g. the 1000 genomes project: www.1000genomes.org
[3] See e.g. www.ncbi.nlm.nih.gov or www.ensembl.org. Eukaryotes are organisms with cells that have a nucleus, in which the DNA is contained. In the other domains of life, the number of sequenced species is even larger: 1710 for prokaryotes (having cells without a nucleus, e.g. bacteria) and 2695 for viruses (September 2011).

of evolution. In this context, models are "minimal" if they are as complex as necessary to have some utility in the interpretation of data, but at the same time as simple as possible to allow for an analytical treatment. In the three chapters following this synopsis, a straightforward application of well-known predictions for evolution under mutation, selection and genetic drift to the analysis of cancer mutations is performed. The central quantity in this analysis is evolutionary fitness, defined as the growth advantage of cells or organisms conveyed by their genes over other organisms carrying competitor genes. This fitness directly depends on the proper function of cells and ultimately on the instruction sets encoded in the DNA. This close link between biological fitness and genotype is reflected in the distribution of genotypes in the pool of all available variants. Mathematical models of evolution are able to make predictions for this connection under appropriate assumptions. One main result of this work is, that the change in fitness that is induced by mutations as predicted by these models is shown to be a useful scale to identify cancer-driver mutations, i.e. mutations that are causally related to cancer progression [6].

In chapter two, the well known one-locus/two-allele model of evolution is reviewed. In this situation of competition between two rivaling alleles in a finite population, the model describes how allele frequencies evolve in time. The prediction for observations when taking samples of such a model depend on the particular combination of the model parameters: population size, selection and mutation rate. When analyzing real life genetic data, these parameters are never known. But they can be inferred from the data. This is done in the second part of chapter two by the mechanisms of Bayesian inference [7]. This method is here employed to find statistically meaningful estimates of the fitness cost of mutations from openly available sequence data. Bayesian inference is a general methodology to estimate model parameters from experimental data in due consideration of all prior information. The gain in knowledge about a system through experimental observation is expressed in the language of probability. The whole process of parameter estimation is demonstrated on a simple but useful example - multinomial sampling from an urn - and should be self-consistent. Driven by the need to cope with scarce data, an important part of the derivation concerns the principles of maximum entropy (MaxEnt) and minimum discrimination information (MinDI). With these guiding principles one can consistently incorporate informative prior information and find meaningful point estimates for the model parameters even for small sample sizes. The conceptual aspects of this part are very general. The results are necessary ingredients for the germline fitness estimation scheme. In the last part of chapter two, it is explicitly shown how to "score" observed missense mutations (not just in cancer) by their inferred effect on cells' function and fitness. As a basis for this inference scheme, representative collections of protein domain sequences are used [8], which are readily available in the Pfam database [9]. In

practice, the scoring pipeline - from mutation data and protein domain alignments to final mutation scores - is carried out by computer programs. The current implementation is very scalable: many mutations in many genes can be scored within hours.
Chapter three puts this mutation scoring scheme to a practical use in the analysis of a specific cancer data set: cancer mutations in protein kinase genes [10]. A subset of these genes with known cancer association is shown to harbor more supposedly harmful mutations in cancer cells than would be expected by chance. This is of immediate practical importance for the discovery of new cancer genes. The analysis in this chapter is of statistical nature. By comparing the scores of the observed cancer mutations with the same characteristics of an ensemble of "random" mutations, the significance of this germline fitness scale for cancer evolution is clearly demonstrated.
In the fourth chapter Hidden Markov Models (HMM) [11] are employed to find signals of *positive* selection in cancer mutation data. Genes that are essential for the cancer development are under strong selection in the tumor. The selection pressure is primarily posed by the immune system of the host. The cancer relevant genes will exhibit a significantly higher rate of missense substitutions, i.e. the take-over of new genetic variants in the cancer cell population. HMM are probabilistic models to relate observed data (the number of mutations in the different genes) to the potential but unknown state of the system that generated this data (the selection status of the different genes in cancer). Internally, the HMM presented in this work uses the very same theoretical predictions for the substitution rates from a minimal evolution model that are the base of the germline fitness scoring scheme. The effectiveness of this method to identify cancer driver genes in the protein kinase mutation set mentioned above is demonstrated as the results from the original study are reproduced [10]. However, the HMM method has a larger capacity and flexibility: the analysis can be extended to e.g. selection in protein domains and in protein kinase sub-domains. The last part of chapter four shows how such an analysis could be implemented.
The fifth chapter of this work goes back to the theme of minimal evolution models. It is devoted to the analysis of a more extended model, that - on top of selection mutation and genetic drift - additionally includes the effects of recombination and epistatic interaction of alleles at different loci. In a two-locus/two-allele set-up, this model describes the competition of four genotypes in a sexually reproducing population. Recombination is the exchange of genetic material by (chromosomal) cross-over in sexually reproducing organisms and epistasis describes the non-additive effect of mutations at interacting loci. Interestingly, within this model it is possible to analyze the problem of adaption to a state of maximal fitness by crossing of a fitness valley. This happens, when a mutation at each locus of the wild type genotype is strongly deleterious but more than compensated by

a secondary mutation at the other locus. Such a specific fitness assignment (sign epistasis [12]) leads to the situation that there are two genotypes with high fitness separated by intermediate states of low fitness. Of central importance is the time needed to cross this valley for a population that starts in one of the the local maxima. As a function of recombination strength, there is a cross-over to a regime where valley-crossing is impeded by recombinatorial reshuffling of genotypes. In infinite populations, fixation of the super-fit double mutant is impossible if recombination is too strong [13, 14]. Only rare large fluctuations in finite populations can lead to adaption. The valley crossing rate is then very small. i.e. exponentially suppressed. This process can be described as stochastic tunneling [15]. Similarly to quantum mechanical tunneling, the allele frequencies of the intermediate evolutionary states - the deleterious single mutants - never reach macroscopic sizes in the population during the adaption process. In this stochastic tunneling regime, the relevant time scale for fixation is derived for a wide parameter range that goes well beyond the point where recombination dominates. Also the critical size, that a double mutant subpopulation needs to reach in order to successfully fixate is calculated explicitly. The analytical predictions are derived with the Moran formulation of evolution. They are validated by numerical simulations using a Wright-Fisher variant of the process. In the appendix, the connection between those two complementary formulations is worked out.

1.1 Glossary of genetics related terms

At this point, it is in order to set some nomenclature used throughout this work. All terms explained in this glossary appear in *italics*.

alignment: A graphical way to compare related *homologous sequences* is to align them, i.e stack them on top of each other with corresponding *loci* (*nucleotides* or *amino acids*) in the same column. If the sequences are very similar, this alignment can be done almost "by eye". If there is a certain degree of divergence between the sequences, the alignment procedure is not at all straightforward and elaborate programs - such as the widely used BLAST [16] - are used for that task. An example of an amino acid alignment can be found on page 54.

allele: If a *genetic* unit can appear in different variants in a population, each of those rivals is called an allele of the specific *locus*. The unit can be one single *nucleotide* or even a whole *gene sequence*.

amino acid: Amino acids are the building blocks of *proteins*. In most life forms, 20 different amino acids are used to build proteins. Each one is denoted by a capital roman letter, from (A) for alanine to (W) for tryptophan.

base pair: The two polymers of *DNA* molecules are connected by pairs of *nucleotides*. There are only the two pairings A:T (or T:A) and C:G (or G:C).

cancer: A disease where the *DNA* of cells is changed in such a way that it leads to their uncontrolled growth. Ultimately, this will lead to an unsustainable situation, where the growth of the cancer cells impairs the function of organs. The DNA of cancer cells is markedly different from those of the *germline*. [17, 18].

cancer gene: There are essentially two ways in which genetic changes can lead to the *cancer* state of a cell: either *genes* that control and suppress cell division are "switched off" - so called tumor suppressor genes - or genes that enhance the cell proliferation are "switched on" - so called oncogenes.

codon: The genetic information is encoded in the *DNA* using the four-letter alphabet of *nucleotides*. However, the *protein* products are chains of symbols from the twenty-letter set of *amino acids*. So there is a mapping from the *nucleotide* to the *amino acid sequence*: the genetic code. Three nucleotides are grouped to form a codon, a short three letter sequence, which is translated into one of the twenty amino acids. This mapping is realized in the process of *translation*. Since three letter sequences could potentially code for $4^3 = 64$ meta-letters, the genetic code is inherent with a high degree of redundancy. There is a 21st symbol coded for by several codons: the stop codon that signals the end point of a coding sequence. The start position of a coding sequence is always located at the first methionine (M) with its codon ATG.

clone: A clone describes the perfect copy of a genome. In the context of *cancer*, a single tumor cell with a large growth advantage conveyed by a *driver mutation* will rapidly multiply and produce copies of itself - an event called clonal expansion. Importantly, all members of a clonal sub-population are genetically identical.

deletion: This kind of *mutation* deletes a whole sub-*sequence* and can have the same potential effects as an *insertion*, i.e. a complete loss of the proper protein function.

diploid: Diploid organisms have two copies of *DNA* in their cells, one from each of their parents. This means, that there are also two copies of every *locus*. If the two copies are identical, the locus is called homozygous, else heterozygous. Cells with just one copy of their DNA are called haploid, such as sperm cells or egg cells.

DNA: Deoxyribonucleic acid is a large molecule present in most cells of all living life forms that stores all information necessary for the assembly and function of an organism. DNA molecules consist of two polymer chains of a sugar and phosphate backbone, called "strands", on which the *nucleotides* are arranged in a

complementary fashion to form a double helix. The two polymers are connected by complementary pairs of nucleotides.

driver mutation: A mutation that is causally linked with *cancer* development. Standard cancer models assume that a tumor goes through a succession of *clonal* expansion stages, where each stage is initiated by one or more driver mutations [19, 20].

fitness: In evolutionary terms, the fitness of an organisms is its rate of growth compared to other organisms. In models of evolution, fitness is always defined relative to that of competitors. The central quantity is often the *selection* coefficient, i.e. the difference in fitness between *alleles*.

fitness landscape: The assignment of *fitness* values to different *alleles* in the space of all possible alleles is called a fitness landscape. Alleles with local fitness maxima (local in the sense of number of mutations as a distance measure) are sometimes called "fitness peaks". If there are multiple peaks, they are usually separated by "fitness valleys". If the fitness assignments change as a function of time, one might speak of fitness "seascapes" [21].

gene: Some of the *sequence* in the *DNA* is in fact a collection of recipes or programs for building machinery needed in the cell. Such a single instruction set sequence is called a gene. In human DNA, there are about $2 \cdot 10^4$ genes [22].

genetic drift: This is a term describing the influence of random reproduction successes in the evolution of a finite population.

genome: The genome is a term for the total of all hereditary information contained in a *DNA* molecule. This entails coding *sequences* for *genes*, inter-genetic sequences and sequences that are important for the regulation of gene expression.

germline: In an adult organism, only the genetic material of a certain line of cells - the germline - could be potentially passed on to the next generation and is thus relevant for evolution. These are the gametes - sperm and egg - and all cells from which they derive, up to the cells from which an organism evolved. As only changes in the germline are passed on to the next generation, they are under direct evolutionary pressure. Likewise, the *DNA* sequence of germline cells can be understood as the result of a billion year long evolutionary process.

homologue: Homologous genes are similar or identical to each other both functionally and *sequence* wise and they share a common ancestor.

insertion: An insertion is not a point mutation. Instead, a whole *sequence* bit is inserted within a *gene* sequence. If the insertion length is not a multiple of three - the length of a *codon* - the reading frame is shifted, which will quickly result in a premature stop *codon*. Otherwise, the gene will still be translated but will have a

longer protein sequence and different folding behavior.

locus: The physical location of a specific genetic unit (i.e. a *gene*, *codon* or *nucleotide*) within a *genome* is called it's locus (plural: loci).

missense: A missense *mutation* (or nsSNP for non-synonymous SNP) changes the *codon* in a way that it now codes for a different *amino acid*, potentially altering the structure and function of the *protein* product.

monomorphic: A group or population is said to be monomorphic if all its members are of the same (geno-)type. This is the opposite of a *polymorphic* population.

mutation: The *DNA* of cells is subject to processes that change the *sequence* in different ways, e.g radiation or errors in DNA copy during cell divisions. These changes are called mutations. Their effect can be the change of a single nucleotide (*point mutation*), an *insertion* or *deletion* or even a large scale DNA rearrangement.

non-sense: A non-sense mutation changes the *codon* to be a premature stop codon. In the *translation* step, the rest of the *gene* after the non-sense mutation is neglected, potentially making the gene product useless.

non-synonymous: See *missense* mutation.

nucleotide: Nucleotides are the basic building blocks of the genetic material of all life forms. These are the four molecules adenine (A), guanine (G), cytosine (C) and thymine (T).

oncogene: In this class of *cancer genes*, the gene *protein* products show an enhanced activity in the cancer cells. Here, it usually suffices to mutate one of the parental *alleles*, hence they are also called dominant cancer genes. Oncogenes are involved in the regulation of cell growth and differentiation. Most of the cancer genes discovered up to now are oncogenes [17], but this is mostly due to a bias in the methods used for their discovery.

orthologue: Orthologous *genes* are related or identical genes in different species.

paralogue: Paralogous *genes* are present at different locations in the same *genome* which are descendant from a common ancestor gene. They usually stem from a gene-duplication event in the evolutionary past of the genome, where an initially single gene is (erroneously) duplicated during cell division in *germline* cells.

passenger mutation: During *cancer* evolution, many processes that prevent random mutations to accumulate are switched off, such as *DNA* damage repair or apoptosis. Thus, cancer cell DNA also harbors a large number of mutations that are not usually seen in healthy cells, but that are not causally linked to the cancer progression.

point mutation: A change of a single *nucleotide* in the *DNA* (also called SNP for single-nucleotide polymorphism), brought about e.g. by biochemical reactions, radiation or DNA-copying errors. Depending on the genetic context of the mutation location, the effect of a mutation is specified as *silent*, *missense* or *non-sense*.

polymorphic: A group or population is said to be polymorphic if its members belong to various different (geno-)types. This is the opposite of a *monomorphic* population.

protein: Proteins are polymers of *amino acids* chained together by peptide bonds. The polymer is not in a linear configuration, but folded in a state of minimum free energy. Proteins are biological machines and involved in virtually all processes in the cell, from metabolism to signal processing. They are the manifestations of the information encoded in the *genes*.

protein domain: *Genes* (and their products) are not the smallest or only functional units of the *genome*. Instead, each gene includes one or more domains, which are recurring highly conserved strings. Most of the domains can be identified with a very specific function. Domains can be found in many different genes within a single genome. The Pfam database [9] is a resource for information about protein domains.

RNA: Ribonucleic acid is a macromolecule used for the intermediate *transcription* step of a *gene's* expression. Compared to *DNA*, the deoxyribose is substituted by ribose and one of the four nucleotides - thymine (T) - is changed to the nucleotide uracil (U).

selection: Selection in evolution manifests itself via differential growth rates (reproductive success) between different *alleles*. The most prevalent form of selection is called "purifying", i.e. where new variants with comparably low growth rates quickly disappear from a population. Sometimes, a new allele is actually better adapted than the prevalent allele. The quick take-over of this new variant is described with positive selection.

sequence: Main task of the *DNA* is the storage of information and regulation. The information is stored in the "sequence" of letters from the four-letter *nucleotide* alphabet. The sequences of both DNA strands carry the same information due to the complementary nature of the *base pairs*.

silent: Within a coding *sequence*, a silent mutation changes a *nucleotide* in a way that the *codon* still maps to the same *amino acid*. The *protein* product is unchanged. This kind of point mutation might still have a measurable effect, e.g. in the efficiency of *transcription*. A better term for it is therefore *synonymous* mutation.

somatic: All cells that are not in the *germline* are called somatic. Mutations in somatic cells are relevant to the individual organism, but not to the *genome's* evolution - unless the somatic change prevents the organism from reproducing. In the context of *cancer*, the two terms germline and somatic are used in a somewhat different way. Here, somatic mutations (not cells) are those *only seen in the cancer cells*, whereas germline mutations are observed in the non-cancer cells. It is understood that the genetic variation across cells within an organism is so small that the *DNA* of any cell is very close to the "germline".

synonymous: See *silent* mutation.

transcription: If a specific *protein* is needed in a cell, the information about that protein encoded in the corresponding *gene* is materialized by a two step process: transcription followed by *translation*. In transcription, a complementary copy of the gene's *nucleotide* sequence is produced in the form of a mRNA (messenger *RNA*) molecule. The mRNA is subsequently *translated* to produce the final *protein*.

translation: The mRNA - a complementary copy of a *gene sequence* produced in the *transcription* step - is converted to the final *protein* product in the translation step. This is carried out by a large molecule called ribosome. Beginning from the start *codon*, it appends all *amino acids* according to the mRNA sequence until it encounters the stop codon. The synthesized protein is released and folds to its final form.

tumor suppressor: In tumor suppressor genes, both parental *alleles* have to be mutated to deactivate the *gene* product (the "two-hit hypothesis"). That is why they are also sometimes called recessive *cancer genes*. These are mostly genes that are involved in *DNA* damage repair, coupling of the cell cycle to DNA damage and apoptosis (the intentional cell death). The most prominent example for a gene with all those functions is TP53 that codes for the p53 *protein*.

Chapter 2

Bayesian inference of germline fitness

"When the facts change,
I change my mind.
What do you do, sir?"

John Maynard Keynes, 1940

2.1 Introduction

How can one quantify the effect of a mutation on a gene's function? In this first chapter, it is shown how minimal evolutionary models can help to answer this question. Ever since the first human genome was sequenced [2, 23], there is an increasing demand for methods to interpret the observed variation data, especially in the context of disease. There are various complementary methods to quantify the functional impact of a mutation [24, 25], e.g. based on indicators of selection in mutation frequencies [26, 27, 10, 8] or based on bioinformatic methods to infer the functional or structural effect of individual mutations [28, 29, 30, 24]. As an example, this can mean to find the consequences of changes in the amino acid sequence of a protein for its biophysical stability or folding characteristics[1] [31]. This analysis involves extensive simulations of the molecular dynamics of the protein polymer chain to find its folding behavior. These simulations impede the scalability of those methods for very large data sets, as they are typical for current sequencing studies.
The methods presented here try to exploit the fact that the germline DNA sequence of each organism is the result of all of its evolutionary past. All possible evolu-

[1] Please see the glossary on page 16 for an explanation of the most important genetics related terms used in this thesis.

tionary "experiments" have been in fact performed countless times. What we can observe today in each gene's sequence conservation pattern - seen across many organisms - is a manifestation of the selection pressures that have been and still are acting on it. In probabilistic terms, there is a distribution of all possible alleles at each locus, which strongly depends on the influences posed by natural selection - favoring some alleles over others. Observing samples from that distribution would be a direct way to assess the evolutionary "fitness" of each allele. More importantly, for a mutation between two alleles, the statistical information derived from such a sample of alleles provides an evolutionarily meaningful scale for the mutation's functional impact. This scale is called the "germline fitness score" of a mutation.
In this first chapter, the conceptual steps necessary to set-up this mutation scoring scheme are laid out. The problem of inferring parameter estimates from samples of an unknown distribution is a classic task in the natural sciences. It is presented here in terms of Bayesian inference [7]. The concrete difficulties in realizing these conceptual ideas for genetic data in practice will be addressed later in the chapter. Methods based on the analysis of sequence conservation patterns are well known in the field [8, 32, 33, 34], their application in the context of cancer mutations however, is a novel approach.

2.2 The one locus two allele model

The whole idea of germline fitness scoring is based on the following observation. The distribution of alleles at a specific genetic locus mirrors the underlying evolutionary process. Using a minimal model of evolution that includes mutation and selection and genetic drift[2], one can predict how that distribution should look like as a function of the alleles' fitnesses. Moreover, it is possible to infer (estimate) the fitness values from a finite sample of the allele distribution.

2.2.1 Moran model of evolution

This is a short presentation of the most basic evolutionary model capable of capturing the effects of mutation and selection - the one-locus/two-allele model of evolution. It is well known and studied and even allows for a full analytical description [3] [35, 36, 37, 38]. The basic ingredients are:

[2]Genetic drift is a term for the random fluctuations in the evolutionary processes due to finite size of the population.
[3]The solution was found by M. Kimura, hence the model will from now on be referred to as the "Kimura model of evolution".

- The evolution of allele frequencies in a population of organisms is modeled as a stochastic Markov process, i.e. a sequence of random events in time where the next set of events along with their probabilities depends only on the current state of the population (and not its entire history).

- The population is supposed to be of constant size N, where usually $N \gg 1$. This population size is fixed for the entire process.

- Each organism of the population carries a genome with a single locus that allows for one of two alleles, e.g. A and B. We call n_i ($i = A, B$) the number of individuals with these alleles and $x_i = n_i/N$ their allele frequency. Because $n_A + n_B = N$, there is really just one degree of freedom, e.g. $n_A =: n$.

- The events that take place in every evolutionary turn are births and deaths of individuals (the events actually take place at random time points themselves, see appendix C.4). This is the Moran model formulation of evolution [4] [39].

- The probability D_i ($i = A, B$) that an individual with allele i dies in the next turn is given by the allele's frequency: $D_i = x_i$. The probability B_i for a birth of allele i is additionally weighted according to its fitness f_i with the weight $w_i = f_i/\bar{f}$ and mean fitness $\bar{f} = f_A x_A + f_B x_B$. This means: $B_i = w_i x_i$.

- To introduce variation, a newborn mutates to the other allele with the probability u. The birth and death probabilities above change accordingly.

Mathematically, the evolution of the population is fully described by the probability distribution $P(n,t)$, which returns the probability that at time t there are exactly n copies of the A allele present in the population. It is useful to imagine an ensemble of identical populations - a population swarm - all evolving randomly according to the same set of rules. The configuration (e.g. depicted in a histogram) of that swarm evolves in time as well. For an increasing swarm size, this histogram approaches the probability distribution $P(n,t)$. The rate of that change of probability at each state is described by the Master equation, here for the Moran birth-death model [39]. The Master equation is but a continuity equation: the rate of probability change at n is equal to the probability *inflow* from $n+1$ and $n-1$ minus the probability *outflow* to the neighboring states. This flow of probability from source to target is simply given by the amount of probability present at the source times the probability per unit time W for a transition to take place to the

[4] There is an alternative formulation attributed to Wright and Fisher, according to which the whole population is assembled anew in every turn [39]. The relationship between both formulations is demonstrated in appendix C.

target. What was said in words translates to the equation:

$$\begin{aligned}\partial_t P(n,t) \;=\; & \left[P(n+1,t)\,W(n+1\to n)+P(n-1,t)\,W(n-1\to n)\right] \\ & -\left[P(n,t)\,W(n\to n+1)+P(n,t)\,W(n\to n-1)\right]\end{aligned} \tag{2.1}$$

For a birth-death process, a transition from n to $n+1$ alleles of type A can be only accomplished if in above protocol an A is born and a B dies. The probabilities for these events depend on the *current* number of A alleles present. Hence we have

$$W(n\to n+1) = B_A(n)\cdot D_B(n), \quad W(n\to n-1) = D_A(n)\cdot B_B(n) \tag{2.2}$$

Altogether, the Master equation above is usually expressed in a more compact form using the discrete shift operators $\mathbb{E}^{\pm}$ that shift the argument of everything they act on by plus or minus one:

$$\mathbb{E}^{\pm} f(n) := f(n\pm 1)$$

With this we have the birth-death Master equation in its usual form:

$$\begin{aligned}\partial_t P(n,t) &= \left[\left(\mathbb{E}^{-}-1\right) B_A(n) D_B(n) + \left(\mathbb{E}^{+}-1\right) D_A(n) B_B(n)\right] P(n,t) \\ x := \frac{n}{N}, \quad D_A(x) &:= x, \quad B_A(x) := \frac{f_A\, x\,(1-\mu)}{f_A\, x + f_B\,(1-x)} + \frac{f_B\,(1-x)\,\mu}{f_A\, x + f_B\,(1-x)}\end{aligned} \tag{2.3}$$

By definition of the model, in every turn one allele must be born and one must die. This is reflected in the identities:

$$B_A(x) + B_B(x) = 1 \quad \text{and} \quad D_A(x) + D_B(x) = 1 \tag{2.4}$$

In this model, the evolutionary forces are supposed to be small, i.e. $f_i = 1+s_i$, with $s_i \ll 1$. The assumption is made that it suffices to work with the transition rates to leading order in the parameters $\{s_i, \mu\}$.

$$B_A(x) = x + (s_A - s_B)\,x\,(1-x) + u\,(1-2x) + \mathcal{O}\left(u^2, s_i^2\right) \tag{2.5}$$

2.2.2 Expansion of the Master equation

A solution to the Master equation in the above form is not known. However, it is also usually not needed. In biologically meaningful evolution models, one is often interested in a qualified limit of a large population size $N\to\infty$. This limit is to be realized with the condition that the products $N\cdot u =: \mu$ and $N\cdot(s_A - s_B) =: \sigma$ are held constant. It is these combined parameters that separate between qualitatively different behaviors of the system [38]. The limit $N\to\infty$ leads to a diffusion approximation of the Master equation, which is called a Fokker-Planck equation

[40, 41]. The time is then conveniently measured on a different scale. For the details of the derivation see appendix C.

$$\partial_\tau P(x,\tau) = \left[-\partial_x\left(\sigma x(1-x) + \mu(1-2x)\right) + \partial_x^2 x(1-x)\right] P(x,\tau) \tag{2.6}$$

$$\sigma := N(s_A - s_B), \quad \mu := Nu, \quad \tau := \frac{t}{N^2} \tag{2.7}$$

This is a linear second order partial differential equation, which is sometimes easier to solve than the set of coupled ordinary differential equations that is the Master equation. Especially for stationary distributions, i.e. $\partial_t P(x,t) = 0$, there are standard techniques to find them from the so called "drift" and "diffusion" terms appearing in the Fokker-Planck equation [41]. These are the terms appearing in the first order and second order parts of the Fokker-Planck operator, respectively. They are here denoted with $F(x)$ and $D(x)$ (since the drift term can be regarded as a *force* and D is the usual notation for a *diffusion* constant). In the above equation, we have

$$\text{drift term}: \quad F(x) = \sigma x(1-x) + \mu(1-2x) \tag{2.8}$$

$$\text{diffusion term}: \quad D(x) = x(1-x) \tag{2.9}$$

The drift term describes the dynamics of the system if fluctuations can be neglected. Indeed, if the second order derivative term in the Fokker-Planck equation above is omitted, we arrive at a Liouville equation. Its solution is a delta peak fixed to the deterministic trajectory $\phi(\tau)$:

$$\partial_\tau P(x,\tau) = -\partial_x F(x) P(x,\tau), \quad P(x,0) = \delta(x - x_0) \tag{2.10}$$

$$\Rightarrow P(x,\tau) = \delta(x - \phi(\tau)), \quad \text{with} \quad \frac{\mathrm{d}}{\mathrm{d}\tau}\phi(\tau) = F(\phi(\tau)),\ \phi(0) = x_0 \tag{2.11}$$

In our case, this deterministic behavior dominates, if both selection and mutation are strong: $1 \ll \mu, \sigma$. If there is even mutation-selection balance - $\mu \ll \sigma$ - the deterministic trajectory approaches the stationary state ϕ_*:

$$\left.\frac{\mathrm{d}}{\mathrm{d}\tau}\phi(\tau)\right|_{\phi_*} = 0 \quad \Rightarrow \quad \phi_* \approx \begin{cases} 1 - \frac{\mu}{\sigma}, & \sigma > 0 \\ \frac{\mu}{\sigma}, & \sigma < 0 \end{cases} \tag{2.12}$$

Note: Sadly, there is a very confusing double usage of the word "drift" in the populations genetics literature. It is used to refer either to *genetic drift* - the random nature of reproductive success - or to the drift appearing in a Fokker-Planck equation. Both aspects could not be more contradictory: whereas genetic drift emphasizes fluctuations, the Fokker-Planck drift describes the dynamics *in absence of fluctuations*!

2.2.3 Allele frequency distribution in equilibrium

The stationary distribution of the Fokker-Planck equation (2.6) can be given analytically [41, 38] (see figure 2.1)

$$P^s(x) = \frac{\mathcal{Z}^{-1}}{x(1-x)} \exp\left(\int_a^x \mathrm{d}x' \frac{\sigma x'(1-x') + \mu(1-2x')}{x'(1-x')} \right) = \mathcal{Z}^{-1} \left[x(1-x)\right]^{\mu-1} e^{\sigma x} \tag{2.13}$$

where $a \in (0,1)$ is an arbitrary reference value and terms involving it are absorbed in the constant normalization factor $\mathcal{Z}$, which is given by the following integral.

$$\mathcal{Z} := \int_0^1 \mathrm{d}x \, \left[x(1-x)\right]^{\mu-1} e^{\sigma x} = \sqrt{\pi} e^{\sigma/2} \sigma^{\frac{1}{2}-\mu} \, \Gamma(\mu) I_{\mu-\frac{1}{2}}\left(\frac{\sigma}{2}\right) \tag{2.14}$$

where $I_\alpha(x)$ is the modified Bessel function of the first kind (see table of mathematical definitions in the preamble). The normalization for the neutral case $\sigma = 0$ is given by

$$\mathcal{Z}_{\sigma=0} = \frac{\Gamma(\mu)\Gamma(\mu)}{\Gamma(2\mu)} \tag{2.15}$$

For low mutation rates $\mu = Nu \ll 1$, most of the probability weight is concentrated near the boundaries $x(1-x) = \mathcal{O}\left(\frac{1}{N}\right)$ [38]. This means that typical trajectories spend most of the time at either boundary, where the population is monomorphic, i.e. all organisms carry the same allele. This can be most easily seen, for the neutral case by integrating the singular parts of the distribution over half the interval each [38] and leaving out the boundary regions $x(1-x) = \mathcal{O}\left(\frac{1}{N}\right)$:

$$\mathcal{Z}_{\sigma=0}^{-1} \left(\int_{1/N}^{1/2} \mathrm{d}x \, x^{\mu-1} + \int_{1/2}^{1-1/N} \mathrm{d}x \, (1-x)^{\mu-1} \right) = \frac{2\left(2^{-\mu} - N^{-\mu}\right)\Gamma(2\mu)}{\mu\Gamma(\mu)\Gamma(\mu)} = \mu \ln\frac{N}{2} + \mathcal{O}\left(\mu^2\right) \tag{2.16}$$

This means that most of the probability weight for $Nu \ll 1$ is indeed concentrated at the monomorphic boundaries $x = 0$ and $x = 1$. For a typical trajectory this means that, infrequently, the majority allele frequency switches to the other extreme, which is called a "substitution event". This carries over to non-neutral situations with $\sigma \gg 1$. Altogether, the result is that in equilibrium the probability distribution is "U-shaped" (see figures 2.1 to 2.3).

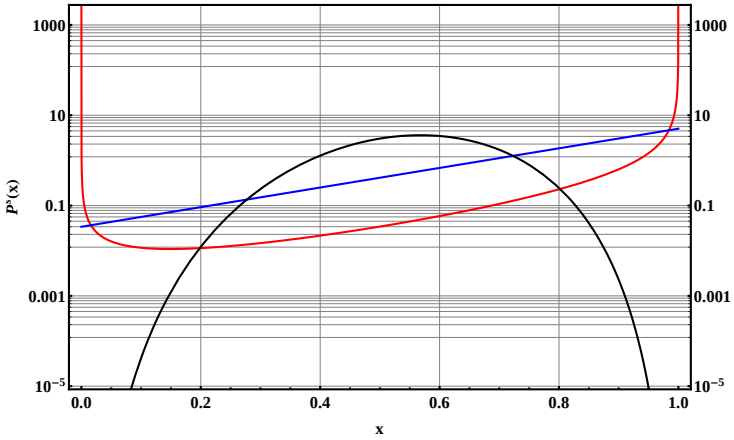

Figure 2.1: Stationary distribution of the one-locus/two-allele model for $\sigma = 5$ and different values of mutation $\mu = 0.1$ (red), $\mu = 1$ (blue) and $\mu = 10$ (black).

2.2.4 Substitution rates in equilibrium

The two substitution events ($A \to B$ and $B \to A$) take place on characteristic time scales, i.e. with typical rates. The most straightforward way to calculate these "substitution rates" is via conditional escape probabilities [41]. Starting in a monomorphic state, e.g. at $x = 0$ (all B), a first A allele appears with a rate Nu. This one A individual will spawn an A-sub-population that might ultimately take over the whole population - instead of dying out - with an escape probability π_{esc}. This conditional escape probability can be evaluated at the level of the Fokker-Planck equation (the boundary at $x = 0$ is not directly involved in that process) and more importantly, further mutations during the escape event can be neglected for a mutation rate that is low enough: $Nu \ll 1$. For general one-dimensional Fokker-Planck equations of the form

$$\partial_t P(x,t) = \left[-\partial_x F(x) + \partial_x^2 D(x)\right] P(x,t) \tag{2.17}$$

with a *drift* term $F(x)$ and a *diffusion* term $D(x)$, the probability $\pi_{L,R}(x_0)$ that a stochastic trajectory starting at x_0 leaves a region bounded by $x_L < x_0 < x_R$ through either end of the interval can be calculated directly using the auxiliary quantity $\psi(x)$ [41].

$$\psi(x) := e^{-\int_a^x \mathrm{d}x' \, \frac{F(x')}{D(x')}} \;\Rightarrow\; \pi_L(x_0) = \frac{\int_{x_0}^{x_R} \mathrm{d}x\ \psi(x)}{\int_{x_L}^{x_R} \mathrm{d}x\ \psi(x)}, \quad \pi_R(x_0) = \frac{\int_{x_L}^{x_0} \mathrm{d}x\ \psi(x)}{\int_{x_L}^{x_R} \mathrm{d}x\ \psi(x)} \tag{2.18}$$

As before, the choice for the reference value a is irrelevant, because terms involving it cancel out in the definitions of $\pi_{L,R}(x_0)$. For the present purpose, we need

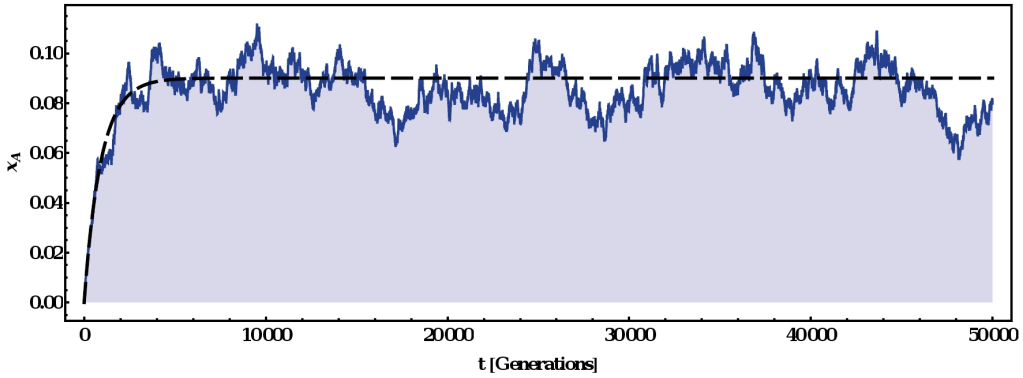

Figure 2.2: Typical trajectory of the one-locus/two-allele model for large mutation $Nu \gg 1$. The black dashed line corresponds to the deterministic trajectory with the same initial condition. In this limit, it is a good approximation to the stochastic trajectory.

the escape probability for an initial A allele ($x_0 = 1/N$) to spawn an A-allele subpopulation of size x instead of dying out, i.e. $x_L = 0$, $x_R = x$. The drift and diffusion terms are taken from the Fokker-Planck equation (2.6), but crucially the mutation rate is set to zero: $u = 0$. This is because for $Nu \ll 1$, the time (in generations) until the next A allele appears due to a mutations is $\mathcal{O}\left(\frac{1}{Nu}\right)$, which is much longer than the time it would take the current clone to grow to fixation *conditional on its success*.

$$\psi(x) = e^{-\int_a^x \mathrm{d}x' \, \frac{\sigma x'(1-x')}{x'(1-x')}} \propto e^{-\sigma x} \tag{2.19}$$

$$\pi_{\mathrm{esc}}(x) := \pi_{x_R=x}(x_0 = 1/N) = \frac{\int_0^{1/N} \mathrm{d}x' \; \psi(x')}{\int_0^x \mathrm{d}x' \; \psi(x')} = \frac{1 - e^{-\sigma/N}}{1 - e^{-\sigma x}} \approx \frac{\sigma/N}{1 - e^{-\sigma x}} \tag{2.20}$$

The total rate $\Gamma_{B\to A}$ for a substitution event from an all-B population to an all-A one is to a good approximation the rate of first arrival of an A allele times its probability of escape to $x = 1$:

$$\Gamma_{B\to A} \approx Nu\,\pi_{\mathrm{esc}}(1) = \frac{u\sigma}{1 - e^{-\sigma}} \tag{2.21}$$

If one takes into account that the mutation rates might not be the same in each direction, one arrives at the following central result:

$$\boxed{\Gamma_{A\to B} \approx \frac{u_{A\to B}(-\sigma)}{1 - e^{\sigma}}, \quad \Gamma_{B\to A} \approx \frac{u_{B\to A}\,\sigma}{1 - e^{-\sigma}}} \tag{2.22}$$

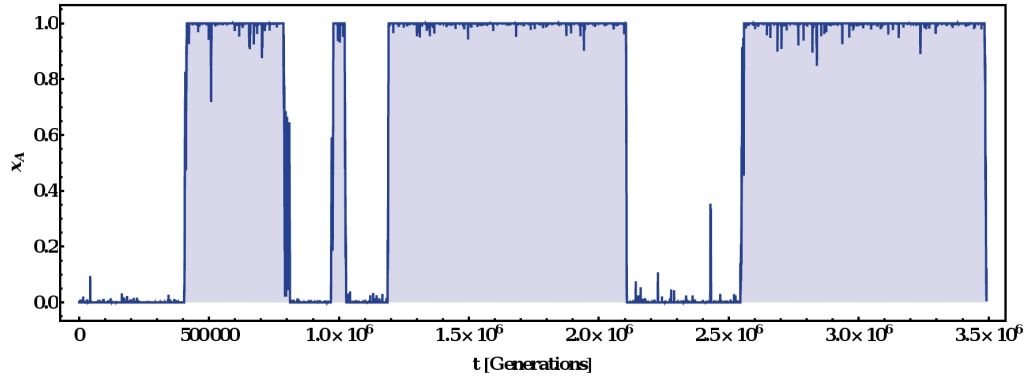

Figure 2.3: Typical trajectory of the one-locus/two-allele model for small mutation $Nu \ll 1$. The deterministic trajectory is clearly not a good approximation. Instead, several substitution events (switches) can be seen. In this simulation, the fitness of the A-allele (upper boundary) is slightly larger than that of the rival, as is reflected in the different amounts of time spent at each boundary.

The ratio of these two directed substitution rates is thus proportional to a rather simple and well known expression, that is central to all derivations in this chapter [36, 42, 43].

$$\boxed{\frac{\Gamma_{B\to A}}{\Gamma_{A\to B}} = \frac{u_{B\to A}}{u_{A\to B}}\, e^{\sigma}} \tag{2.23}$$

Note: The main argument for the simplified derivation of the substitution rate above, was that the time needed for a *successful* A allele clone to ultimately reach fixation is shorter than the time scale for further mutations $\frac{1}{Nu}$. Then we could neglect the mutation process for the subsequent fixation. If the A allele is preferred ($\sigma > 0$), then the conditional fixation time is approximately given by the deterministic time $\mathcal{O}\left(\frac{\ln N}{\sigma}\right)$. But even in the case of $\sigma < 0$, when deterministically the A allele would not fix at all, the time needed for fixation by a large fluctuation is still given by the same deterministic value. This is because the most likely "escape-by-fluctuation-path" (for one dimensional systems) is exactly the time-*reversed* path [44] (also called "anti-deterministic path").

2.2.5 Sampling distribution in equilibrium for $Nu \ll 1$

Up to this point we treated the one-locus/two-allele model as if the model parameters - N, u and s - were known. For *given* parameters, we stated the important quantities and observables, e.g. the stationary distribution $P^s(x)$. Starting with the

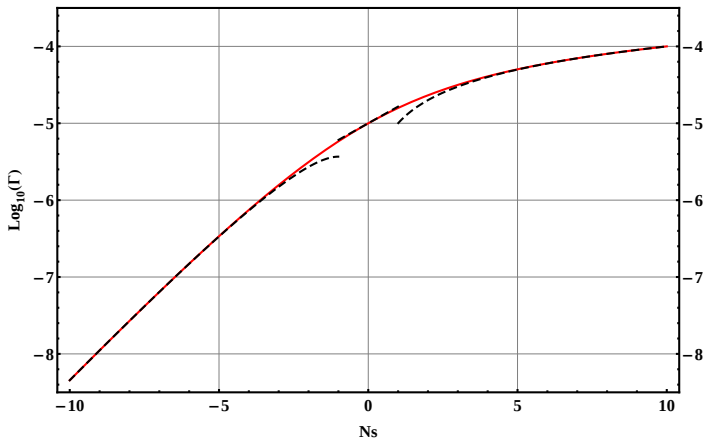

Figure 2.4: The substitution rate $\Gamma(u,\sigma)$ as given by equation (2.21) (red line, with $u = 10^{-5}$) and compared to the three approximations (black dashed) $\Gamma \approx u\sigma e^{-\sigma}$ (for $\sigma \ll -1$), $\Gamma \approx ue^{\sigma/2}$ (for $|\sigma| \ll 1$) and $\Gamma \approx u\sigma$ (for $\sigma \gg 1$).

next section, the task will be actually to infer/estimate those model parameters in situation when they are *unknown*. This can be done if "experimental data" for this model is available. This data could be, for example, the time series for a particular trajectory of allele frequencies, such as in figure 2.3. If (and that is a big if) the time span covered by the data was long enough, such that many substitution events are observed, one could use eq. (2.23) for the inference: the time spent in each monomorphic state depends on the selection coefficient σ and the mutation rates. If the mutation rates were known from independent sources, one would thus be able to estimate the selection coefficient σ directly from such a single trajectory. The problem of course is that ordinary germline evolution - which formed the DNA of humans and all other animals - proceeds on time scales which are absolutely inaccessible to contemporary researchers. On this scale, all the DNA data that was collected in the last decade is virtually from a single point in time. All that we have is but a snapshot of evolution[5]. But this shall not be a problem, as long as there are *many independent and identically distributed* (i.i.d.) data points available. Consider again the trajectory in figure 2.3 and imagine there were many - in fact infinitely many - such trajectories overlaid in the picture. All these trajectories are supposed to follow exactly the same rules, such that they form a statistical ensemble which represents the stationary distribution $P^s(x)$. If you were to take a snapshot of that ensemble at any point in time, you would still be able

[5]Of course, I am not talking about "experimental evolution", nowadays performed with bacteria, where tens of thousands of generations can be observed in a matter of years.

to reconstruct the model parameters. But instead of using time spans spent in the monomorphic states by *one* representative, you would use the proportions of the *ensemble* at the two boundaries at that single time point.
In reality, the ensemble as a whole is also not available. What is usually available is just a small number of representatives from the ensemble of all populations. Or even less: from every populations in the sample just a single individual each. In the substitution regime ($Nu \ll 1$), the population is at any given point in time most likely to be monomorphic, anyway. Sampling more individuals from the same population would be redundant. For every single population, θ_A is defined as the probability that a randomly chosen individual is of allele type A. It is for $Nu \ll 1$ given by:

$$\theta_A := \int_0^1 \mathrm{d}x\, x P^s(x) \approx \frac{\Gamma_{B\to A}}{\Gamma_{A\to B}+\Gamma_{B\to A}} + \mathcal{O}(Nu), \quad \theta_B := 1-\theta_A \tag{2.24}$$

where the stationary distribution was approximated by a sum of two delta peaks at $x = 0$ and $x = 1$ and a "polymorphic part" with weight $Nu \ll 1$. As we have seen before, the main contribution comes from the monomorphic state at $x = 1$. Now what is the probability to find exactly k_A alleles of type A and k_B of type B when taking a randomly chosen individual from each one of K populations? It is just the binomial distribution:

$$P(k_A, k_B \mid \theta_A, \theta_B) = \frac{(k_A+k_B)!}{k_A!\,k_B!}\theta_A^{k_A}\,\theta_B^{k_B} \tag{2.25}$$

Now, the allele counts (k_A, k_B) constitute the "data". After what was said before, it should be clear that the bias seen in such data reflects exactly the evolutionary bias stemming from the fitness difference and mutation rates:

$$\frac{k_A}{k_B} \xrightarrow{\text{sample size K}\to\infty} \frac{\theta_A}{\theta_B} = \frac{\Gamma_{B\to A}}{\Gamma_{A\to B}} = \frac{\mu_{B\to A}}{\mu_{A\to B}}\, e^{\sigma} \tag{2.26}$$

If further independent knowledge of the bare mutation rates $u_{A\to B}$, $u_{B\to A}$ was available, one would be able to extract the sought-after fitness effect of a mutation $B \to A$ directly

$$\ln\frac{k_A}{k_B} - \ln\frac{u_{B\to A}}{u_{A\to B}} \xrightarrow{\text{sample size K}\to\infty} \sigma \tag{2.27}$$

The problem is, that one cannot have an infinitely large sample. The true fixed state probabilities $\theta_{A,B}$ are unattainable, they can only be approximated from a finite sample of size K: $\theta_i \approx k_i/K$. Moreover, if the sample size K is in fact small - as is often the case - it might well happen that one does not observe an allele ($k_i = 0$). This would lead to immediate problems in equation (2.27). The next sections

try to address this issue with the method of Bayesian inference. Practically, in order to estimate selection coefficients from biological DNA sequence data, we need a method that provides reasonable estimates even for small sample sizes. Moreover, the bare mutation rates $u_{i\to j}$ must be included as external information. The Bayesian method is tailor made for such tasks.

2.3 Bayesian inference

The task set in the last section is a stereotypical problem of statistical inference: how can one deduce the unknown parameters of a mathematical model from a finite sample of it? The theory to address this problem is known as Bayesian inference [7, 45]. Before explaining its main concepts, some nomenclature and notation is needed:

model: A "model" is usually a prediction of measurable events in the form of a probability distribution. A simple example used throughout this chapter is the multinomial sampling distribution for drawing colored balls from an urn (with replacement).

parameters: The "model" always depends on a set of parameters, which are denoted by $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$. These parameters characterize the model fully but are usually unknown. In the urn example, the parameters would be the actual frequencies of the different colors in the urn.

sample: The sample could also be called "measurement", "experiment" or "observation". This is the set of data that is directly accessible to the experimenter by a measurement etc. The data can in principle be everything, depending on the specifics of the model. The sample is in general denoted by X. In this work, it will almost always be a set of integer numbers, denoted by $\boldsymbol{k} \in \mathbb{N}^n$. The model attaches a probability $P(\boldsymbol{k}|\,\boldsymbol{\theta})$ to all possible samples $\boldsymbol{k}$. The notation is meant to emphasize that this prediction depends on the value of $\boldsymbol{\theta}$.

posterior: The posterior distribution $P(\boldsymbol{\theta}\,|\,\boldsymbol{k})$ expresses the knowledge about the value of the parameter $\boldsymbol{\theta}$ *after* having made the observation $\boldsymbol{k}$. Importantly, it is a *bona fide* probability distribution for $\boldsymbol{\theta}$. Its spread reflects the amount of uncertainty left after an experiment. If the sample size is small, this spread can be significant. The posterior distribution is usually the end result of an inference problem. It contains everything that is known about the model at that time.

prior: It becomes clear immediately, that the posterior distribution expresses a

gain of knowledge about the model through the sampling experiment. For logical consistency, the state of knowledge *before* the experiment must be stated as well. All information about the model available before the experiment (denoted by I_0) is encapsulated in the prior distribution $P(\boldsymbol{\theta}\,|\,I_0)$. The prior information can be e.g. some specific values or boundary conditions.

point estimate: In applications of Bayesian inference, a concrete value for the model parameter is often desired. This current best estimate is denoted by $\hat{\boldsymbol{\theta}}$ and must be derived from the posterior distribution. Although there are sometimes obvious candidates, such as the posterior mean or modal value, the particular choice is subject to the needs of the user. The derivation of point estimates is not strictly part of the Bayesian inference scheme. Later in this chapter, a particular method to find point estimates more systematically will be presented.

The central equation lying at the heart of Bayesian inference is Bayes' theorem for conditional probabilities for events A and B (not to confuse with the allele labels of previous sections).

$$P(A\,|\,B)\,P(B) = P(B\,|\,A)\,P(A) \quad \Rightarrow \quad P(A\,|\,B) = \frac{P(B\,|\,A)\,P(A)}{P(B)} \tag{2.28}$$

Applied to the problem of model parameter estimation introduced above, it connects the posterior distribution with the prior knowledge and the measurement. (The symbol in the last term signifies that the parameter space may be either discrete or continuous.)

$$P(\boldsymbol{\theta}\,|\,\boldsymbol{k}, I_0) = \frac{P(\boldsymbol{k}\,|\,\boldsymbol{\theta})\,P(\boldsymbol{\theta}\,|\,I_0)}{P(\boldsymbol{k}\,|\,I_0)}, \quad P(\boldsymbol{k}\,|\,I_0) = \int\!\!\!\!\!\!\sum_{\boldsymbol{\theta}} P(\boldsymbol{k}\,|\,\boldsymbol{\theta})\,P(\boldsymbol{\theta}\,|\,I_0) \tag{2.29}$$

The Bayesian methodology is a consequent application of the laws of logic to the problem of inference [7]. It is related but not equivalent to other approaches for parameter estimation, such as maximum likelihood.

INFO **The maximum likelihood method** provides a point estimate of the unknown parameter θ by *maximizing* the sample probability $P(X \mid \theta)$ for a given measurement X with respect to θ. It is thus looking for the value that makes the observation at hand most likely. For this task, it considers the sample probability above as an ordinary function of θ - also called the likelihood-function $\mathcal{L}(\theta) := P(X \mid \theta)$. From the Bayesian perspective, this method is the special case of a uniform prior: $P(\theta \mid I_0) = \text{const}$. In this case, the likelihood-function is proportional to the posterior distribution and the modal value - where the maximum is attained - of both coincide. This modal value is but one commonly used point estimate. For the likelihood-function, it is the only one available. The uniform prior is also called "uninformed" prior for reasons that will shortly become clear.

The most attractive feature of the Bayesian ansatz is that it provides a probability distribution over all possible model parameters that includes all accessible knowledge about the model. It is thus well suited to handle even very small sample sizes. But this probabilistic nature is also its main weakness. In practice, the estimation of parameters is often just an intermediate step. Usually, one then wants to continue working with the best parameter estimates available. This step is often called the "decision step" [7]. But the Bayesian analysis does not provide point estimates. The Bayesian final output is always a distribution[6]. If the sample size is *large enough* (which needs to be quantified), the posterior distribution will be peaked and centered around the true parameter value and a reasonable point estimate would be the mean or modal value (also called the maximum a posteriori, MAP). But for small sample sizes, these estimates might not be very representative of the posterior distribution. This decision problem is addressed by the principle of minimum discrimination information [46], which will be discussed later in this chapter.

2.3.1 A simple example: The highest number in the urn

To demonstrate the essential steps of the Bayesian inference protocol, we will now present a very simple set-up: consider an urn that includes balls which are *numbered*, such as in lottery drawings. The prior information is exactly the following: there is at least one ball, but not more than $N_{\max}$ balls in the urn. The exact number of balls is $1 \leq N \leq N_{\max}$. The balls are numbered successively from 1 to N without gap. The task is to take a single ball from the urn and infer from that data N, the

[6] This is in my opinion the main reason, why methods such as maximum likelihood are still widely used

highest number in the urn. We will now translate this set-up in the mathematical quantities introduced in the last section.

- The **parameter** of the model is the total number of balls in the urn (or equivalently the highest number printed on any ball), which we will denote with N, i.e. we have $\theta := N$.

- The **model** is given by the probability to draw the ball with the number k from the urn, i.e. we have for the general outcome $X := k$ and

$$P(X \mid \theta) := P(k \mid N) = \begin{cases} \frac{1}{N}, & 1 \leq k \leq N \\ 0, & k > N \end{cases} \tag{2.30}$$

- The **prior** distribution must reflect the fact that $I_0 : \ 1 \leq N \leq N_{\max}$. We can immediately guess the prior in this case:

$$P(\theta \mid I_0) := P(N \mid I_0) = \begin{cases} \frac{1}{N_{\max}}, & 1 \leq N \leq N_{\max} \\ 0, & N > N_{\max} \end{cases} \tag{2.31}$$

- The **posterior** distribution for a *particular* draw of the number k must now be calculated as in equation (2.29).

$$P(N \mid k, I_0) = \frac{P(k \mid N)\, P(N \mid I_0)}{P(k \mid I_0)} \tag{2.32}$$

  The only ingredient missing is the normalization factor in the denominator of the right hand side:

$$P(k \mid I_0) = \sum_{N=1}^{\infty} P(k \mid N)\, P(N \mid I_0) = \frac{1}{N_{\max}} \sum_{N=1}^{N_{\max}} P(k \mid N) \tag{2.33}$$

$$= \frac{1}{N_{\max}} \sum_{N=k}^{N_{\max}} \frac{1}{N} =: \frac{1}{N_{\max}} \left(H_{N_{\max}} - H_{k-1}\right) \tag{2.34}$$

  where in the last equation we used the common notation for the harmonic number $H_n := \sum_{m=1}^{n} \frac{1}{m}$. Altogether, we have now the formula for the posterior distribution:

$$P(N \mid k, I_0) = \begin{cases} \left(H_{N_{\max}} - H_{k-1}\right)^{-1} \frac{1}{N}, & k \leq N \leq N_{\max} \\ 0, & N > N_{\max} \end{cases} \tag{2.35}$$

The set-up and the calculations were admittedly very simple, but they demonstrate some of the essential characteristics of Bayesian inference. It is especially worth noting, that the end result - the posterior distribution - is automatically properly normalized. It is also *intuitively* the correct result: drawing a ball with a "24" printed on it rules out the possibility that the highest number in the urn is anything *below* 24. The posterior accordingly attaches zero probability to these cases.
The question of which value of N should now be used as a **point estimate** $\hat{N}$ is outside the realm of orthodox Bayesian inference. In any case, it should be noted that the posterior distribution in this example does by no means *suggest* a particular choice. The highest (modal) value is achieved at $N = k$. The posterior mean value is nowhere close to that.

$$\langle N \rangle = \frac{N_{\max} - (k-1)}{H_{N_{\max}} - H_{k-1}} \tag{2.36}$$

For example, with $N_{\max} = 100$ and $k = 24$, we would have $\langle N \rangle \approx 53$ (see figure 2.5).
As a last comment, the maximum likelihood approach would suggest, according to equation (2.30), to take $\hat{N} = k$. It is the one "parameter" value compatible with the outcome that maximizes the likelihood-function $\mathcal{L}(N) = \frac{1}{N}$.

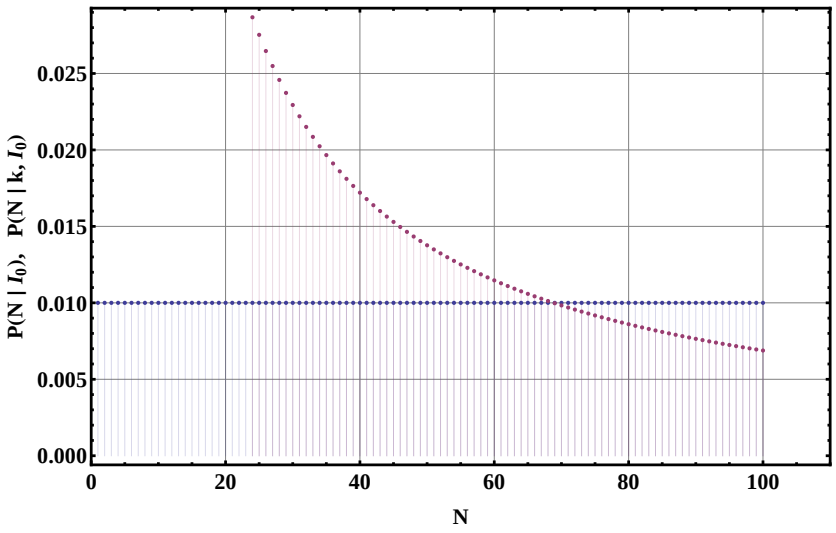

Figure 2.5: The prior and posterior distribution of the example of Bayesian inference explained in the text for $N_{\max} = 100$ and $k = 24$.

2.3.2 Bayesian inference of multinomial weights: a case study

As a second example of more immediate importance, consider the problem of estimating the frequencies $\boldsymbol{\theta} = \{\theta_i\}_{i=1\ldots n}$ of balls of n different colors in a given (big)

urn from a finite sample of its content. This particular example is central to the estimation of germline fitness effects from samples of biological DNA sequences later and will be discussed in some detail. The set-up is adapted from the standard textbook by Durbin *et al.* [47].
Assume first, that you are given the prior information I_0: the presented urn is but one of many similar urns. All these urns have one thing in common: they originate from the same "urn factory". The frequencies $\boldsymbol{p} = \{p_i\}_{i=1\ldots n}$ of the colors when pooling *all* existing urns of this type is supposed to be known (e.g. by knowledge of the amounts of different dyes used in the urn factory). But the process of assigning balls to urns that is carried out in the factory is completely unknown. Practically this means that, if you were forced to make an estimate $\hat{\theta}_i$ of the frequency of color i for your particular urn before *any* sample, your answer would be based on the prior knowledge I_0 alone, i.e. your guess would be $\hat{\theta}_i = p_i$.
To be more precise, if there were only two colors - black and white - and you were told that $p_1 = p_2 = 0.5$, then this could mean any of the following: (a) in every individual urn - including the one in front of you - there is presumably an equal number of black and white balls; (b) for every urn that contains only white balls, there is another urn containing only black balls such that the global balance is maintained; (c) any configuration in between these two extremes. Before *measuring* the configuration of the urn presented to you by a sample of its content, there is no way of telling what the actual frequencies in it are. You only know that the first ball you draw from the urn is as likely to be black as to be white. This is exactly how any prior information I_0 in this set-up should be understood.

Translated into the language introduced previously, the **model** is here a multinomial distribution, i.e. the probability of drawing $\boldsymbol{k} = \{k_i\}$ balls of colors $i = 1, \ldots, n$ when sampling from the urn[7], assuming that the color frequencies $\boldsymbol{\theta} = \{\theta_i\}$ are *known*:

$$P(\boldsymbol{k}|\,\boldsymbol{\theta}) := \frac{K!}{\prod_{i=1}^{n} k_i!} \prod_{i=1}^{n} \theta_i^{k_i}, \quad K := \sum_{i=1}^{n} k_i \tag{2.37}$$

The **prior** distribution should, above all, be a distribution over the space of all possible frequency configurations. This space is the $(n-1)$-dimensional simplex Δ^{n-1} with the general definition

$$\Delta^n := \left\{ \boldsymbol{\theta} \in \mathbb{R}^{n+1} \;\middle|\; \sum_{i=1}^{n+1} \theta_i = 1,\ \forall i = 1, \ldots, n+1 :\ 0 \le \theta_i \le 1 \right\} \tag{2.38}$$

[7] The urn is assumed to be so large that there is no distinction between drawing with or without replacement, although the derivation can be modified to take this into account [7].

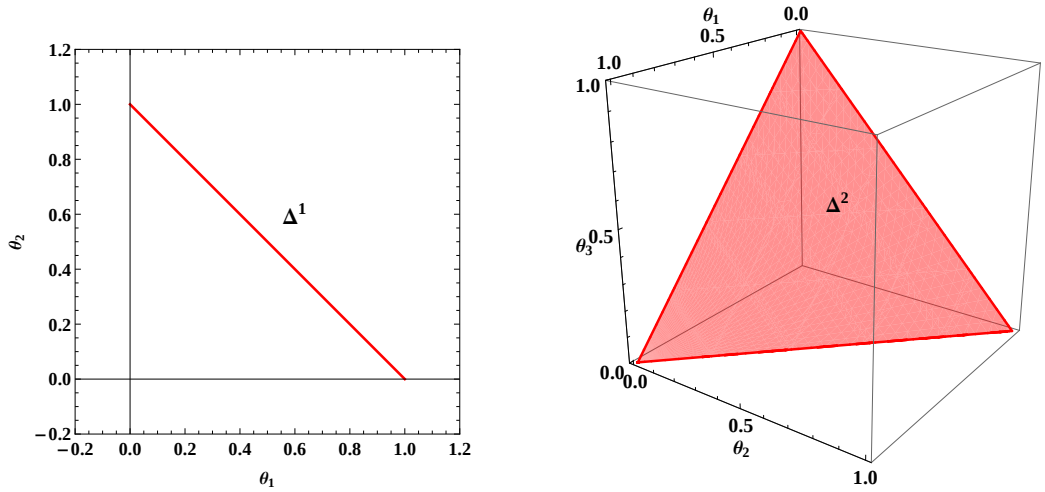

Figure 2.6: Illustrations of the simplex $\Delta^1 \subset \mathbb{R}^2$ (left) and the simplex $\Delta^2 \subset \mathbb{R}^3$ (right).

The simplex Δ^{n-1} is a $(n-1)$-dimensional object embedded in the n-dimensional space $\mathbb{R}^n$. There are different ways to model a prior distribution on the simplex, but the standard approach for a multinomial distribution model is to use a prior distribution of Dirichlet type to encode any information [47].

$$\mathrm{Dir}(\boldsymbol{\theta} \,|\, \boldsymbol{\alpha}) := \frac{1}{\mathrm{Beta}(\boldsymbol{\alpha})} \prod_{i=1}^{n} \theta_i^{\alpha_i - 1} \tag{2.39}$$

$$\mathrm{Beta}(\boldsymbol{\alpha}) := \frac{\prod_{i=1}^{n} \Gamma(\alpha_i)}{\Gamma(A)}, \; \boldsymbol{\alpha} \in \mathbb{R}^n_{\geq 0}, \; A := \sum_{i=1}^{n} \alpha_i$$

The prior Dirichlet distribution depends itself on real and positive parameters $\boldsymbol{\alpha} = \{\alpha_i\}_{i=1\ldots n}$. These parameters must be adjusted according to the prior information: $\boldsymbol{\alpha} = \boldsymbol{\alpha}[I_0]$. The normalization of the Dirichlet distribution is expressed in terms of the n-dimensional beta function $\mathrm{Beta}(\boldsymbol{\alpha})$. To repeat, we here choose the prior distribution to be of Dirichlet type, i.e.

$$P(\boldsymbol{\theta} \,|\, I_0) = \mathrm{Dir}(\boldsymbol{\theta} \,|\, \boldsymbol{\alpha}[I_0]) \tag{2.40}$$

INFO

The Dirichlet distribution: In principle, any probability distribution on the simplex would qualify as a prior. For example, a "histogram" of previous measurements from related models could serve as a prior. But more often than not, the actual prior information I_0 is not that specific. Usually, one only knows about the possible *range* of the parameter (here the simplex) and some background values, which would be the "best guess" before an experiment.
The class of Dirichlet distributions is compatible with most prior information of that kind. Furthermore, this class is flexible in the sense that it includes distributions that are sharply peaked ($\forall i : \alpha_i \gg 1$), uniform ($\forall i : \alpha_i = 1$) or localized at the edges ($\forall i : \alpha_i \ll 1$). But most importantly, this prior allows for analytical calculations and it is *conjugate* to the multinomial distribution, which means that for any multinomial sample the posterior distribution will again be of Dirichlet type. The mean of the i-th component of $\boldsymbol{\theta}$ follows from the normalization:

$$\langle \theta_i \rangle = \int_{\Delta^{n-1}} \mathrm{d}^n\theta \; \theta_i \; \mathrm{Dir}(\boldsymbol{\theta} \,|\, \boldsymbol{\alpha}) = \frac{\mathrm{Beta}(\boldsymbol{\alpha} + \boldsymbol{e}_i)}{\mathrm{Beta}(\boldsymbol{\alpha})} = \frac{\alpha_i}{A} \tag{2.41}$$

with the unit vector $(\boldsymbol{e}_i)_j = \delta_{ij}$. It is a special property of the Dirichlet distribution, that all the mean values are invariant under a scaling of the parameters $\boldsymbol{\alpha} \to a \cdot \boldsymbol{\alpha}$, $a \in \mathbb{R}_{>0}$. Thus, a Dirichlet distribution is not unambiguously specified by its mean values alone.
To calculate the entropy of the Dirichlet distribution, we also need the logarithmic mean $\langle \ln(\theta_i) \rangle$, which can be expressed through the digamma function $\psi(x) = \frac{\mathrm{d}}{\mathrm{d}x}\Gamma(x)$:

$$\langle \ln(\theta_i) \rangle = \int_{\Delta^{n-1}} \mathrm{d}^n\theta \; \ln(\theta_i) \; \mathrm{Dir}(\boldsymbol{\theta} \,|\, \boldsymbol{\alpha}) = \psi(\alpha_i) - \psi(A) \tag{2.42}$$

The second step of the inference program is the measurement of a sample $\boldsymbol{k}$ of size $K := \sum_{i=1}^{n} k_i \geq 1$, and the update of the parameter probability distribution of the frequencies $\boldsymbol{\theta}$ to the **posterior** [47].

$$P(\boldsymbol{\theta} \,|\, \boldsymbol{k}, I_0) = \frac{P(\boldsymbol{k} \,|\, \boldsymbol{\theta})\, P(\boldsymbol{\theta} \,|\, I_0)}{P(\boldsymbol{k} \,|\, I_0)} \tag{2.43}$$

As in the previous example, we still need to calculate the normalization factor $P(\boldsymbol{k} \,|\, I_0)$. Together with the sampling probability eq. (2.37) and the Dirichlet

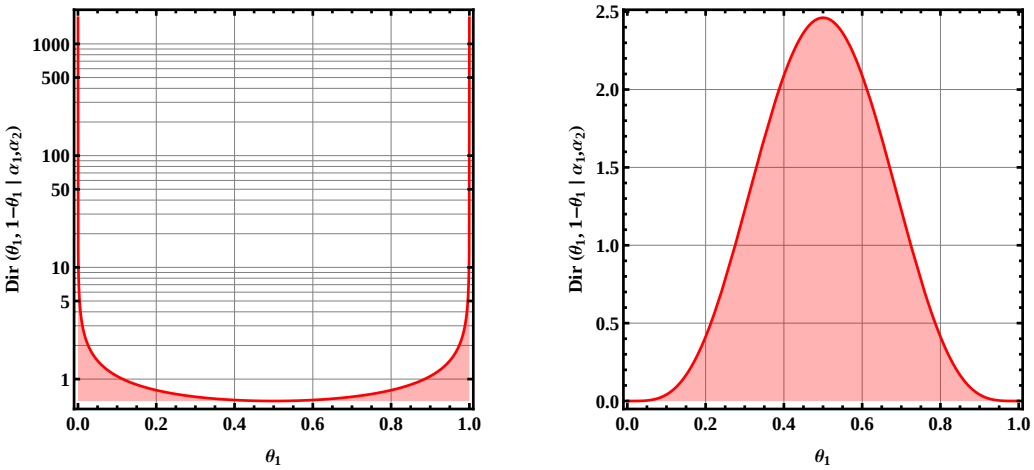

Figure 2.7: Two Dirichlet distributions on the one-dimensional simplex Δ^1 (two colors) for $\boldsymbol{\alpha} = (0.5, 0.5)$ (left) and $\boldsymbol{\alpha} = (5, 5)$ (right). Note that the mean value is identical.

prior eq. (2.40) this yields the posterior.

$$P(\boldsymbol{k}|I_0) = \int_{\Delta^{n-1}} \mathrm{d}^n\theta \; P(\boldsymbol{k}|\boldsymbol{\theta})\, P(\boldsymbol{\theta}|I_0) \tag{2.44}$$

$$= \int_{\Delta^{n-1}} \mathrm{d}^n\theta \; \frac{K!\cdot\Gamma(A)}{\prod_{i=1}^n k_i!\cdot\Gamma(\alpha_i)} \prod_{i=1}^n \theta_i^{k_i+\alpha_i-1} \tag{2.45}$$

$$= \frac{\Gamma(A)\,K!}{\Gamma(A+K)} \prod_{i=1}^n \frac{\Gamma(\alpha_i+k_i)}{\Gamma(\alpha_i)\,k_i!} \tag{2.46}$$

Inserting this factor and the definitions above immediately leads to the result

$$P(\theta|\boldsymbol{k}, I_0) = \mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha}+\boldsymbol{k})\,. \tag{2.47}$$

This is in principle the end result of the Bayesian inference protocol. It is remarkably easy to remember: the prior information acts by adding *pseudo-counts* $\boldsymbol{\alpha}[I_0]$ to the actual measurement $\boldsymbol{k}$, which can be understood as imaginary samples from a background or null-model distribution reflecting I_0. For very large sample sizes, this regularization becomes ever less important.

Note: The concept of pseudo-counts is well known in the context of sequence analysis [47, 48] and there has been some debate about an appropriate choice for them [47]. Suggestions went from $\alpha_i = p_i$ [49] to $\alpha_i = \sqrt{K}\,p_i$ [49, 34] and to more complicated constructs using substitution probabilities [48]. We will see in the next sections that the Bayesian method suggests a particular choice of pseudo-counts.

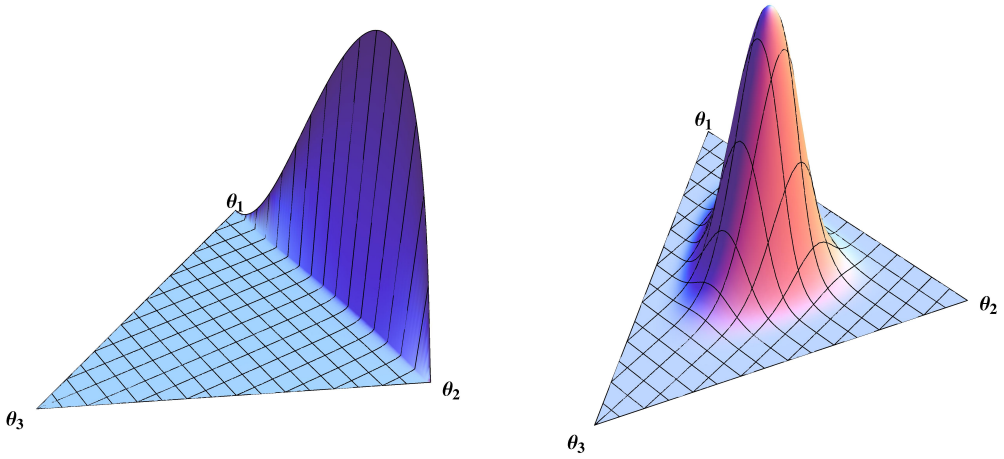

Figure 2.8: Two Dirichlet distributions on the two-dimensional simplex Δ^2 (three colors) for $\boldsymbol{\alpha} = (2.1, 3.1, 0.5)$ (left) and $\boldsymbol{\alpha} = (10, 8, 7)$ (right).

To complete the inference program, there now remain two problems: how should one choose the prior Dirichlet parameters $\boldsymbol{\alpha}[I_0]$ for given I_0? And what should one use as a point estimate $\hat{\boldsymbol{\theta}}$ for the frequencies $\boldsymbol{\theta}$? The posterior mean value is not necessarily the best choice for all sample sizes:

$$\langle \theta_i \rangle_{\boldsymbol{k}, I_0} = \frac{\alpha_i + k_i}{A + K} \xrightarrow{\forall i: k_i \gg 1} \frac{k_i}{K} \xrightarrow{K \to \infty} \theta_i \tag{2.48}$$

2.3.3 Choosing the prior I: the principle of maximum entropy

How should one choose the prior parameters $\alpha_i[I_0]$? Remember, that I_0 states that the global background frequencies are given by $\boldsymbol{p} = \{p_i\}$ (refer to the discussion in the beginning of the last section). One arguably reasonable choice would be to have Dirichlet parameters $\boldsymbol{\alpha}$ such that the *prior mean value* is p_i, i.e.

$$\text{choose } \boldsymbol{\alpha}[I_0], \text{ such that } \quad \langle \theta \rangle_{I_0} = \frac{\alpha_i}{A} = p_i \quad \Rightarrow \quad \alpha_i = A\, p_i \tag{2.49}$$

This prescription leaves on quantity undetermined: the sum $A = \sum_i \alpha_i$[8]. There is nothing more in the prior information I_0 to set this value. One ansatz to solve this problem is to use the principle of maximum entropy (MaxEnt) as suggested by Jaynes [45, 7, 50]:

[8]This quantity A is called *concentration parameter* in the case of uniform $\{p_i\}_{i=1\ldots n}$.

Principle of maximum entropy (discrete case): The one distribution which represents the current state of knowledge I about a system and its constraints best, is the one which maximizes the entropy $H(\theta)$

$$H(\theta) := -\sum_{\theta} P(\theta \mid I) \ln P(\theta \mid I) \tag{2.50}$$

Principle of maximum entropy (continuous case): The one distribution which represents the current state of knowledge I about a system and its constraints best, is the one which maximizes the relative entropy $H(\theta)$

$$H(\theta) := -\int \mathrm{d}\theta \; P(\theta \mid I) \ln \frac{P(\theta \mid I)}{m(\theta)} \tag{2.51}$$

where $m(\theta)$ is called the "invariant measure" by Jaynes, proportional to the limiting density of discrete points. In most cases, it is simply proportional to the uniform distribution on the space in question.

In the two versions of the principle, the "current state of knowledge" I can be either the prior information I_0 alone or in conjunction with experimental data. Coming back to the problem of the undetermined parameter A of the Dirichlet prior, the MaxEnt principle suggests to choose the one A (if it exists) which maximizes the entropy $H(A)$.

$$A_{\mathrm{MaxEnt}} := \operatorname*{argmax}_{A} \; H(A), \quad \text{with } \boldsymbol{\alpha}[I_0] = A\,\boldsymbol{p} \tag{2.52}$$

$$\begin{aligned}
H(A) &:= -\int_{\Delta^{n-1}} \mathrm{d}^n\boldsymbol{\theta} \; P(\boldsymbol{\theta} \mid I_0) \; \ln P(\boldsymbol{\theta} \mid I_0) && (2.53)\\
&= -\int_{\Delta^{n-1}} \mathrm{d}^n\boldsymbol{\theta} \; \mathrm{Dir}(\boldsymbol{\theta} \mid \boldsymbol{\alpha}[I_0]) \; \ln \; \mathrm{Dir}(\boldsymbol{\theta} \mid \boldsymbol{\alpha}[I_0]) && (2.54)\\
&= \ln \mathrm{Beta}(\boldsymbol{\alpha}) - \sum_{i=1}^{n} (\alpha_i - 1) \langle \ln \theta_i \rangle && (2.55)\\
&= \ln \mathrm{Beta}(\boldsymbol{\alpha}) - \sum_{i=1}^{n} (\alpha_i - 1) \left(\psi(\alpha_i) - \psi(A)\right) && (2.56)\\
&= \ln \mathrm{Beta}(\boldsymbol{\alpha}) - \sum_{i=1}^{n} (\alpha_i - 1) \, \psi(\alpha_i) + (A - n) \, \psi(A) && (2.57)\\
&= (A - n) \, \psi(A) - \ln\Gamma(A) + \sum_{i=1}^{n} \left[\ln\Gamma(A\,p_i) - (A\,p_i - 1) \, \psi(A\,p_i)\right] && (2.58)
\end{aligned}$$

where $\psi(x) = \frac{d}{dx} \ln\Gamma(x)$ is the digamma function and the dependence of the entropy on the parameter A is made explicit in the last line. This entropy must be

maximized with respect to A. In the special case of *unbiased* prior weights, i.e. $p_i = \frac{1}{n}$, the parameter A_{MaxEnt} can be found analytically:

$$0 \stackrel{!}{=} \frac{\mathrm{d}}{\mathrm{d}A} H(A) = (A-n)\,\psi'(A) - \sum_{i=1}^{n} p_i\,(p_i A - 1)\,\psi'(p_i A) \tag{2.59}$$

$$\stackrel{p_i = 1/n}{=} (A-n)\left[\psi'(A) - \frac{1}{n}\,\psi'\left(\frac{A}{n}\right)\right] \tag{2.60}$$

This is trivially fulfilled for $A = n$. Thus, we have

$$I_0 : \ \forall i : \ p_i = \frac{1}{n} \quad \Rightarrow \quad A_{\text{MaxEnt}} = n \quad \Rightarrow \quad \alpha_i = 1 \tag{2.61}$$

This corresponds to the uniform distribution on the simplex, which is of course what we expected intuitively. To repeat, if there is no bias in the prior information, the maximum entropy principle tells us to add one pseudo-count to the sample for every color in the urn. However, for general prior frequencies $\boldsymbol{p}$, one needs to solve eq. (2.59) for A, which can only be done numerically. If one insists on using the mean value of the parameter distribution as a point estimate, then consistency demands that $\alpha_i = A\,p_i$ and the "confidence" parameter A must be found as described here. But so far, there is no special reason why the mean value should be used as a point estimate in the first place. The next section will address this issue more closely.

2.3.4 Choosing a point estimate: the principle of minimum discrimination information (MinDI)

After drawing the sample $\boldsymbol{k}$, we are left with the posterior distribution $P(\boldsymbol{\theta} \,|\, \boldsymbol{k}, I_0)$. What are guidelines to decide for a point estimate $\hat{\boldsymbol{\theta}}$ of the true (but unknown) parameter $\boldsymbol{\theta}$? One ansatz [7] is to employ loss functions $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ for this purpose. These loss functions are mostly subjective, for they express a quantity to be minimized in choosing a concrete estimate. In any case, the prescription for the decision step is to take the one estimate $\hat{\theta}$ (if it exists) that minimizes the mean loss to be expected under the posterior.

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}} \langle L \rangle(\boldsymbol{\theta}), \quad \langle L \rangle(\boldsymbol{\theta}) := \int \mathrm{d}^n \boldsymbol{\theta}'\, L(\boldsymbol{\theta}, \boldsymbol{\theta}')\, P\left(\boldsymbol{\theta}' \,\middle|\, \boldsymbol{k}, I_0\right) \tag{2.62}$$

A typical loss function is e.g. the square distance to the true parameter [7].

$$\text{e.g.} \quad L_2(\boldsymbol{\theta}, \boldsymbol{\theta}') = \sum_{i=1}^{n} (\theta_i - \theta_i')^2 \tag{2.63}$$

This particular choice leads to the mean value of the posterior distribution as the point estimate [7]. In any case, all loss functions are - to some degree - subject to the demands of the experimenter who wants to use the point estimates later on. One way to introduce a more objective loss function is to use the discrimination information [51].

$$L_{\mathrm{DI}}(\boldsymbol{\theta}, \boldsymbol{\theta}') := D_{\mathrm{KL}}\left(P(\boldsymbol{k}|\,\boldsymbol{\theta})\,\middle|\,P\left(\boldsymbol{k}|\,\boldsymbol{\theta}'\right)\right) \tag{2.64}$$

INFO

The Kullback-Leibler divergence $D_{\mathrm{KL}}(p\,|q)$ between two distributions p and q is used to measure their similarity [52]. For discrete or continuous one-dimensional distributions, it is defined by:

$$D_{\mathrm{KL}}(p\,|q) := \sum_i p_i \ln \frac{p_i}{q_i} \tag{2.65}$$

$$D_{\mathrm{KL}}(p\,|q) := \int_{-\infty}^{\infty} \mathrm{d}x\; p(x) \ln \frac{p(x)}{q(x)} \tag{2.66}$$

with a straight-forward generalization to higher dimensions. It is $D_{\mathrm{KL}}(p\,|q) = 0$, if and only if $p = q$, else it is $D_{\mathrm{KL}}(p\,|q) > 0$. However the measure does not qualify as a distance, since it is not symmetric in p and q (although one could easily construct a symmetric variant) and does not fulfill the triangle inequality.

The Kullback-Leibler divergence has an information-theoretic meaning in terms of coding efficiency [52]. In this context, q is some kind of "approximation" to the "true" distribution p. The Kullback-Leibler divergence - using base-two logarithms - is then the expected number of extra bits that have to be used when encoding a datum x with a code that is optimal for the distribution q, instead of using a code for the "correct" distribution p.

This abstract interpretation becomes more useful, if we assign the prior and posterior distributions of Bayesian inference to the roles of q and p, respectively. If a new piece of information I_1 is available, then $D_{\mathrm{KL}}(P(\theta\,|\,I_1, I_0)\,|P(\theta\,|\,I_0))$ is the amount of *useful* information that is gained by I_1 over the prior information I_0. The Kullback-Leibler divergence is therefore a central quantity of Bayesian inference.

The loss function $L_{\mathrm{DI}}(\boldsymbol{\theta}, \boldsymbol{\theta}')$ has a very concrete interpretation: it is the amount of information that a subsequent sample $\boldsymbol{k}$ would yield in order to *discriminate* between the two models $P\left(\boldsymbol{k}|\,\boldsymbol{\theta}'\right)$ and $P(\boldsymbol{k}|\,\boldsymbol{\theta})$ (obviously, we have $L(\boldsymbol{\theta}, \boldsymbol{\theta}) = 0$). When choosing a conservative point estimate $\hat{\boldsymbol{\theta}}$, the objective must be to minimize this potential for discrimination. Of course, if the true parameter $\boldsymbol{\theta}$

were *known*, one would simply set $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$. In reality, all that is actually known after the experiment is in the posterior distribution. The quantity to be minimized is then the expected discrimination information under the posterior, i.e.

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \int_{\Delta^{n-1}} \mathrm{d}^n\boldsymbol{\theta}'\, L_{\mathrm{DI}}(\boldsymbol{\theta}, \boldsymbol{\theta}')\, P\left(\boldsymbol{\theta}' \,\middle|\, \boldsymbol{k}, I_0\right) \tag{2.67}$$

This is actually a consequent application of the principle of minimum discrimination put forward by Kullback [46] (see also the reviews [53, 54, 55] and references therein) which - for our purposes - can be stated as follows:

Principle of minimum discrimination information (MinDI): Let $P(X \mid I_0)$ be a distribution over the random variable X depending on some (prior) information I_0. If a new piece of information I_1 becomes available, a new distribution $P(X \mid I_1, I_0)$ must be chosen to describe the model for X which is - under the constraints posed by I_1 - as hard to discriminate from the original distribution as possible, i.e. one that minimizes $D_{\mathrm{KL}}\left(P(X \mid I_1, I_0) \,|\, P(X \mid I_0)\right)$.

This principle can be seen as a generalization of the maximum entropy principle in the following sense: in the absence of *any* information, the least informative distribution is certainly the *uniform* distribution on the space in question. In the light of "prior" information I_0 (now meaning prior to an actual subsequent sample) - whatever I_0 is - the MinDI principle suggests to find a new distribution which complies with I_0 but is as "close" to the uniform distribution as possible in the above sense. But this constrained minimization of the Kullback-Leibler divergence between the two is *equivalent* to the maximization of their relative entropy in Jaynes' MaxEnt principle eq. (2.51).
The application of the MinDI principle to Bayesian parameter estimation is mentioned by Soofi [51]. A comprehensive application of these concepts to the problem of multinomial sampling under Dirichlet priors could however not be found, although all the central aspects and ingredients are present in the literature, especially in [55].
Coming now back to the problem of estimating the frequencies of colors by sampling from an urn, the above prescription (2.67) leads to concrete point estimates for the color frequencies. After the sample $\boldsymbol{k}$, the Bayesian posterior distribution is $P(\boldsymbol{\theta} \mid \boldsymbol{k}, I_0) = \mathrm{Dir}(\boldsymbol{\theta} \mid \boldsymbol{\alpha}[I_0] + \boldsymbol{k})$. We have not yet specified the pseudo-counts $\boldsymbol{\alpha}[I_0]$, which will be the topic of the next section. At this point, we are just interested in the point estimate $\hat{\boldsymbol{\theta}}$ that follows from the MinDI principle. Referring to appendix A.3 for the simple but lengthy calculations, we here give the result that the principle of minimum discrimination information (MinDI) dictates as point estimates the values:

$$\text{MinDI}: \quad \hat{\theta}_i = \frac{e^{\psi(k_i+\alpha_i)}}{\sum_{j=1}^{n} e^{\psi(k_j+\alpha_j)}} \xrightarrow{\forall j:\, k_j \gg 1} \frac{k_i+\alpha_i}{K+A}, \quad K = \sum_{j=1}^{n} k_i,\ A = \sum_{j=1}^{n} \alpha_i \tag{2.68}$$

where $\psi(x)$ is again the digamma function. The asymptotic expansion of the digamma function $\psi(x) = \ln(x) - \frac{1}{2x} + \mathcal{O}(x^{-2})$ connects the MinDI estimate above to the conventional posterior mean value, but only for large counts k_i! The estimate above can therefore be seen as a generalization of the more commonly used posterior mean value estimate for small count numbers k_i.

The mean MinDI loss itself, incurred by this particular choice of the parameter is:

$$\text{MinDI}: \quad \langle L_{\text{DI}} \rangle = \psi(K+A) - \ln\left(\sum_{i=1}^{n} e^{\psi(k_i+\alpha_i)}\right) \tag{2.69}$$

This mean loss is a scale for the *quality* of the estimate. The biggest caveat in using point estimates from Bayesian inference is that we actually sacrifice all the information in the posterior distribution. In summarizing the posterior by a point estimate, one needs to quantify the trustworthiness of the estimate separately. This mean loss is one way to do exactly that.

Finally, we are now in the position to connect the derivations up to this point to the original evolutionary model. Recall, that sampling one representative individual from each population in an ensemble of i.i.d. populations amounts to multinomial sampling in the limit of low mutation rates $Nu \ll 1$. The multinomial weights in this set-up will reflect the *selection difference* between the different alleles. Such samples can then be used to infer this selection difference between allele i and j in our simplified evolutionary set-up (see equation (2.27)), i.e. with mutation rates $u_{i\to j}, u_{j\to i}$, we would set

$$\Delta s_{i,j} := s_i - s_j = \ln\frac{\hat{\theta}_i}{\hat{\theta}_j} - \ln\frac{u_{j\to i}}{u_{i\to j}} \tag{2.70}$$

If the prior (allele) frequencies $\boldsymbol{p}$ where chosen to reflect what one would expect under *neutral* evolution, i.e. $\sigma = 0$, then they could be used to replace the ratio of bare mutation rates:

$$\frac{u_{j\to i}}{u_{i\to j}} = \frac{p_i}{p_j} \tag{2.71}$$

Inserting this and the point estimates $\hat{\boldsymbol{\theta}}$ for the actual allele fixed state probabilities $\boldsymbol{\theta}$ suggested by the methods of this section, we get

$$\boxed{\text{MinDI}: \quad \Delta s_{i,j} \;=\; \psi(k_i+\alpha_i) - \psi(k_j+\alpha_j) - \ln\frac{p_i}{p_j}} \tag{2.72}$$

The only piece missing in this formula is the form of the Dirichlet priors $\boldsymbol{\alpha}$. This will be the topic of the next section.

2.3.5 Choosing the prior II: implications of MinDI

How does one encode the prior information I_0 in the Dirichlet priors $\boldsymbol{\alpha}[I_0]$? Logic dictates that *before* we are given any sampling data $\boldsymbol{k}$, our best guess $\hat{\boldsymbol{\theta}}$ for the unknown weights $\boldsymbol{\theta}$ would be based on the prior information I_0 alone, i.e. $\hat{\boldsymbol{\theta}} = \boldsymbol{p}$. If we insist on a Dirichlet prior distribution and be consistent with MinDI guiding the estimation decision at all times, we must choose the Dirichlet weights $\boldsymbol{\alpha}[I_0]$ to ensure

$$\text{pre}-\text{data}: \quad \hat{\theta}_i(\boldsymbol{\alpha}) = \frac{e^{\psi(\alpha_i)}}{\sum_{j=1}^{n} e^{\psi(\alpha_j)}} \stackrel{!}{=} p_i, \quad i = 1,\ldots,n \tag{2.73}$$

rather than to simply set the prior mean values proportional to the p_i, i.e. $\boldsymbol{\alpha} = A\,\boldsymbol{p}$. Additionally, the one degree of freedom left by the set of equations (2.73) must be used to ensure the MinDI principle for the prior, i.e. the constrained minimization of the Kullback-Leibler divergence of the prior w.r.t. the uniform distribution. Therefore, the task is to solve the following set of equations for the quantities $(\{\alpha_i\}, C)$, where C is a proportionality constant that still needs to be determined.

$$e^{\psi(\alpha_i)} \stackrel{!}{=} C\,p_i, \quad i = 1,\ldots,n \tag{2.74}$$

$$C \stackrel{!}{=} \underset{C'}{\text{argmin}}\, D_{\text{KL}}(C') \tag{2.75}$$

$$\text{with} \quad D_{KL}(C) := D_{\text{KL}}\left(P(\boldsymbol{\theta}\,|\,I_0)\,|\,U_{\Delta^{n-1}}(\boldsymbol{\theta})\right) \tag{2.76}$$

where in the last equation, $D_{\text{KL}}(C)$ is defined as the Kullback-Leibler divergence of the prior distribution w.r.t. the uniform distribution on the simplex Δ^{n-1}.

$$U_{\Delta^{n-1}}: \mathbb{R}^n \to \mathbb{R}: \quad \boldsymbol{\theta} \mapsto U_{\Delta^{n-1}}(\boldsymbol{\theta}) = \begin{cases} \Gamma(n), & \boldsymbol{\theta} \in \Delta^{n-1} \\ 0, & \text{else} \end{cases} \tag{2.77}$$

It was actually already calculated in eq. (2.58) (save for a minus sign and an additive constant coming from the normalization of the uniform distribution).

$$D_{\text{KL}}(C) = (n-A)\,\psi(A) + \ln\frac{\Gamma(A)}{\Gamma(n)} - \sum_{i=1}^{n}\left[\ln\Gamma(\alpha_i) - (\alpha_i - 1)\,\psi(\alpha_i)\right] \tag{2.78}$$

The dependence of D_{KL} on C is implicit through the dependence of the $\boldsymbol{\alpha} = \boldsymbol{\alpha}(C)$ via the first set of equations (2.74). This task can be solved numerically for any prior $I_0 = \boldsymbol{p}$ to yield the Dirichlet weights $\boldsymbol{\alpha}[\boldsymbol{p}]$. One important analytic result is that the equations (2.74) and (2.75) imply

$$\forall\, I_0 = \boldsymbol{p}: \quad A = \sum_{i=1}^{n} \alpha_i[\boldsymbol{p}] = n \tag{2.79}$$

for *any* prior of that form. This includes our previous finding, that an uninformed (uniform) prior with $p_i = 1/n$ will yield $\alpha_i\left[p_j = 1/n\right] = 1$. This returns as a prior distribution the uniform distribution on the simplex, as it should be. Thus, we can replace the equations (2.74) and (2.75) by the new set of equations

$$e^{\psi(\alpha_i)} \stackrel{!}{=} C\, p_i, \quad i = 1, \ldots, n \tag{2.80}$$

$$A = \sum_{i=1}^{n} \alpha_i \stackrel{!}{=} n \tag{2.81}$$

Proof. All we need for the proof of statement (2.79) are the two conditions (2.74) and (2.75). From the first one we get:

$$\frac{\mathrm{d}}{\mathrm{d}C}(C\, p_i) = p_i \stackrel{!}{=} \frac{\mathrm{d}}{\mathrm{d}C} e^{\psi(\alpha_i)} = C p_i\, \psi'(\alpha_i) \frac{\mathrm{d}\alpha_i}{\mathrm{d}C} \quad \Rightarrow \quad \psi'(\alpha_i) \frac{\mathrm{d}\alpha_i}{\mathrm{d}C} = \frac{1}{C} \tag{2.82}$$

This together with the second condition implies

$$0 \stackrel{!}{=} \frac{\mathrm{d}}{\mathrm{d}C} D_{\mathrm{KL}}(C) = (n - A)\, \psi'(A) \frac{\mathrm{d}A}{\mathrm{d}C} + \sum_{i=1}^{n} (\alpha_i - 1)\, \psi'(\alpha_i) \frac{\mathrm{d}\alpha_i}{\mathrm{d}C} \tag{2.83}$$

$$= (n - A)\, \psi'(A) \frac{\mathrm{d}A}{\mathrm{d}C} + \frac{1}{C}(A - n) \tag{2.84}$$

$$= \frac{(n - A)}{C} \left(\sum_{i=1}^{n} \frac{\psi'(A)}{\psi'(\alpha_i)} - 1 \right) \tag{2.85}$$

The function $\phi(x) := \frac{1}{\psi'(x)}$ is concave on $(0, +\infty)$ [56, 57], therefore Jensen's inequality [58] yields immediately

$$\phi\left(\sum_{i=1}^{n} \alpha_i \right) \geq \sum_{i=1}^{n} \phi(\alpha_i) \quad \Rightarrow \quad \sum_{i=1}^{n} \frac{\psi'(A)}{\psi'(\alpha_i)} \leq 1 \tag{2.86}$$

where equality in the last line holds only in the particular case of uniform weights $\left(\forall i = 1 \ldots n:\ \alpha_i = \frac{A}{n}\right)$. Thus we always have

$$e^{\psi(\alpha_i)} = C\, p_i \quad \text{and} \quad \frac{\mathrm{d}}{\mathrm{d}C} D_{\mathrm{KL}}(C) = 0 \quad \Rightarrow \quad A = \sum_{i} \alpha_i = n \tag{2.87}$$

□

In summary, for any prior - not just the uniform one - the principle of minimum discrimination information dictates that *in total n* (the number of colors) pseudo-counts must be added to the sample. The individual pseudo-counts α_i per color must be found by solving the coupled equations (2.80) and (2.81) simultaneously and - in general - numerically. The final point estimate for the color frequencies in the urn is then given by equation (2.68).

INFO

Jensen's inequality is a statement for convex (and concave) functions. Let $U \subset V$ be a subset of a vector space V and $\phi : U \to \mathbb{R}$ a real valued function on U, then ϕ is called *convex*, if for any two points $x_1, x_2 \in U$ and for any $t \in [0,1]$, it is

$$\phi(t\,x_1 + (1-t)\,x_2) \;\leq\; t\,\phi(x_1) + (1-t)\,\phi(x_2) \tag{2.88}$$

The function ϕ is called *concave* if $-\phi$ is convex, i.e. the "$\leq$" is replaced by a "$\geq$". Jensen's inequality for the convex function ϕ states that for any set $\{x_1, \ldots, x_n\} \subset U$ and positive weights $(a_1, \ldots, a_n) \in \mathbb{R}^n_{\geq 0}$, it is

$$\phi\left(\frac{\sum_{i=1}^n a_i x_i}{\sum_{i=1}^n a_i}\right) \leq \frac{\sum_{i=1}^n a_i\,\phi(x_i)}{\sum_{i=1}^n a_i} \tag{2.89}$$

where equality holds only if all weights a_i are equal. The corresponding inequality for concave functions is again with the relation in the other direction.

2.3.6 Example: inference of binomial weights

Consider the problem of estimating the frequencies in a urn with balls of *two* colors ($n = 2$). There is really just one parameter to estimate, i.e. $\boldsymbol{\theta} = (\theta_1, \theta_2) = (\theta_1, 1-\theta_1)$. Without further prior knowledge, MinDI dictates that the prior distribution is U_{Δ^1}, the uniform distribution on Δ^1.

$$I_0 :\; p_1 = p_2 = 0.5 \quad \xrightarrow{\text{MinDI}} \quad \alpha_1 = \alpha_2 = 1 \quad \Rightarrow \quad P(\boldsymbol{\theta} \,|\, I_0) = U_{\Delta^1}(\boldsymbol{\theta}) \tag{2.90}$$

On the other hand, the prior information I_0 might tell us that there is *globally*, i.e. in the set of all urns of that kind, an imbalance in favor of color 1, e.g. $p_1 = 0.6$ and $p_2 = 0.4$. In this case, we get numerically from equations (2.74) and (2.75):

$$p_1 = 0.6,\; p_2 = 0.4 \quad \Rightarrow \quad \alpha_1 \approx 1.12203,\; \alpha_2 \approx 0.87797 \tag{2.91}$$

This will skew the prior distribution to larger values of θ_1 in a way to ensure that the MinDI point estimate is $\boldsymbol{\theta} = (0.6, 0.4)$ (see figure 2.9).

To visualize the dependence of the inference on the sample size K, consider an urn with *known* frequencies $\boldsymbol{\theta} = (0.9, 0.1)$. As a prior, assume this time the uniform $p_1 = p_2 = 0.5$. Now, random samples of increasing size are taken from the underlying *true* binomial distribution. For each value of the total sample size K, 10^3 random sample draws are generated and for each individual sample the MinDI point estimate $\hat{\boldsymbol{\theta}}$ is calculated according to equation (2.68) and the mean loss $\langle L_{\mathrm{DI}} \rangle$ according to equation (2.69). The following plots show the means for both quantities over the 10^3 samples. For samples of size $K > 10$, the average point estimate quickly approaches the true value of θ_1. The mean loss incurred by each estimate drops on the same scale.

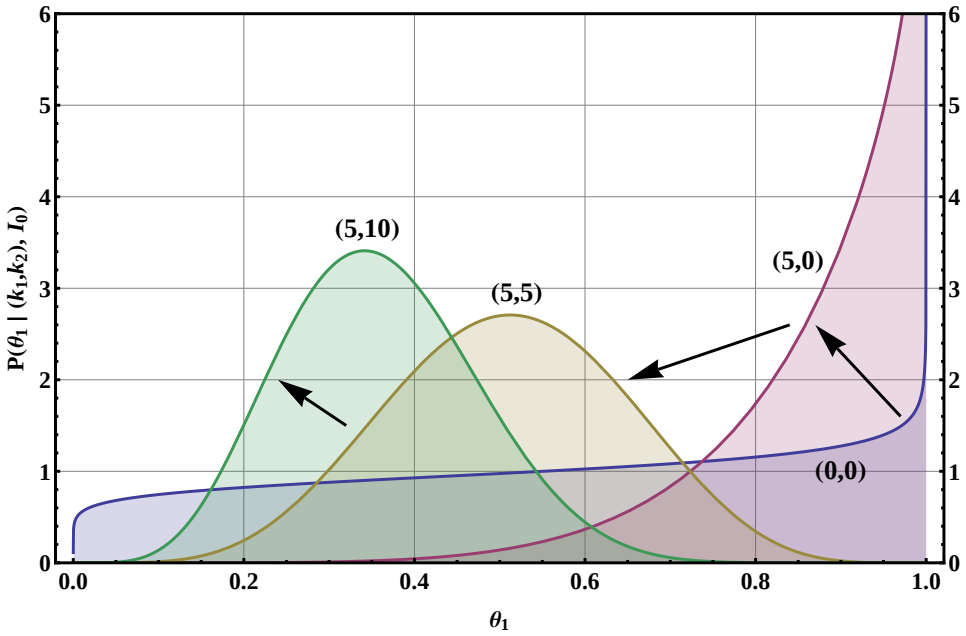

Figure 2.9: Prior distribution (blue) of the binomial weight θ_1 for the biased prior information $I_0 = \{p_1 = 0.6, p_2 = 0.4\}$ as given by MinDI (see text).The other curves show the evolution of the posterior distribution, if the first five draws are all from color one, the next five from color two and the last five from color two as well.

2.4 Germline fitness from protein domain alignments

Now that we have seen how to estimate frequencies of categories (colors) in a big population (urn) from finite samples, the aim of this section is to draw the connection to the problem of germline fitness inference. "Germline fitness" means

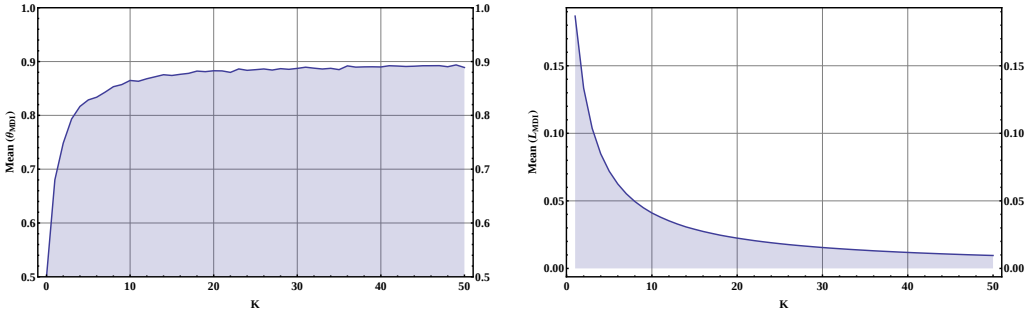

Figure 2.10: Visualization of the dependence of the MinDI point estimate (left) and loss $\langle L_{\mathrm{DI}} \rangle$ (right) on sample size K (see text) for a true value of $\theta = 0.9$.

broadly the fitness landscape that shaped (and still shapes) the makeup of an organism's genome (in evolutionary relevant germline cells, that is). It was argued before, that in the limit of small mutation rates, a certain locus will be mostly monomorphic within a population, such that sampling a single allele at that locus is as good as sampling many. At any given point in time, one needs an *ensemble* of i.i.d. populations to make statistical inferences about the fitness landscape. Table 2.1 points out the analogies and differences between the urn model and the evolutionary set-up.

The remainder of this section will address the most important problem in this table: where can one find statistically meaningful samples from a specific evolutionary model? What is meant by that? To use probabilistic inference, a sample of (statistically) independent genomic loci is needed. In a strict sense, this is impossible! Because any two genes share a common ancestor in the past, all genomes are to a certain degree correlated. Of course, this correlation decreases with the time that two homologous loci are separated, e.g. by a gene duplication or speciation event. Obviously, orthologous loci (e.g. a specific gene) in evolutionary close species - such as human and chimp - are much more correlated than paralogous loci in different genes that stem from a gene duplication event in the distant past and which still carry out similar function, i.e. experience similar selection. There is no clean solution to this conundrum. One good source for evolutionarily meaningful samples are protein domains. To our knowledge, Moses and Durbin [8] were the first to put them to use in that context. We will quickly formulate some arguments in their favor.

urn model	evolution model
• urn	• The set of all loci that experience the same fitness landscape.
• ball	• A population of identical loci (e.g. across many organisms).
• ball color spectrum	• The set of all possible alleles of the locus in question.
• color of a ball	• Frequencies of the alleles. In this sense, a "ball" is not always uni-color but can be in between two "colors" during substitution events.
• sampling distribution (multinomial)	• Stationary allele frequency distribution in equilibrium. For $Nu \ll 1$, it is effectively multinomial.
• sample of K balls	• Sample of K homologous and i.i.d. genomic loci.

Table 2.1: Comparison of the urn model to the set-up of germline fitness inference.

2.4.1 Why protein domains?

One should really start with: what are protein domains? One way to describe them is that they are the *functional atoms* of the genome. Protein domains are recurrent sequence motifs that appear as subsequences in almost all genes [9, 47, 59]. Protein domain sequences are categorized in "families". Domains from one family can often be identified with a certain function, such as posphorylation[9] mediated by the protein kinase domain [9, 60] . Protein domains constitute the most conserved part of the genome. This is mainly due to the fact, that their importance to the cell through their correct function is an excellent lever for natural selection. By the same token, protein domains are also the oldest structures in the genome. For example, there are protein kinase domains found in all eukaryotes, such as worm fly and yeast [60]. Because domain families broadly correspond to one specific function, one can argue that there is a specific fitness landscape for each domain.

[9]This is the transfer of a phosphate group from one molecule, usually ATP, to an amino acid resulting in its conformational change.

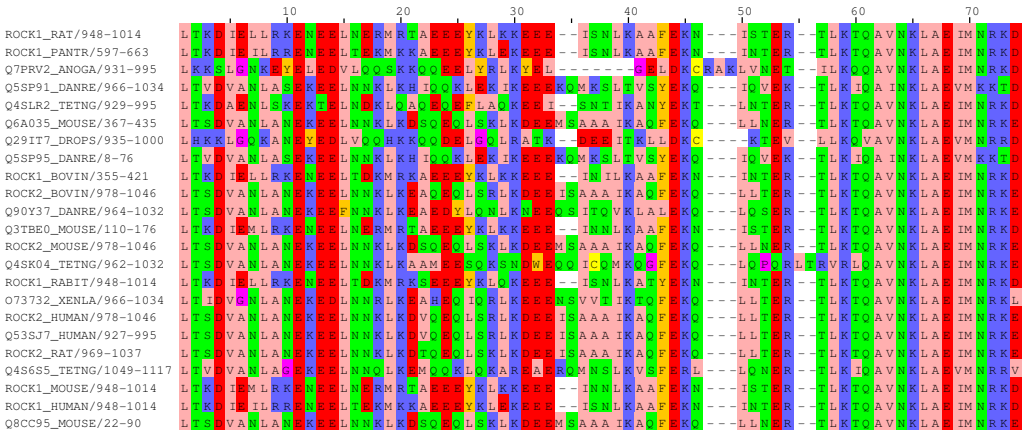

Figure 2.11: The seed alignment for the rho binding protein domain family (Pfam-ID PF08912). Colors correspond to amino acid groups (clustered by their biochemical similarity). The picture is extracted from the Jalview alignment viewer (http://www.jalview.org).

2.4.2 The Pfam database

The reference resource for information about protein domains is the Pfam (protein family) database [9]. Since the latest release 25.0 in April 2011, there are 12,273 families in the database. The database provides comprehensive information about each domain family: function, structure, phylogeny, interactions and relevant references, to name only a few. For our purposes, the most valuable data is an alignment of amino acid sequences for each domain, the so called "seed". Internally, the Pfam database uses a probabilistic program called HMMER that looks for realizations of sequence motifs - called "profiles" - in sequence space [61, 62, 47, 9]. These profiles are fingerprints describing how a typical sequence from that protein domain family should look like. The profiles are derived from an assortment of representative sequences from that family, which are aligned, i.e. "stacked" on top of each other in a fashion where corresponding loci are in the same column (see figure 2.11). Although the seed alignments in Pfam were constructed for a different purpose, one can still use them for the inference scheme presented here. Their purpose actually *is* to be as representative as possible for the domain (function/fitness landscape/distribution) in question, else they would be useless in finding new family members. That's really all one can ask for.

2.4.3 Assumptions and limitations

Now that a source for statistically meaningful samples of sequences is established, a number of assumptions and limitations must be addressed before one goes about the actual germline fitness inference.

- Even with their special purpose, the sequences in the seed alignments of protein domains will still be correlated and hence not statistically independent. One could in principle measure the amount of correlation and - using more elaborate evolutionary models - correct for it. This direction will not be pursued here.

- Since the homologous sequences in an alignment have separated, their function - and with it the selection pressures - might have changed. Although sequences from the same domain family are associated with the same function, their particular task might depend on the genomic and cellular environment they are in.

- Furthermore, the function of a domain family *itself* might have changed over evolutionary time. Altogether, for the purpose of fitness inference this amounts to assuming a stationary (over time) and homogenous (over domain instances) fitness landscape.

- The Kimura model described in the beginning of this chapter describes the evolution of a single locus with two alleles (of different fitness). If one wants to use its predictions, one has to neglect all effects which are outside its realm. In particular:

  - One can only treat a single locus at a time. Therefore, inferences about fitness are made for every locus (column) in a domain motif (alignment) separately. Non-local effects, such as epistasis and recombination are neglected. A different way to express this is that the sequence probabilities are factorizable, i.e. the probability $Q(\boldsymbol{a})$ of the (amino acid) sequence $\boldsymbol{a} = (a_1, \ldots, a_L)$ can be written as $Q(\boldsymbol{a}) = \prod_{i=1}^{L} q(a_i)$. This is consistent with the assumption of infinite recombination.
  - The domain alignments are given as amino acid sequences. Thus, one would have to consider a 20-allele model (the size of the amino acid alphabet used in the genetic code). However, if one assumes that the system is in equilibrium and detailed balance [41] holds at every locus, then the *pairwise* ratio of fixed state probabilities will still be given by the ratio of corresponding substitution rates (as in equation (2.23)).

  - The *neutral* mutation process $u_{i\to j}$ are assumed to be shared by all loci equally. This would mean, that all loci are subject to the same mutational processes (radiation, biochemical reactions and repair/copy errors) in the same way.

- An obvious limitation of using protein domains as a source of germline fitness estimates is that one cannot make statements about loci *outside* of protein domains. About 50% of all protein coding sequence is part of a domain [9]. Thus, about half of the genome is missed.

- With this set-up, only missense (amino acid changing) mutations receive a non-zero score. Silent mutations have score zero (which is reasonable), but non-sense mutations (premature stop-codons) cannot be scored with this ansatz, for they induce a non-local effect.

2.4.4 Background frequencies

One last ingredient missing in the inference protocol is a statement about the bare, i.e. neutral mutation rates between alleles (amino acids). Biochemically, the 20 amino acids are a quite heterogeneous group [59]. For example, a mutation cannot simply change the negatively charged, polar aspartic acid (D) into a non-polar, aromatic phenylalanine (F). On the other hand, it is much simpler to mutate aspartic acid to the very similar glutamic acid (E). The mutation D→F requires at least two nucleotides in the codon to change independently, whereas two of the three possible mutations at the third position in the codon of D result into the mutation D→E. But at this point only *effective* mutation rates are of importance. Does one really need the bare neutral mutation rates themselves? No, as we have seen earlier, one needs only the pairwise ratios of the neutral fixed state probabilities (see equation (2.72)). Under above assumptions, this ratio is exactly equal to the ratio of neutral mutation rates. For this purpose one can take the allele frequencies *in the entire genome*. This effectively averages over all loci and their specific selection forces, leaving only the neutral mutation biases. Table 2.2 lists the amino acid frequencies in the human genome [22]. The 20 letter alphabet of amino acids will be denoted with

$$\mathcal{A} := \{A, C, D, E, F, G, H, I, K, L, M, P, Q, R, S, T, V, W\} \tag{2.92}$$

These *background* or *prior* frequencies are exactly what one would guess to be present at a given locus before actually sampling the allele distribution and gaining locus-specific information. They correspond in all respects to the prior frequencies used in the more conceptual part of the Bayesian inference of multinomial weights earlier.

letter	amino acid	p_i	α_i	letter	amino acid	p_i	α_i.
A	alanine	0.076	1.32	R	arginine	0.065	1.19
N	asparagine	0.032	0.78	D	aspartic acid	0.043	0.92
C	cysteine	0.025	0.69	Q	glutamine	0.045	0.95
E	glutamic acid	0.065	1.19	G	glycine	0.072	1.28
H	histidine	0.026	0.70	I	isoleucine	0.040	0.88
L	leucine	0.098	1.58	K	lysine	0.055	1.07
M	methionine	0.022	0.65	F	phenylalanine	0.035	0.82
P	proline	0.070	1.25	S	serine	0.081	1.38
T	threonine	0.052	1.03	W	tryptophan	0.014	0.53
Y	tyrosine	0.025	0.69	V	valine	0.057	1.09

Table 2.2: Amino acid frequencies (and associated MDI pseudo-counts) in the human genome [22].

2.4.5 Instruction set for germline fitness inference

Collecting all the pieces from this chapter, concrete instructions on how to estimate the germline fitness effect of a mutation can be set up. Keeping in mind all the caveats mentioned in the last section, this quantity is called a (germline fitness) "score" to express its relation to (but not its equality with) the true fitness cost of a mutation.

- Every protein domain family defines one scoring environment. The following is carried out for every single domain family separately.
- For a family's "seed" alignment (as provided by the Pfam database [9]), calculate first for each column a vector of scores $\{s(a_i)\}_{a_i \in \mathcal{A}}$, with

$$\forall a_i \in \mathcal{A}: \; s(a_i) = \ln\left(\frac{\hat{\theta}_i}{p_i}\right) = \psi(k_i + \alpha_i) - \ln\left(\sum_{a_j \in \mathcal{A}} e^{\psi(k_j + \alpha_j)}\right) - \ln(p_i) \quad (2.93)$$

  where $\psi(x)$ is the digamma function, k_i is the count number of amino acids of type i in the column and α_i is the pseudo-count (see table 2.2).
- Given a concrete mutation in a given gene (e.g. BRAF V600E, i.e. a mutation V→E at amino acid position 600 in the BRAF gene), find the domain family that the locus is part of (if it exists). Then align the target sequence (here the domain instance of the tyrosine kinase family in the gene BRAF) to the seed profile to find the corresponding locus (column) to be used for scoring the mutation.

- Given that you are at the correct locus (column), assign to a point mutation $a_i \to a_j$ ($a_i, a_j \in \mathcal{A}$) the score difference $\Delta s(a_i, a_j)$ by taking the score difference between the two alleles.

$$\Delta s(a_i, a_j) = s(a_j) - s(a_i) = \psi(k_j + \alpha_j) - \psi(k_i + \alpha_i) - \ln \frac{p_j}{p_i} \tag{2.94}$$

$$\Delta s(a_i, a_j) \xrightarrow{k_i, k_j \gg 1} \ln \frac{k_j + \alpha_j}{k_i + \alpha_i} - \ln \frac{p_j}{p_i} \tag{2.95}$$

What this procedure amounts to is essentially a generalization of a position weight matrix (PWM, also called position-specific scoring matrix PSSM) for every domain family seed alignment. PWMs are well known concepts of sequence analysis [32, 33, 34, 48, 63, 64, 65] and were used for motif discovery before the advent of more powerful methods such as profiles and Hidden Markov Models [47]. Initially, they were mostly understood as information theoretic constructs but their relation to evolutionary concepts is now clear [8, 42, 43, 66].

2.5 Testing the reliability of inferred fitness values

At this point, it is in order to mention a possibility to test the correlation of the above fitness *score* with the true *biological* fitness cost [8]. Referring to table 2.1, a piece of data *not* used in the derivation of the scoring scheme was the actual internal state of a population of identical loci (a single "ball"). It is explicitly assumed that the allele found in the target sequence is fixed in the whole population, otherwise the predictions from the evolution model based on substitution rates could not be used. Fortunately, in real-life genetic studies many individuals (patients) are sequenced. Thus one can afford to ask the following question: What is the probability $P_{\mathrm{p}}(m, \sigma)$ to find a polymorphism of fitness effect σ in a sample of m individuals? The Kimura model makes a concrete analytical prediction for this probability [67].

$$P_{\mathrm{p}}(m, \sigma, u) := \sum_{k=1}^{m-1} P(k; m, \sigma, u) := \sum_{k=1}^{m-1} \binom{m}{k} \int_0^1 \mathrm{d}x\, x^k (1-x)^{m-k} P_{\mathrm{fw}}(x) \tag{2.96}$$

$$\text{with} \quad P_{\mathrm{fw}}(x) := \frac{2Nu}{1 - e^{-\sigma}} \frac{1 - e^{-\sigma(1-x)}}{x(1-x)} \tag{2.97}$$

$$\boxed{P_{\mathrm{p}}(m, \sigma, u) = 2Nu \sum_{k=1}^{m-1} \frac{m}{k(m-k)} \frac{(1 - e^{-\sigma} F_1(k, m, \sigma))}{(1 - e^{-\sigma})} =: 2Nu\, f(m, \sigma)} \tag{2.98}$$

P_{fw} is the "forward spectrum" or the limiting density of mutant allele frequencies [67]. $F_1(k, m, \sigma)$ is the Kummer confluent hypergeometric function. Terms with

$k = 0$ and $k = m$ in the sum above are left out, because they would correspond to *monomorphic* samples. Here, the integral is not carried out over the stationary distribution $P_s(x)$ of the process, but rather over the limiting density of *newly introduced* mutants, before they have even reached fixation for the first time. To find this quantity, the original Kimura model is solved with a *reflective* boundary at $x = 1$, such that fixation of the newly segregating variant is prohibited [67]. Importantly, the probability to find a site polymorphic in a population sample of size m is proportional to the scaled mutation rate Nu and a function f depending on σ and m only. This allows for the cancellation of the bare (and unknown) mutation rate altogether by considering the rescaled quantity [8].

$$\frac{P_{\mathrm{p}}(m,\sigma,u)}{P_{\mathrm{p}}(m,0,u)} = \frac{f(m,\sigma)}{f(m,0)} \tag{2.99}$$

This quantity can be approximated in the real polymorphism data by dividing the histogram of polymorphic sites as a function of fitness score Δs by the corresponding histogram of all possible scorable mutations in the target sequence and scaling by the value at $\Delta s = 0$. If the score Δs were completely uncorrelated to germline fitness, the data points would fall on a horizontal line, whereas the theoretical prediction is a sigmoidal curve with lower (higher) polymorphism probability for negative (positive) fitness costs. The comparison of any data to this curve is a meaningful test for the main task of the germline fitness score: predicting the germline fitness cost of mutations. Of course, one cannot expect a perfect agreement in the face of the numerous simplifying assumptions that were being made in the derivation of the scoring scheme (see section 2.4.3).

2.6 Discussion

The focus of this chapter was on the conceptual steps necessary to derive a statistically and evolutionarily meaningful estimate of the fitness cost of mutations. This cost was defined as the change in the fitness contribution of a locus that a mutation induces with respect to a concrete mathematical model of evolution: the one-locus/two-allele model. Within this minimal model, the expected frequency distribution of alleles according to their different fitness values can be stated analytically. This prediction is then used to estimate - with the methods of Bayesian inference - the allele-frequencies in the entire gene pool from a finite sample of it.

Conceptually, this problem was reduced to the set-up of sampling balls from an urn and inferring the unknown frequencies $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ of n different colors. A key element of the Bayesian inference technique is the incorporation of prior information I_0. As guidelines, the principle of maximum entropy (MaxEnt) and the

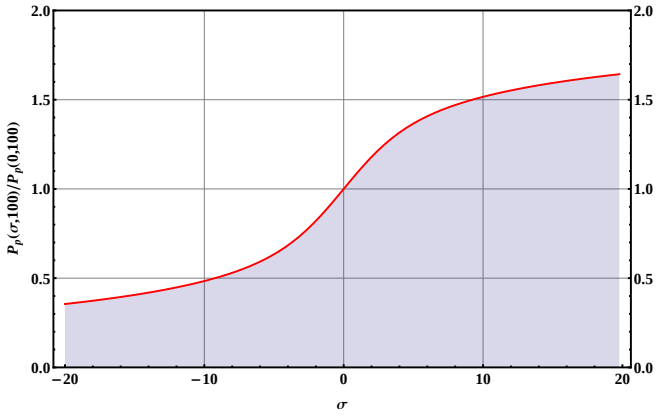

Figure 2.12: Rescaled polymorphism density for a sample of $m = 100$ individuals. The probability to find a site polymorphic is larger than the neutral expectation for beneficial mutations ($\sigma > 0$) and smaller for deleterious mutations ($\sigma < 0$.)

more general principle of minimum discrimination information (MinDI) are employed to find the most objective (conservative) way to express the prior information in terms of a probability distribution over the sought-after model-parameters. As is customary for multinomial sampling problems, the prior distribution was modeled as a Dirichlet distribution, which itself depends on a set of parameters, usually called "pseudo-counts" $\boldsymbol{\alpha}[I_0]$. The biggest advantage of this particular choice is that the posterior distribution - describing the state of knowledge after a sample of counts $\boldsymbol{k} = (k_1, \ldots, k_n)$ from the model - is again of Dirichlet type, but with updated Dirichlet-parameters: $\boldsymbol{\alpha} \to \boldsymbol{\alpha} + \boldsymbol{k}$.

The MinDI principle can be shown to yield objective point estimates $\hat{\boldsymbol{\theta}}$ for the desired model parameters $\boldsymbol{\theta}$ from the posterior distribution. From the Bayesian perspective, MinDI provides an objective loss function that can be used to find point estimates. This in turn suggests a particular choice for the prior pseudo-counts - a choice that is consistent with MinDI. A particularly noteworthy result of the calculation is that the "confidence" carried by the pseudo-counts, i.e. their total sum $A = \sum_i \alpha_i$, is *under all priors* - uninformative or informed - equal to the number of colors: $A = n$.

Both MinDI and its application in Bayesian inference are known concepts in the literature [53, 51], although their use seems not to be widely spread. This present work tried to present the application to the important model of Bayesian inference under multinomial sampling in a comprehensive manner. It is doubtful that the calculations presented here can not be found somewhere in the literature. This

exposition might hopefully serve as a practical tutorial of the essential concepts of Bayesian inference.

The Bayesian inference program was actually only a stepping stone in the greater scheme of things of this work. Its purpose is to provide estimates for the fitness difference of different alleles found in samples of DNA sequence data. The Pfam database [9] was presented as a useful resource for meaningful samples of allele distributions in the form of protein domain alignments. It was shown, how to use this extensive and ever-growing data source to construct an evolutionarily informed mutation-scoring scheme. The limiting assumptions under which one can interpret actual biological sequence data by the predictions of the minimal model of evolution described in this chapter, are explicitly stated. The mutation-scoring scheme will be employed in the next chapter in the analysis of somatic cancer mutations.

Chapter 3

Germline fitness scoring of cancer mutations

"There are two possible outcomes:
if the result confirms the hypothesis,
then you've made a measurement.
If the result is contrary to the hypothesis,
then you've made a discovery."

Enrico Fermi

3.1 Introduction

In the last chapter, the germline fitness scoring scheme was presented as an evolutionarily meaningful way to quantify the functional impact of a mutation. This method will now be put to use in the analysis of somatic cancer variation. Why is this important, and why should it work at all?
Cancer is a disease with origin on the genetic level of cells. The genome of malignant cancer cells is mutated in such a way that they have a fitness advantage over neighboring normal cells. The tumor, i.e. the population of cancer cells, thus eventually grows at the expense of the surrounding tissue, damaging organs and ultimately spreading to the whole body (metastasis). Populations genetic terms such as "fitness" and "population" are appropriate, since cancer progression is often discussed in terms of Darwinian evolution [68]. All necessary ingredients are present: mutational processes (often massively enhanced by loss-of-repair mutations) and selection via competition with the healthy tissue cells and pressure from the immune system. This cancer-specific evolutionary process is not well understood quantitatively [69, 70] (and it may well be cancer-type specific) and there are many avenues to formulate minimal mathematical models of cancer evolution [71, 72, 73, 74, 75, 76].

Most of the recent advances in cancer research were driven by an increasing amount of large scale cancer sequence data with nucleotide resolution [10, 17, 18, 27, 77]. Although cancer genomics has been successful in identifying somatic variants in cancer, it also exposed a serious problem [17]: there are only a few *recurring* mutations that appear in many tumor samples and can be robustly identified as cancer driver mutations, such as e.g. BRAF V600E. Most cancer samples show a much more heterogeneous mutation pattern. Additionally, since most cancers become clinically relevant only in their late developmental stages, their sequence data show an excess of passenger mutations, i.e. mutations that are not directly responsible for cancer progression but were acquired and maintained by the tumor. Importantly, passenger mutations are still cancer *somatic* for they are not found in healthy cells. Thus, one is faced with the statistical challenge to extract the relevant driver mutations from a large background of passenger mutations.

One approach to find genes containing cancer driver mutations is to look for signals of positive selection (beneficial for the tumor that is). Genes under positive selection will show an increase in the rate of substitutions [10]. However, this prediction depends sensitively on the neutral model one is comparing to [78]. The Hidden Markov Model presented in chapter 4 falls into this category of methods. A second complementary approach to find driver mutations is to quantify the functional impact of somatic mutations with bioinformatic measures, e.g. based on conservation, biophysical and structural considerations [28, 29, 30, 79, 80]. These bioinformatic methods are not limited to the analysis of cancer variation. In fact, most were originally introduced to understand germline and common disease variation [80, 81, 82]. A recent review of current scoring methods can be found in [83].

It is important to understand that the germline fitness scoring of cancer mutations [6] falls in the second category for the following reason. Cancer evolution and germline evolution are clearly different. There is *a priori* no reason, why somatic mutations should should be subject to the same evolutionary pressures that shaped the conservation pattern of the germline genome. The natural question then is: Does it help to know the germline fitness effect of somatic mutations in order to identify cancer driver mutations? This question is addressed by the project presented here and published in [6].

The next section will describe the examined data set and the methods used. This is followed by a presentation of the results. Key findings are that it is useful to introduce integrated observables, i.e. scores for loci in the amino acid sequence and scores for genes, and that germline fitness is a meaningful scale for mutations in genes with cancer association, especially for tumor suppressor genes [6].

3.2 Materials and methods

3.2.1 Data set

A widely studied data set of cancer variation in human protein kinase genes [60] as given in [10] is analyzed here. In particular, the data set consists of:

- Reference coding sequences of the 518 human protein kinase genes that were sequenced in the original study [10] (both nucleotide and amino acid sequences). Protein kinase genes are an ideal set of candidate driver genes, for kinases are involved in all essential cell processes, especially in signalling and metabolic pathways. Mutations in kinase genes are known to cause disease and are implicated in cancer [60, 84]. Mutations were defined as deviations from this reference sequence.

- A set of somatic mutations found in 210 cancer samples. Because of the limitations of the scoring method described earlier, only missense mutations were considered.

- A set of germline mutations from the same set of patients. These are variants found in healthy tissue samples. This data was used to test the reliability of the germline scoring method as described in section 2.5. Because some of the variants are in fact polymorphisms (new segregating variants in the patient population), the ancestral allele was decided by comparing to the orthologous chimpanzee reference sequence (and not the human reference, which is just an arbitrary standard) [85]. Of the 142 germline variants, for which this "outgroup polarization" did not decide the ancestral allele, it is estimated that for no more than 10 variants using the human reference leads to an error. These variants were included nevertheless.

- As an external piece of information, a list of candidate cancer genes from [86] (suppl. table 4c therein) was included, for which information about copy number loss and gain in cancer samples is available. This information is used to assign some genes to one of two categories: candidate tumor suppressor (if the rate of loss is greater than the rate of gain) and candidate oncogenes (vice versa). This classification is motivated by figure 3.a in [86]. The score statistics for both categories of cancer genes is studied. Also the set of all cancer associated genes, i.e. the union of above subsets, is considered separately to allow for the possibility that this classification scheme is not adequate. The list should then still be enriched for cancer driver genes.

	opportunity $[10^5]$			somatic			germline		
	all	t. supp.	onco	all	t. supp.	onco	all	t. supp.	onco
total	29.37	3.63	3.68	620	100	83	2423	277	264
scored	14.26	1.78	1.87	324	56	49	1018	125	102

Table 3.1: Number of (available) missense mutations in the different categories. The two categories of candidate tumor suppressor genes and candidate oncogenes are shown separately (for the classification criterion see text).

3.2.2 Formulation of the null model

To assess the relevance of the germline fitness scale for somatic cancer mutations, the mutation data is compared to the following null-model: all cancer mutations are random with respect to germline fitness, i.e. their locations and effects are uncorrelated to the evolutionary conservation of the target sequence. To test the data against this hypothesis, all possible missense mutations away from the reference sequence are constructed *in silico* (by a computer program). This pool of potentially available mutations is called "mutational opportunity space" $\mathcal{M}$. Please note, that the null hypothesis does not make a statement about the actual mutation processes, such as incidence of UV-light etc. These processes most likely *are* ignorant about the target. The statement is about the variation seen in the evolved cancer, i.e. which mutations are tolerated by or relevant to cancer progression.

The mutational opportunity space is used to generate a large number of synthetic replicas of the original mutation set. These replicas are supposed to share all essential characteristics of the true data set, e.g. the total number of missense mutations and the biases in the six mutational channels [10, 18] (see figure 3.1)).

channel	opportunity	somatic	germline
A:T>T:A	17%	7%	5%
A:T>C:G	19%	3%	5%
A:T>G:C	16%	10%	21%
C:G>G:C	19%	13%	11%
C:G>A:T	16%	10%	9%
C:G>T:A	13%	57%	49%

Table 3.2: Mutational biases.

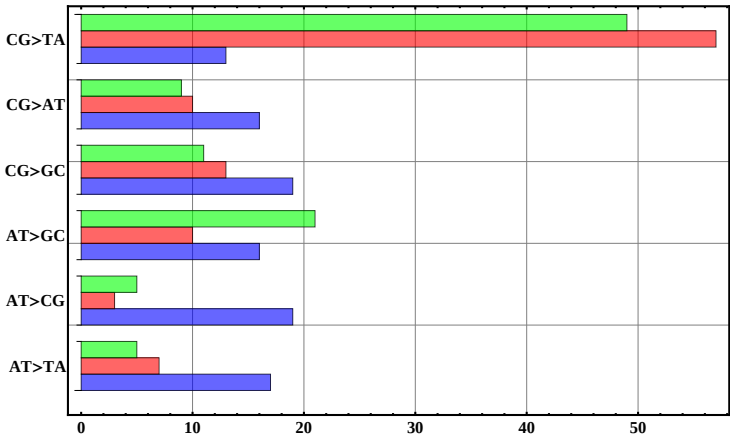

Figure 3.1: Mutation channel biases (in %) in the kinase cancer data: germline variation (green), somatic variation (red) and mutational opportunity (blue). The distribution of mutations over the channels varies with cancer type [10].

3.2.3 Levels of integration

The germline fitness score estimates the functional impact of individual point mutations. In order to compare the impact between different loci or genes as seen in the set of cancer samples, proper genomic observables are needed.

Note: To avoid confusion, the term "locus" means here a specific amino acid position in a gene's translated sequence. Of course, the mutation itself happens in the DNA, such that only nucleotides are ever mutated. But a non-silent mutation does have an effect for the resulting amino-acid sequence of a gene's protein product. In this sense, one can consider the mutation to "fall" on a certain amino acid locus.

For each of the $m = 210$ samples (patients), the kinome sequence (set of kinase genes) of individual k is denoted with $\boldsymbol{a}_k := \{a_{1,k}, \ldots, a_{N^l,k}\}$, where N^l is the length of each kinome, i.e. the total number of loci. Each mutation $a_{i,\mathrm{ref}} \to a_{i,k}$ away from the reference kinome $\boldsymbol{a}_{\mathrm{ref}}$ and at locus i for patient k is assigned the score

$$\Delta s_i(a_{i,\mathrm{ref}}, a_{i,k}) := s_i(a_{i,k}) - s_i(a_{i,\mathrm{ref}}), \quad k = 1, \ldots, m,\ i = 1, \ldots, N^l \tag{3.1}$$

according to equation (2.72). The superscript l always designates a genomic observable on the locus level. The total effect per locus i in the entire set of patients is then defined as:

$$\Delta s_i^l := \sum_{k=1}^{m} \Delta s_i(a_{i,\mathrm{ref}}, a_{i,k}) \tag{3.2}$$

The locus score is a projection over samples. An individual-based analysis would be preferable but is clearly not warranted by the size of the data set. This will hopefully change in the future [5].
Likewise, one can define gene-level observables by integrating the locus-level scores and scaling by l_j^g, defined as the number of available missense mutation in gene j (proportional to the gene's amino acid sequence length).

$$\Delta s_j^g := \frac{1}{l_j^g} \sum_{i \in \text{gene}_j} \Delta s_i^l, \quad j = 1, \ldots, N^g \tag{3.3}$$

where N^g is the number of genes that were sequenced in the data set. The rescaling by the target size allows to compare the scores of genes of different lengths. Δs_j^g is really the effect per locus in gene j. One could in principle define more genomic observables by partitioning the kinome differently, e.g. by genetic pathway or by domain family. We found the locus and gene level to be most informative for this data set.
Finally, a count score c_i^l (c_j^g) is also introduced on the locus (gene) level, that assigns a score of 1 to each mutation, i.e.

$$c(a_{i,\text{ref}}, a_{ik}) := 1 - \delta\left(a_{i,\text{ref}}, a_{ik}\right) \tag{3.4}$$

All results for projected observables must be contrasted against the performance of counting mutations (per locus or per gene) alone.

3.2.4 Scoring of target loci

It has been stated several times before, that the germline fitness score for every missense mutation is derived from a single column in a domain alignment. One can still try to incorporate information about the neighboring sequence around a mutation target. By evaluating the mean germline fitness score of the *reference sequence* in a window of size $l_w = 2w + 1$ ($w \geq 0$) centered around the target locus, one gets a scale for how representative the target really is for that domain.

$$S_i^w := \frac{1}{l_w} \sum_{a_j \in w_i} s_j(a_{j,\text{ref}}), \quad w_i := \left\{a_{i-w,\text{ref}}, \ldots, a_{i,\text{ref}}, \ldots, a_{i+w,\text{ref}}\right\} \tag{3.5}$$

This "locus score" is a property of the target (the reference sequence) alone, not of a specific mutation. In particular, it can be used to weight the mutation scores: $f(S_i^w)\,\Delta s_i(a_{i,\text{ref}}, a_{i,k})$. An interesting candidate for the weight function is the exponential function $f(S) = e^S$. This is motivated by the observation, that the score for a certain amino acid at locus j approximates for the log odds ratio of the

actual fitness-landscape dependent allele weight θ_j and the neutral background frequency p_j (see e.g. equation (2.93)):

$$s_j(a_{j,\mathrm{ref}}) \xrightarrow{K\gg 1} \ln\left(\frac{\theta_j}{p_j}\right) \tag{3.6}$$

$$f(S_i^w) := e^{S_i^w} = \exp\left(\frac{1}{l_w}\sum_{a_j\in w_i} s_j(a_{j,\mathrm{ref}})\right) \xrightarrow{K\gg 1} \left(\prod_{a_j\in w_i}\frac{\theta_j}{p_j}\right)^{1/l_w} \tag{3.7}$$

This choice of $f(S)$ gives for large enough sample sizes K (seed alignment depth) the geometric mean of the likelihood ratio between the locus-specific distribution $\boldsymbol{\theta}$ and the neutral background distribution $\boldsymbol{p}$. For values larger than one, the subsequence is on average more likely to conform to the domain family model than to be a random sequence. If one looks at the distribution of this locus score for $w = 10$ over all amino acid positions in the kinome (see figure 3.2), we see that its mean is larger than one ($\langle e^{S^{10}}\rangle \approx 2.67$), which means that a typical locus in the aligned part of the kinome is actually more likely to be in a sequence block which is well-aligned to the domain model, as it should be when the alignment program works properly.

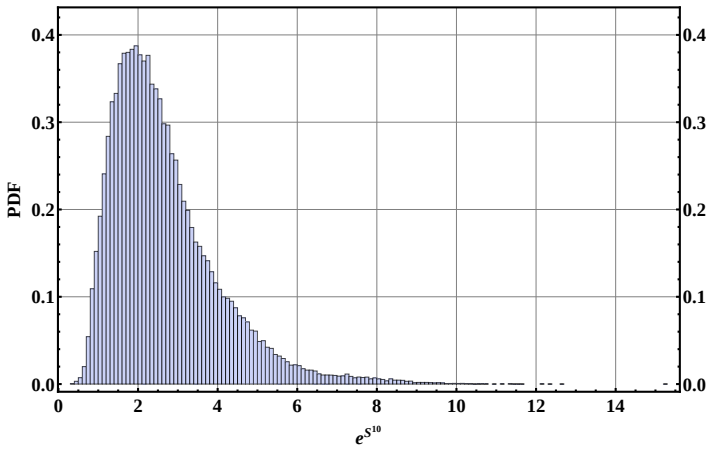

Figure 3.2: Distribution of the locus score $\exp\left(S^{10}\right)$ in mutational opportunity space $\mathcal{M}$. The mean is at $\langle\exp\left(S^{10}\right)\rangle \approx 2.67$.

3.2.5 Analysis pipeline

The scoring pipeline is implemented as a suite of Perl routines. As input, it takes the reference (target) sequence of the data set in both nucleotide and amino acid

space and separate lists of mutations to be analyzed, here the list of somatic and germline mutations. Then the following steps are carried out in sequence:

1. The HMMER program [61] is used to find instances of protein domains in the reference sequence. These domain "hits" are then aligned to the corresponding Pfam seed alignment also with the help of HMMER. This assigns sites in the reference to loci in the domain models.

2. For all domain families with "hits" in the reference, the position specific scoring matrices are constructed.

3. The mutational opportunity space is constructed consisting of all possible missense mutations away from the reference.

4. The opportunity space and the lists of real mutations from the data are then scored. The locus score is also computed at this stage. Since only a subset of the kinome sequence is within a protein domain, only mutations in this subset can be scored. Only this scorable part of opportunity space is used for the analysis.

5. A large number (10^5) of synthetic replicas of a prototype mutation set (somatic or germline) is drawn randomly, but with correct mutation channel biases from the opportunity space. All genomic observables are evaluated and statistics are computed.

Note: In conditioning the analysis on the subset of protein domains within the kinome, we omit a significant signal: The number of mutations falling onto protein domains is already larger than expected by chance (p-value $3.1 \cdot 10^{-2}$). This observation of mutation clustering in protein domains was made earlier [79, 87]. However, we are at this point not interested in mutation clustering but rather in the quantification of their functional effect.

The end output of the analysis pipeline are several distributions of synthetic means for all genomic observables and all scores. Also a list of genes with their score significance is produced.

3.3 Results: germline mutations

3.3.1 Reliability test

First of all, one needs to make sure that the scoring scheme is fulfilling its alleged purpose: to be a measure of the germline fitness effect. Thus, the missense mutations found in the germline variation set were compared to the null hypothesis of

random mutations. A first straightforward way is to compare the score distributions of point events in the germline and null set (see the cumulative distribution in figure 3.3). Clearly, germline mutations are not random w.r.t. to the fitness score. As expected, mutations with a large negative effect (supposedly strongly deleterious) are suppressed. But please be reminded that the scoring methods regards all mutations to be fixed (substitutions) and this is clearly not the case for germline variation. A way to test the reliability of the scoring scheme was described in [8] and here in section 2.5. The corresponding figure 3.4 shows a correlation between the data and the theoretical prediction. If the score were uncorrelated to the true fitness cost, then the data would be on a flat horizontal line. This is clearly not the case. Obviously, the fitness score underestimates the fitness cost of highly deleterious mutations: they appear much more infrequently than predicted by the model. Nevertheless, the degree of agreement is still quite surprising in the light of all the simplifications and rather strong assumptions that went into the scoring scheme (see section 2.4.3). This basically confirms the findings of Moses and Durbin in [8] for this particular data set.

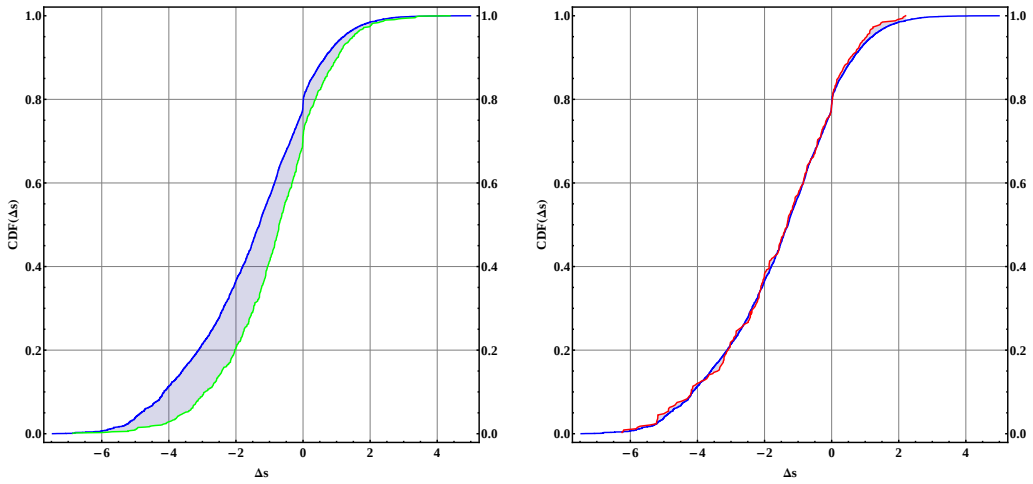

Figure 3.3: Cumulative distribution of germline (left, green) and somatic (right, red) mutation scores vs. the random null model (blue). Mutations with negative fitness effect would be significantly suppressed in the germline, if they were substitutions.

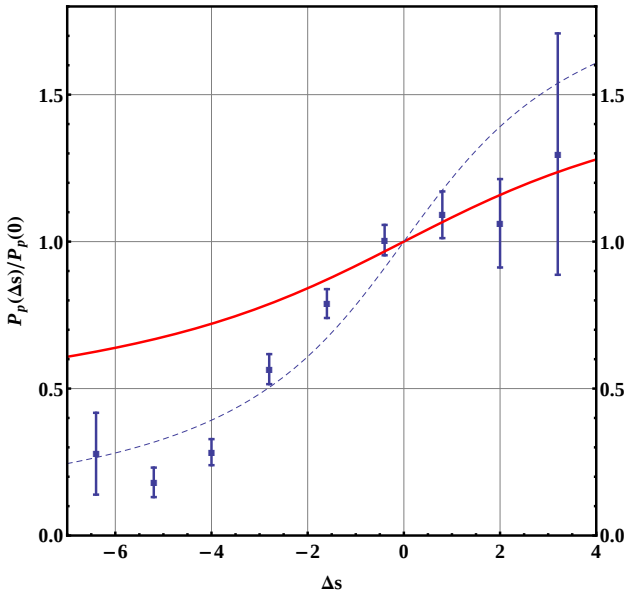

Figure 3.4: Comparison of the germline (scaled) polymorphism probability $P_p(\Delta s)/P_p(0)$ as given by equation (2.98) (red, with $m = 210$) against the same quantity from the data (blue boxes). Error bars are evaluated with $\Delta N_{\text{bin}} = \sqrt{N_{\text{bin}}}$. The dotted line is an error-weighted fit of the data to the sigmoidal function $f(x) := A\left(1 + \frac{\arctan(Bx)}{\pi/2}\right)$ to obtain the scale value at zero (the plot shows $f(x)/f(0)$). (This is an *ad hoc* choice for a stereo-type sigmoidal function without any strong relation to the model.)

3.4 Results: somatic mutations

3.4.1 Most somatic mutations are passengers

The distribution of scores in the set of somatic mutations on the level of point events (without integrating over genomic units) is not significantly different from the random null model (see figure 3.3). This leads to two conclusions:

1. Most somatic mutations are indeed random w.r.t. germline fitness. This is consistent with the picture that most somatic mutations are actually passengers [17, 10, 88]: random mutations that are not essential for the cancer development. They were picked up along the way and never repaired/selected against. This also means that:

2. Cancer cells tolerate a significant mutational load, i.e. mutations that are usually not tolerated by germline cells. This signifies that selection pressures on cancer cells are fundamentally different from those acting on germline cells.

3.4.2 Somatic mutation in cancer genes

The distribution of scores for the genomic units "locus" and "gene" is now considered. To make more concrete statements about the significance of deviations from the null model, the distributions themselves are not compared directly[1]. Rather, the *means* of the distributions are compared. For 10^5 synthetic replicas of the somatic (and germline) mutation set, the mean scores $\langle \Delta s^l \rangle$ and $\langle \Delta s^g \rangle$ are measured over all kinase genes and conditioned on the subsets of tumor suppressor and oncogenes. The means are also conditioned on the union of both subsets, i.e. the set of genes with known cancer association. Then the data mean is compared with the distribution of synthetic means. This yields two quantities: (i) a p-value for the deviation of the data mean from the synthetic means and (ii) an effect size of the data mean, i.e. the data mean divided my the mean of synthetic means. The p-value is a measure for the significance of a deviation from the null model. This depends strongly on the size of the data. The effect size, on the other hand, is a measure for the strength of the deviation and should not depend on data size.

p-value: The probability that under the null-hypothesis an outcome can appear which is at least as "extreme" as the real data point. Here, this probability is estimated from the location of the real data in a histogram of 10^5 samples from the null-hypothesis.

effect size: Defined as the ratio of the mean value of above histogram to the real data value. It is a measure for the size of the deviation, not its significance.

The results of this experiment can be found in table 3.3 (and in table 3.4 for germline variation) and in figure 3.5. The outcome can be summarized as follows:

1. Germline variation is with very high significance not random w.r.t. the fitness scoring scale. Especially in tumor suppressor genes, observed mutations have on average a much smaller deleterious effect than random mutations. Again, this is mainly a sensibility test, for the scheme considers all mutations as substitutions. The more meaningful test compares with the expected polymorphism spectrum (see previous sections).

2. The mean impact per locus in somatic mutations in all kinase genes is somewhat more deleterious than expected by chance (p-value 0.03, effect size 1.14). However, the significance is not strong and the effect size negligi-

[1] Although one could employ measures such as Kullback-Leibler divergence or the Kolmogorov-Smirnov test, which are often too conservative.

ble. This hints at the fact that the somatic mutation set is enriched for high scoring mutations but consists mostly of random passengers [17].

3. The conditional means over cancer-associated genes (as taken from [86]) reveal two important observations: (i) cancer-associated genes harbor many more mutations than expected. This is especially true for tumor suppressor genes (first two rows in table 3.3); and (ii) these mutations have a much stronger germline-deleterious effect than expected. The deviation significance using Δs increases by two orders of magnitude compared to the count score alone.

4. Trying to locate the strong signal in cancer genes, the subsets of (candidate) tumor suppressor genes and oncogenes are analyzed. They both show similar behavior: for both sets, weighting each mutation by its germline fitness score Δs increases the significance of the data by about one order of magnitude compared to the count score alone.

5. The locus score $e^{S^{10}}$ is not clearly relevant for mutations in cancer-genes. If at all, it is relevant for mutations in tumor suppressor genes, where they are somewhat more likely to fall onto conserved and thus functionally relevant targets. The signal is somewhat stronger than the mutation count alone. In oncogenes, there is no such signal.

6. Using the combined weight $e^{S^{10}}\Delta s$, does not further increase the significance of the tumor suppressor data. It seems that there is a large degree of redundancy in both measures Δs and $e^{S^{10}}$.

7. Somatic mutations in candidate oncogenes have a slightly stronger germline fitness effect than the null. Using the score Δs brings the deviation towards the significant regime (p-values of about $3 \cdot 10^{-2}$). They do however show no preference for model-conforming targets.

8. One can evaluate a p-value for each of the 518 protein kinase genes by comparing the data gene score with the 10^5 synthetic replica scores. Of all kinase genes, the gene MAP2K4 stands out with a corrected p-value of 0.025 (after Bonferroni correction[2] for multiple testing of 518 genes, the count score alone is not significant). This method thus predicts MAP2K4 without additional information to be a gene with more germline deleterious

[2]When testing many different and independent hypotheses with a single experiment, one needs to correct the significance levels via e.g. the Bonferroni correction, which means simply decreasing the p-value thresholds with the number of tests. This is necessary, because some of the "significant" outliers might be just the ones we expected by chance anyway.

mutations in cancer than by chance. The role of MAP2K4 as a tumor suppressor in oncogenesis is well recognized [89] and was also noted in [86], from where we derived the cancer gene classification.

		somatic mutations							
		all genes		cancer associated		tumor suppressor		oncogenes	
score	level	p-val.	eff.size	p-val.	eff.size	p-val.	eff.size	p-val.	eff.size
c	locus	N/A	N/A	0.005	1.24	0.005	1.37	0.19	1.11
	gene	0.05	0.95	0.02	1.22	0.02	1.34	0.21	1.12
Δs	locus	0.03	1.14	0.00007	1.65	0.0008	1.80	0.02	1.51
	gene	0.15	1.09	0.0002	1.67	0.001	1.88	0.03	1.47
$e^{S^{10}}$	locus	0.13	1.03	0.003	1.31	0.002	1.50	0.23	1.12
	gene	0.23	0.97	0.01	1.27	0.006	1.45	0.28	1.09
$e^{S^{10}}\Delta s$	locus	0.02	1.17	0.0003	1.69	0.0009	1.95	0.05	1.43
	gene	0.12	1.11	0.0006	1.67	0.0007	2.02	0.10	1.34

Table 3.3: Genomic observables for somatic mutations in all kinase genes, candidate tumor suppressor genes and candidate oncogenes.

3.5 Comparison to other scoring schemes

The results of the germline fitness score Δs are compared to two other widely used mutation scoring methods: the SIFT score s_{SIFT} [26] and the HMMER3 E-values and bit scores s_{HMM} [90]. The SIFT program "sorts intolerant from tolerant" substitutions. Given an alignment it gives a probability score for a substitution, where SIFT < 0.05 is the cutoff for intolerant, i.e. deleterious mutations. (We used the latest version SIFT v4.0.3 together with BLIMPS v3.8.) The HMMER program yields scores for the "goodness-of-fit" of a sequence to a given model (via its HMM profile, here as given from the domain family seed alignment). The difference in bit score between the alignment of the reference sequence $\mathrm{seq}_{\mathrm{ref}}$ to the HMM model and that of the mutated sequence $\mathrm{seq}_{\mathrm{mut}}$, i.e.

$$\Delta s_{\mathrm{HMM}}/\log 2 := \log_2 \frac{\mathrm{P}(\mathrm{seq}_{\mathrm{mut}} \in \mathrm{HMM})}{\mathrm{P}(\mathrm{seq}_{\mathrm{ref}} \in \mathrm{HMM})} - \log_2 \frac{\mathrm{P}(\mathrm{seq}_{\mathrm{mut}} \in \mathrm{Null})}{\mathrm{P}(\mathrm{seq}_{\mathrm{ref}} \in \mathrm{Null})}. \tag{3.8}$$

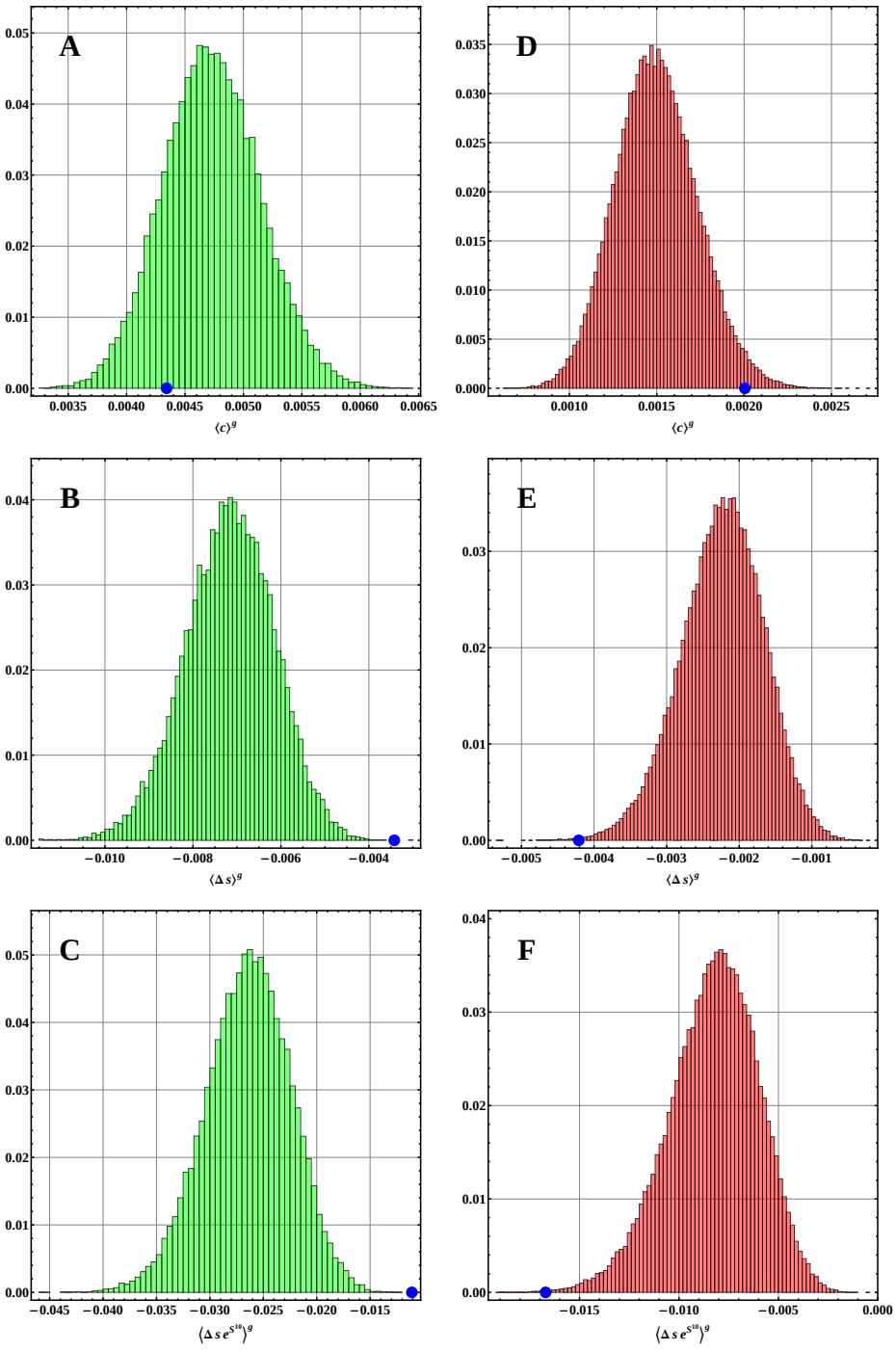

Figure 3.5: Distribution of gene observable means (histograms) against the data mean for tumor suppressor genes (blue dot). Germline data (A-C): germline mutations are not enriched (or suppressed) in tumor suppressor genes (A), they are however significantly less deleterious (B) and are not likely to fall onto highly conserved targets (C). Somatic data (D-F): There is a surplus of mutations in candidate tumor suppressor genes (D), weighting them with their germline fitness effect increases the significance by a factor of 20 (E) (the p-value drops from 0.02 to 0.001, the effect size grows from 1.34 to 1.88). The combined weight increases significance and effect size a bit more (F) ($10^{-3} \rightarrow 7 \cdot 10^{-4}$, $1.88 \rightarrow 2.02$).

		germline mutations					
		all genes		tumor suppressor		oncogenes	
score	level	p-val.	eff. size	p-val.	eff. size	p-val.	eff. size
c	locus	N/A	N/A	0.24	1.05	0.002	0.78
	gene	0.00006	0.93	0.18	0.92	0.0001	0.70
Δs	locus	$< 10^{-5}$	0.60	0.00003	0.51	0.0001	0.56
	gene	$< 10^{-5}$	0.52	0.00004	0.48	0.00005	0.50
$e^{S^{10}}$	locus	$< 10^{-5}$	0.91	0.25	0.94	0.0001	0.69
	gene	$< 10^{-5}$	0.85	0.04	0.84	$< 10^{-5}$	0.61
$e^{S^{10}}\Delta s$	locus	$< 10^{-5}$	0.53	$< 10^{-5}$	0.43	0.00006	0.52
	gene	$< 10^{-5}$	0.47	$< 10^{-5}$	0.42	0.00001	0.47

Table 3.4: Genomic observables for germline mutations in all kinase genes, candidate tumor suppressor genes and candidate oncogenes.

is evaluated internally by the HMMER program and not externally reproducible ("Null" is here a random background model). But conceptually, this is a quantity which is obviously very similar to the germline fitness score. However, it is derived from a complicated and not very transparent probabilistic model (the underlying HMM), whereas the germline fitness score is closely connected to a very concrete evolutionary model. This allows to compare findings to definite predictions and thus to better interpret the data.

To better compare results, both alternative scores are based on the *same* alignments as the germline fitness score. It is noted, that one essential part of the SIFT program is to construct for each query sequence a new alignment from a database such as UniProt/TrEMBL via the alignment algorithm Psi-BLAST [26]. This will naturally lead to different alignments and other scores. The limiting assumptions mentioned earlier in the construction of the germline fitness score still hold.

The results for both quantities are shown in table 3.5. In summary, both HMMER and SIFT score perform comparably well to the germline fitness score in discriminating between the germline and the null model. For somatic variation in the class of candidate tumor suppressor genes, s_{SIFT} gives a lower performance than both Δs and Δs_{HMM}.

		germline all		somatic tsupp.		germline tsupp.	
score	level	p-val.	eff. size	p-val.	eff. size	p-val.	eff. size
s_{SIFT}	locus	$< 10^{-5}$	0.61	0.02	1.61	0.0001	0.46
	gene	$< 10^{-5}$	0.55	0.02	1.73	0.00008	0.43
Δs_{HMM}	locus	$< 10^{-5}$	0.51	0.006	1.75	$< 10^{-5}$	0.43
	gene	$< 10^{-5}$	0.45	0.003	1.91	$< 10^{-5}$	0.39

Table 3.5: Results for s_{SIFT} and Δs_{HMM} scores.

3.6 Somatic mutations in TP53

The standard example of a tumor suppressor gene is TP53. Its recessive role in cancer development was discovered as early as 1989 [91]. The p53 protein, for which the gene codes, is an important regulator of cell division. It is crucially involved in DNA damage repair or - if need be - apoptosis, i.e. programed cell death. Not surprisingly, mutations in TP53 are very common in many cancer types. As a transcription factor, it is especially prone to missense mutations altering its binding efficiency.

Here the fitness score statistics of mutations in TP53 as reported in the COSMIC database [92] are analyzed. Only those mutations are included with confirmed somatic status, from actual tumor tissue and - importantly - from "systematic screens" (as indicated by COSMIC). It is assumed that this label means that the corresponding samples were systematically sequenced in the entire coding sequence. With this filtering, there are 341 missense mutations, of which 337 could be scored. The domains found in the TP53 gene are: the p53 DNA-binding domain (PF00870), the p53 tetramerisation motif (PF07710) and the p53 transactivation motif (PF08563). The statistical analysis is completely analogous to the previous sections. The data is tested against the null-hypothesis of random mutations. In silico, 10^5 replicas from the null are simulated and the mean of means is compared between data and null (since we are considering only one gene, it suffices to look at the locus level.)

The result is striking: The reported mutations are much more germline deleterious than expected by chance. The significance is well below 10^{-5}. The effect sizes are 1.27 for the Δs score alone and 1.37 for the combined score $e^{s^{10}}\Delta s$ (see figure 3.6).

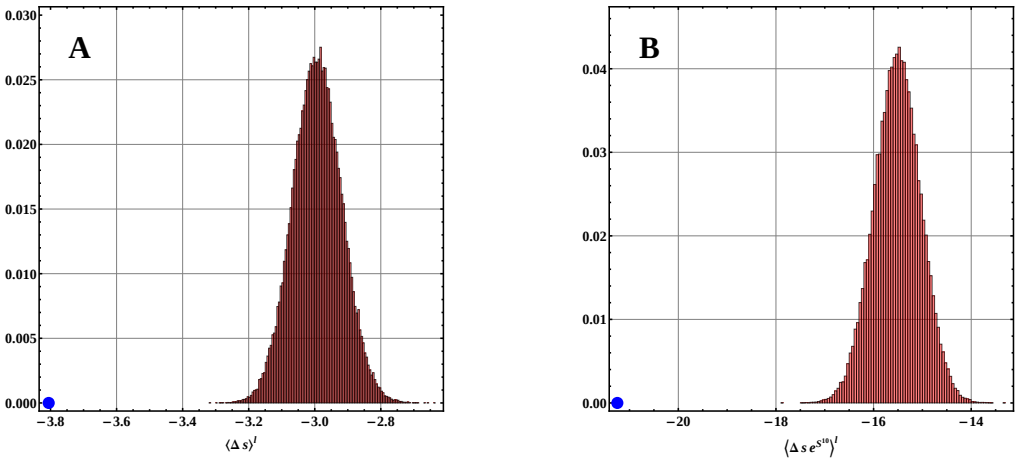

Figure 3.6: Simulation results for the tumor suppressor gene TP53. The mutations reported in COSMIC are significantly more germline deleterious than expected by chance ($p < 10^{-5}$) (A). The combined score increases the signal strength further (B).

3.7 Discussion

In the beginning of this chapter, the following question was asked: how much does it help to know the germline fitness effect of a mutation to understand cancer variation? In particular, does it help to identify cancer driver mutations and driver genes? Protein domains as functional atoms of the genome yield sensible evolutionary conserved sequence models. The domain family seed alignments provided by the Pfam database make it relatively easy to derive conservation based scores that well correlate with the true fitness cost of variation seen in real data [8]. Applied to cancer mutations, this germline fitness scale was shown to be of importance for genes with know cancer association, especially for tumor suppressor genes. These genes are deactivated in cancer cells, which is of course the most likely effect of a strongly germline deleterious mutation. It is tempting to identify the mutations with strongest effect with the cancer drivers. More realistically, this scale will help to prioritize the somatic mutations found in large scale sequencing studies for follow-up analyses.

The method of scoring cancer mutations presented in this work was recently applied in two clinical cancer sequencing studies. In a study on renal carcinoma, mutations in the frequently mutated gene PBRM1 were shown to have a stronger germline fitness effect than random mutations [93]. This score thus added to the identification of PBRM1 as a second cancer driver gene for clear cell renal cancer carcinoma. In a second study on certain blood cancers [94], the gene SF3B1 was identified to be recurrently mutated in patients. However, the mutations observed

were significantly *less* germline deleterious than expected by chance, which lead the authors to speculate that the protein encoded by SF3B1 retains structural integrity in cancer cells, albeit with altered function.
For oncogenes, somatic mutations are also more at the germline-deleterious end of the fitness spectrum. Intuitively, it is much less clear how a mutation confers an activating effect in oncogenes. The mutation might still be deleterious in the sense that it is usually not seen in the germline, irrespective of its functional effect. The weak signal that can be detected in the data pointing in this direction supports this argument. In [80] it is argued, that some of the activating mutations might actually have a germline-beneficial effect (a positive score Δs). In the present data, however, no enrichment of positive scores in the candidate oncogenes could be found. Most likely, the nature of mutations in oncogenes is not easily accessible by the use of germline conservation scales alone. It is also acknowledged, that the classification scheme used to partition the cancer-associated genes might not capture the gene status correctly.
Other studies have shown that known putative cancer driver mutations are more likely to fall onto conserved regions than passengers or neutral polymorphisms [95, 96, 97, 98]. The presented results are consistent with these findings, but the main contribution of this analysis lies in the following clarifications:

- The germline fitness score is connected to a concrete and well-established evolutionary model [8, 42, 43] and thus eligible for comparisons to theoretical predictions. This is important in order to evaluate its performance in its main task: the prediction of germline fitness effects of mutations.

- The evolutionary model automatically provides a list of necessary assumptions for applying it. Its limits are thus well defined.

- It is important to clearly state the null model that any data is compared to. In testing a null hypothesis, it is crucial to know what could have potentially been observed. This demand makes a large majority of cancer mutation data in repositories such as COSMIC [92] not presently applicable for this kind of analysis, for they lack the explicit statement of the opportunity space. (This will likely change in the close future, as the COSMIC providers will include more extensive information about the origin of the uploaded data.)

- The most sensible null model is random w.r.t. the score. Both germline and somatic mutation data can separately be tested against this hypothesis. In particular, they should not be tested against each other.

With the efforts of the International Cancer Genome Consortium [5], much larger data sets will be available for statistical analysis of this kind. In this project, comprehensive data from hundreds of samples for 50 cancer tumor types are collected

and provided to the public. The analysis of these data with the germline fitness score is an immediate task, once they are available. A new era of data driven cancer research will be marked by the advent of cancer polymorphism frequency data (i.e. the frequencies of variants within a single tumor). This will open up new avenues for population genetics motivated analyses and will allow to put new evolutionary models of cancer evolution to the test.

Chapter 4

Cancer mutations and Hidden Markov Models

"Causa latet, vis est notissima."[1]

Ovid, *Metamorphoses*

4.1 Introduction

How can one find mutations that are "driving" the cancer progression in a large pool of random "passenger" variation? The germline fitness score and similar conservation-based methods [26, 99] try to answer this question by estimating the potential damaging effect of a mutation. It was mentioned before that a complementary ansatz is to look for signals of *cancer-beneficial* selection in mutation data [10, 27]. Whatever the true evolutionary model for cancer is, it is certain that mutations beneficial for the cancer progression will fix with a higher rate in the tumor cell population than cancer-neutral mutations.

Note: It is important not to confuse the two different fitness perspectives of a mutation: a cancer-beneficial mutation can be germline-deleterious, as we have seen earlier. But the same can be true for a cancer-neutral mutation (most random passenger mutations reduce fitness, see last chapter). And cancer-deleterious mutations might have no meaning for germline fitness at all.

When sampling an ensemble of cancer tumor cell populations from different patients, the loci under positive selection in cancer will show an increased rate of

[1] "The cause is hidden. The effect is visible to all."

missense mutations. The ratio dN/dS of non-synonymous to synonymous (neutral) substitution rates is the quantity that is usually used to measure this signal [100, 101, 78]. This method of looking for regions and genes with enhanced missense substitution rates has been applied successfully to cancer mutation data [27, 10]. In the context of protein kinase genes, it was shown in [29] (and in [102] for congenital disease SNPs) that driver mutations cluster in functionally relevant sub-domains of the kinase domain. It is exactly this identification of "cancer driver mutation hotspots" [98] that is of immediate importance to understand cancer cell biology.

This chapter will present a new computational method to find signals of selection in mutation data. The framework uses Hidden Markov Models [11]. These models are powerful tools to find the optimal set of model parameters for censored or incomplete data. It is no surprise that the algorithms and inference methods used by Hidden Markov Models are "Bayesian at heart" and fit well in the general framework of this thesis. The following sections will introduce Hidden Markov Models, their basic tasks and how they solve them. Then concrete model adapted to mutations under selection will be formulated. This program will then be used to analyze one more time the protein kinase cancer data from Greenman et al. [10] considered in the last section. The analysis will have the status of a proof of principle. The findings of [10] are reproduced, which consist of a list of genes sorted by their probability to be under positive selection in cancer. Moreover, the mean size of this selection pressure is found. In the conclusive remarks, future potential uses of this method will be addressed.

4.2 HMM: General formulas and algorithms

Hidden Markov Models (HMM) are a computational method for so called "unsupervised statistical learning", i.e. finding automatically the optimal set of model parameters for given data [103, 104, 105, 106]. HMM are routinely used in speech recognition [11] and in sequence discovery [47, 9]. The main idea of HMM is to consider a temporal sequence of observations, e.g. the audio signal recorded by a microphone, as the output from an underlying but inaccessible "hidden" sequence of system states, e.g. the content of the recorded speech. Both the sequence of hidden states and the emission of signals are assumed to follow a probabilistic Markov model with unknown intrinsic parameters. The two main problems that are solved by the HMM method are [11]:

1. What are the parameters of the model?

2. Which sequence of hidden states is the most probable given the signal?

In the context of cancer mutations, the genome of cancer cells is regarded as consisting of regions under positive selection and regions under no selection in cancer evolution. This cancer-selection status of a given sequence site is *a priori* unknown (hidden). But the observation of mutations - both silent and missense - hint at the true status. Thus, the mutation data can be considered as the"emitted" signal of the true sequence state. The parameter to be learned would be the selection strength (equivalent to dN/dS). The location of hidden "selected" sequence sites (genes, domains etc.) would be a valuable piece of information to find cancer driving forces.

The following short exposition of HMM starts by introducing the quantities, formulas and algorithms that are the core of the statistical learning procedure. Throughout this chapter, the following notation of [11] is used:

- $\mathcal{S} = \{S_1, S_2 \dots S_N\}$ is the set of available (hidden) states of the HMM
- $\mathcal{V} = \{v_1, v_2 \dots v_M\}$ is the set of observable symbols. Crucially, there is no $1:1$ relation between hidden states and emitted symbols, otherwise there would be nothing to learn.
- $Q = q_1 q_2 \dots q_T \in \mathcal{S}^T$ is a certain state path with T (time) steps.
- $O = O_1 O_2 \dots O_T \in \mathcal{V}^T$ is a certain observed symbol sequence.
- $\theta = (A, B, \pi)$ the *parameter* set, consisting of:
  - transition probabilities $A = \{a_{ij}\}$, i.e. the probability that the system makes a transition from state $S_i \to S_j$ in a (time) step. ($\sum_j A_{ij} = 1$)
  - emission probabilities $B = \{b_k(l)\}$, i.e. the probability that the (hidden) system being in state S_k emits the symbol v_l.
  - and an entry distribution $\pi = \{\pi_i\}$, i.e. the probability to enter a state sequence into the state S_i.

The three fundamental problems solved by HMM are [11]:

1. What is the probability of a given observation, i.e given (O, θ), what is the path probability $P(O|\theta)$? This is solved by the forward-backward algorithm.
2. Given a concrete observation and a set of parameters θ, what is the most probable hidden state path Q? This path is found by the Viterbi algorithm. The Viterbi algorithm is not directly used in this project, but of principal importance. For completeness, it is described in appendix B.

3. Given an observation O only, what is the parameter set θ that maximizes the observation probability $P(O \mid \theta)$? This is arguably the most difficult task and it is solved by the Baum-Welch algorithm.

Much of the detail of the HMM method to solve these fundamental problems can be found in Rabiner's introductory article [11]. A detailed exposition of the HMM method would go beyond the scope of this work. After the following pedagogical example, the elements of the HMM procedure are concisely presented.

4.2.1 Example: hidden coin tosses

The most simple example of the prototypical situation where a HMM is employed is the following. Consider the situation where you are presented the results of successive coin tosses. The coin itself and the person tossing it are not directly visible, e.g. they are behind a screen. Only the results - head (H) or tail (T) - are offered, e.g. on a display. Now consider that you are told that the person behind the screen actually has two coins, one of which (C1) is "fair" - with equal probability of head and tail - and the other coin (C2) is unfair - with a bias in the two outcomes. Thus we have $\mathcal{S} = \{C1, C2\}$. Moreover, you are told that in between two tosses, the hidden person secretly tosses a *third* coin (C3) - biased or not - to decide whether to switch coins for the next turn, e.g. H means "keep current coin" and T means "switch to other coin".
To the observer, only the sequence of length T (not to confuse with the "tail" symbol) of results of the coin tosses is visible. In terms of the above notation, this is the observation sequence $O \in \{\mathrm{H}, \mathrm{T}\}^T$. The task for the observer is now to determine (i) the biases of the one unfair coin and the secret "switch" coin and (ii) the most likely state sequence $Q = q_1 q_2 \ldots q_T$ of coins that were actually used to produce the observation sequence O. A typical sequence of events might look like this:

time	t:	1		2		3		T-1		T
state	Q:	C1	$\xrightarrow{\text{C3=H}}$	C1	$\xrightarrow{\text{C3=T}}$	C2	…	C1	$\xrightarrow{\text{C3=T}}$	C2
		↓		↓		↓		↓		↓
observation	O:	H		H		T		H		T

A perfect reconstruction of the state sequence Q is impossible. The HMM is after all a probabilistic method. It will find the *most likely* coin biases and state sequence *given the data O*. It is clear that the quality of this estimate is the better the longer the observation sequence O is. We will refer to the picture elaborated here again and again in the subsequent application of the HMM method to the cancer mutation problem. Everything that is essential is already present in this minimal set-up.

4.2.2 The forward-backward algorithm

If the actual state sequence Q would be available, the estimation of the model parameters - emission and transition probabilities, e.g. the biases of the coins C2 and C3 in the coin toss set-up - would be simple enough. For example for the probability of head for C2, one could take all instances of C2 in Q and calculate the proportion of heads in the corresponding observations. We know already, that this estimation procedure should be really carried out with the Bayesian techniques introduced in chapter two, but this is not the focus of this chapter.
The problem is of course that the state sequence Q is *not* available. What one can calculate however, is a probability $P(q_t = S_i | O, \theta)$ for a certain state S_i (C1 or C2) at a certain time t *given* the total observation O and a current best estimate for the parameters θ. As we will see shortly, the parameter estimation procedure is carried out iteratively, so let us assume here that θ is an (educated) initial guess. The forward-backward algorithm not only returns the above state sojourn probability, but also the probability $P(O|\theta)$ of the total observation given the current θ. Crucially, this is the quantity that needs to be maximized to find the most likely θ and Q. This total observation probability is calculated iteratively using the "forward variable".

$$\alpha_t(i) := P(O_1 O_2 \dots O_t, q_t = S_i \mid \theta)$$

This is the probability that the hidden state at time t is S_i in the face of the partial observation $O_1 O_2 \dots O_t$ and conditioned on the parameter θ. The forward variable follows the simple recursion:

1. $\alpha_1(i) = \pi_i b_i(O_1), \quad \forall i = 1, \dots N$
2. $\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij} \right] b_j(O_{t+1}), \quad \forall j = 1, \dots N, \ \forall t = 2, \dots T-1$
3. $P(O \mid \theta) = \sum_{i=1}^{N} \alpha_T(i)$

The value of the forward variable at the end-point T is exactly the total observation probability. Analogously, one can introduce a "backward variable"

$$\beta_t(i) := P(O_{t+1} O_{t+2} \dots O_T \mid q_t = S_i, \theta) \tag{4.1}$$

with corresponding recursion

1. $\beta_T(i) = 1, \quad \forall i = 1, \dots N$
2. $\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j), \quad \forall i = 1, \dots N, \ \forall t = T-1, \dots 1$

Both forward and backward variables will be needed in the solution to the third problem in the above list - estimating the model parameters θ. From the definitions, it follows directly that

$$P(q_t = S_i | O, \theta) \propto \alpha_t(i)\,\beta_t(i). \tag{4.2}$$

4.2.3 The Baum-Welch algorithm

The most difficult problem is the parameter re-estimation: what is the most probable model parameter θ given the observation sequence? As was argued earlier, if one would know the exact hidden state sequence Q, one could easily estimate transition and emission parameters, e.g. by the Bayesian techniques described earlier. The state sequence Q is after all nothing but a finite sample of the underlying probability model. The problem is that one cannot observe the hidden states themselves directly, but only a *probability* for each state sojourn and each state transition at each time can be given.

$$\gamma_t(i) := P(q_t = S_i \,|\, O, \theta) \qquad = \mathcal{N}_t^{-1}\,\alpha_t(i)\,\beta_t(i) \tag{4.3}$$

$$\xi_t(i,j) := P\big(q_t = S_i, q_{t+1} = S_j \,|\, O, \theta\big) = \mathcal{L}_t^{-1}\,\alpha_t(i)\,a_{ij}\,b_j(O_{t+1})\,\beta_{t+1}(j) \tag{4.4}$$

where the normalization factors $\mathcal{N}, \mathcal{L}$ are chosen, such that $\sum_{i=1}^{N}\gamma_t(i) = 1$ and $\sum_{i,j=1}^{N}\xi_t(i,j) = 1$. With these probabilities of state sojourn and state transition, the parameters can be estimated as:

Baum-Welch Update Formulas:

$$\bar{\pi}_i = \gamma_1(i) \tag{BW 1}$$

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)} \tag{BW 2}$$

$$\bar{b}_i(k) = \frac{\sum_{t=1}^{T}\gamma_t(i)\,\delta(O_t - v_k)}{\sum_{t=1}^{T}\gamma_t(i)} \tag{BW 3}$$

It is important to realize, that the Baum-Welch equations are re-estimations, i.e. *updates* of the parameters. One needs to start with a plausible initial value. Furthermore, the above estimations could be endowed with Bayesian pseudo-counts. Often - and also in the case of this work - the observation length T is so large that pseudo-counts barely play a role. We neglect them at this point and stay

with the standard formulation of HMM. Altogether, the usual procedure of HMM parameter estimation is to start with an initial guess and then:

1. calculate the α's and β's in a forward-backward run
2. calculate the path probability $P(O|\theta)$ and test for convergence
3. update the parameters and go to step 1.

After this iteration has converged, one can carry out the Viterbi algorithm to find the most likely sequence Q^* of hidden states.

4.2.4 Example: Poisson emission process

It is instructive to derive the Baum-Welch update formulas for an important but non-standard example of HMM with Poisson emission probabilities:

$$b_k(l) = \mathrm{Pois}(l, \mu_k) := \frac{\mu_k^l e^{-\mu_k}}{l!}, \quad l \in \mathbb{N}_0 \tag{4.5}$$

In this example the observed symbols are positive integers. Later, this quantity will be the number of observed mutations at a certain genetic locus in a collection of samples. The Baum-Welch update rules are a special case of the more general EM-algorithm (expectation-maximization) [107]. This powerful algorithm can be stated in two steps (with the notation of observation sequence O, state sequence S and parameter set θ):

E-step: Calculate the function $Q\left(\theta \mid \theta^{(i)}\right) := \sum_{\text{all } Q} P\left(Q \mid O, \theta^{(i)}\right) \ln P(Q, O \mid \theta)$

M-step: Find the new parameter values as $\theta^{(i+1)} = \underset{\theta}{\operatorname{argmax}}\, Q\left(\theta \mid \theta^{(i)}\right)$

Note: For historical reasons, both the EM-function above and the hidden state sequence are denoted with Q. It should be clear from the context, which one of the two is meant.

Under this procedure, the likelihood of the model $P(O \mid \theta)$ is guaranteed to increase [107]. Actually, it suffices to find a new θ with $Q\left(\theta \mid \theta^{(i)}\right) > Q\left(\theta^{(i)} \mid \theta^{(i)}\right)$ [11]. This is then called a generalized EM (GEM) algorithm. In our case, the first step is always straightforward. The second step, finding the maximum w.r.t. to θ will in general lead to a coupled set of equations that is not always possible to solve exactly. If the emission probabilities were not Poisson-like but simple discrete probabilities over a finite set of possible emission symbols, this

EM-procedure would return the standard update rules (BW 1) - (BW 3).

Now consider the Poisson model above. For a *known* state sequence $Q = q_1 q_2 \dots q_T$, one can calculate the path probability.

$$P(Q, O | \{\mu_i\}, A, \pi) = \pi_{q_1} b_{q_1}(O_1) a_{q_1 q_2} b_{q_2}(O_2) \dots a_{q_{T-1} q_T} b_{q_T}(O_T) \tag{4.6}$$

$$= \pi_{q_1} \left(\prod_{t=1}^{T-1} a_{q_t q_{t+1}} \right) \left(\prod_{t=1}^{T} b_{q_t}(O_t) \right) = \pi_{q_1} \left(\prod_{t=1}^{T-1} a_{q_t q_{t+1}} \right) \left(\prod_{t=1}^{T} \frac{\mu_{q_t}^{O_t} e^{-\mu_{q_t}}}{O_t!} \right) \tag{4.7}$$

Nothing more is needed to calculate the EM Q-function. Anticipating that only the update formula for the Poisson emission rate is of interest here, all terms that do not include the $\{\mu_i\}$ explicitly are dropped. These terms are irrelevant in the second maximization step.

$$Q(\{\mu_i\} | \{\mu_i'\}) = \sum_{\text{all } Q} P(Q | O, \{\mu_i'\}) \ln P(Q, O | \{\mu_i\}) \tag{4.8}$$

$$= \sum_{\text{all } Q} P(Q | O, \{\mu_i'\}) \sum_t \left[-\mu_{q_t} + O_t \log \mu_{q_t} \right] + \dots \tag{4.9}$$

$$= \sum_i \sum_{\text{all } Q} P(Q | O, \{\mu_i'\}) \sum_t \delta(q_t - S_i) \left[-\mu_i + O_t \log \mu_i \right] \tag{4.10}$$

$$= \sum_i \left[-\mu_i \sum_t \gamma_t(i) + \log \mu_i \sum_t \gamma_t(i) O_t \right] \tag{4.11}$$

In the last step, the state sojourn probabilities $\gamma_t(i)$ from the forward-backward step are inserted. These depend implicitly on the old parameter set $\{\mu_i'\}$. Differentiating Q with respect to the $\{\mu_i\}$ yields the update formula for the Poisson emission rates.

$$\bar{\mu}_i = \frac{\sum_t \gamma_t(i) O_t}{\sum_t \gamma_t(i)} \tag{4.12}$$

4.3 HMM for cancer mutations

After the presentation of the HMM technique in the last section, we will now connect to the problem of cancer selection signal detection. In the last sections, a very important ingredient of HMM was left out: the "model" itself, i.e. the topology of the states, their connectivity and the probabilistic nature of emissions. Similar to the Bayesian inference technique, there must always be an input into the machinery, which then readily produces estimates with respect to that input. The model design depends on the concrete problem at hand. There are few mechanical rules how to set up a HMM for a given data set [108]. One guiding principle surely is:

> The amount of data governs the complexity that a model should have. Simpler models (few parameters) are preferred over complex models (many parameters) as long as they are able to explain the data.

This principle could be called the "Occam's razor of HMM" and can be quantified by Bayesian arguments [109]. This principle of optimal model design is very interesting by itself. However, this work is concerned with the very concrete problem of mutations "emitted" from loci that are either under selection in cancer or not. Thus, the HMM design will be fixed in a reasonable manner and estimates are found accordingly.

Without further delay, the cancer-HMM design will now be formulated:

- Every locus in the genome is in one of two states: cancer-selected or cancer-neutral (analogously to the two coins C1 and C2 in the earlier HMM example). Note that this is a statement about the cancer-evolutionary effects of mutations at that locus. Here, the locus is considered to be a nucleotide residue somewhere in the reference sequence for the analyzed (sequenced) part of the genome (e.g. the kinome). The position of a locus is denoted by t (it is in this sense equivalent to the time coordinate of standard HMM).

- Mutations appear at each locus with the same rates $\{\mu_j\}$ in both states. The mutation rates vary for the six different base-pair mutation channels. One could use a single mutation rate μ, but the strong biases are known (see section 3.2.2 and [10]) and the model is built to account for that.

- The rate with which mutations fix in the tumor and are thus subsequently observed in a sequencing experiment depends on their cancer-fitness effect σ. For cancer-neutral loci, there is no modification and the substitution rate is equal to the bare mutation rate. For cancer-selected loci, the substitution rate is modified by a factor: $\mu_j \to \mu_j \frac{\sigma}{1-e^{-\sigma}}$. This choice is the prediction eq. (2.22) from the minimal evolutionary model presented in the first chapter.

- Silent mutations have no effect on protein function, so their observation rate is always the neutral one μ_j, depending on mutation channel.

- At each locus, all mutations are projected over all samples (patients). By doing this the per-sample resolution is lost. But the signal of larger missense mutation rates in cancer-relevant regions is retained.

- Not all mutations are possible at each locus. For each base-pair, only three mutations are possible. The set of totally six mutation channels (of which

always only three are available) is denoted by

$$\begin{aligned}\mathcal{C} := \{&A:T>T:A,\ A:T>C:G,\ A:T>G:C,\\ &C:G>G:C,\ C:G>A:T,\ C:G>T:A\}\end{aligned} \tag{4.13}$$

The availability of mutation channels depends on the nucleotide sequence. For this purpose the filter quantities $\omega_t^{n(s),j}$ are defined as

$$\omega_t^{s(n),j} := \begin{cases} 1, & \text{channel } j \text{ is open at } t,\ \text{mutation is } (\text{non}) - \text{synonymous} \\ 0, & \text{else} \end{cases}$$

And the projected quantities are defined as:

$$\sum_{j\in\mathcal{C}} \omega_t^{s(n),j} =: \omega_t^{s(n)}, \quad \sum_{j\in\mathcal{C}} \sum_{k=n,s} \omega_t^{k,j} =: \sum_{j\in\mathcal{C}} \omega_t^{j} =: \omega_t = \omega_t^n + \omega_t^s \tag{4.14}$$

At the moment, we have $\omega_t = 3$, i.e. only three mutations are possible at each base pair. The following table shows the ω_t^j for the leucine (L) codon. The amino acid in the brackets is the result of a mutation in that channel. Only in the last base are the mutations synonymous.

		A:T>T:A	A:T>C:G	A:T>G:C	C:G>G:C	C:G>A:T	C:G>T:A
L	C	0	0	0	1 (V)	1 (I)	1 (F)
	T	1 (H)	1 (R)	1 (P)	0	0	0
	T	1 (L)	1 (L)	1 (L)	0	0	0

- Likewise, the observations O_t, i.e. the mutation counts, are split into channels and outcomes: $O_t \to O_t^{s(n),j} \in \mathbb{N}_0$.
- The state-dependent emission probabilities are the probabilities of observing O_t mutations after projecting over all samples. They are modeled as Poisson distributions, where the Poisson rates are the state-dependent substitution rates modified by the channel availability.

Table 4.1 summarizes the state-dependent emission properties of the cancer-HMM.

4.3.1 Baum-Welch update formulas for the cancer-HMM

In the general case, one has not one but $N_s \geq 1$ i.i.d. observations from the same HMM. In our context, these could be the individual mutation data per patient or

	Substitution Rate		HMM Emission Probability	
State	silent	missense	silent	missense
Neutral	μ_j	μ_j	$\mathrm{Pois}\left(O_t^{s,j}, \omega_t^{s,j}\mu_j\right)$	$\mathrm{Pois}\left(O_t^{n,j}, \omega_t^{n,j}\mu_j\right)$
Selected	μ_j	$\frac{\mu_j\sigma}{1-e^{-\sigma}}$	$\mathrm{Pois}\left(O_t^{s,j}, \omega_t^{s,j}\mu_j\right)$	$\mathrm{Pois}\left(O_t^{n,j}, \omega_t^{n,j}\frac{\mu_j\sigma}{1-e^{-\sigma}}\right)$

Table 4.1: Emission properties of the cancer HMM

per individual genes, domains or other partitions of the genome. To carry out the HMM-procedure and find the optimal parameters for a given set of observations $\{O_k\}_{k=1\ldots N_s}$, we need to specify the Baum-Welch update formulas. The general formulas for transition rates and entry probabilities are modified by averaging over all samples.

Update of transition rates $\{a_{ij}\}$ and entry probabilities $\{\pi_i\}$ for a set of samples $\{O_k\}_{k=1\ldots N_s}$:

$$\bar{a}_{ij} = \frac{\sum_{k=1}^{N_s}\sum_{t=1}^{T_k}\xi_{k,t}(i,j)}{\sum_{l=1}^{N}\sum_{k=1}^{N_s}\sum_{t=1}^{T_k}\xi_{k,t}(i,l)}, \quad \bar{\pi}_i = \frac{\sum_{k=1}^{N_s}\gamma_{k,1}(i)}{\sum_{j=1}^{N}\sum_{k=1}^{N_s}\gamma_{k,1}(j)} \tag{4.15}$$

Baum-Welch update of mutation and selection strengths

For the mutation rates and the selection strength, there is no closed solution for the maximum of the Q-function. In the implementation used in this work, the μ- and σ-dependent part of the Q-function is maximized numerically. This partial function is given by:

$$\boxed{Q(\mu,\sigma\,|\,\mu',\sigma') := \sum_{k=1}^{N_s}\sum_{j\in\mathcal{C}}\sum_{i=1,2}\left[\langle O_k^j\rangle_i \ln\mu_j + \langle O_k^{n,j}\rangle_i \ln\phi_i - \mu_j\left(\langle\omega_k^{s,j}\rangle_i + \langle\omega_k^{n,j}\rangle_i\phi_i\right)\right]} \tag{4.16}$$

$$\langle O_k^{s(n),j}\rangle_i := \sum_{t=1}^{T_k}\gamma_{k,t}(i)\,O_{k,t}^{s(n),j}, \quad \langle\omega_k^{s(n),j}\rangle_i := \sum_{t=1}^{T_k}\gamma_{k,t}(i)\,\omega_{k,t}^{s(n),j} \tag{4.17}$$

$$O_{k,t}^{j} := O_{k,t}^{n,j} + O_{k,t}^{s,j}, \quad \phi_1 = 1, \quad \phi_2 = \frac{\sigma}{1-e^{-\sigma}} \tag{4.18}$$

This is the function that needs to be maximized with respect to the $\{\mu_j\}_{j\in\mathcal{C}}$ and σ.

4.4 Cancer genes in the human kinome

The cancer mutation data studied in Greenman et al. [10] and considered in chapter 3 with respect to the germline fitness cost of mutations will now be subjected to the cancer-HMM. The method should be able to find regions in the kinome under positive selection in cancer. This study shall serve as a proof-of-principle in that it tries to reproduce the findings of [10]: a list of kinase genes ordered by their probability to be selected for in cancer. The following assumptions and simplifications are made:

- The cancer status (neutral or selected) does not change within a gene. There are no transitions between the two states within a gene. For the HMM, this means $a_{ij} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}_{ij} = \text{const}$. These transition rates are not updated.
- It follows, that the state sojourn probabilities do not change within a gene: $\gamma_{k,t}(i) = \gamma_k(i) = \text{Prob}(\text{gene } k \text{ is in state } i)$. This quantity will be used to order the genes in the above sense.
- The mutation data for each gene is treated as a sample observation from the same HMM. This means that mutation and selection parameters are not found for each gene separately, but rather a *mean* selection strength σ (averaged over all selected genes) and mean mutation rates.

For this particular purpose, the elaborate HMM procedure might seem like an "overkill". Especially because the HMM is not used to its full capacity, namely spatially resolving the hidden state sequence within genes. This task however is not easily solvable with the limited size of the available data set. The above set-up is but a first order analysis and larger data sets will allow higher resolution studies. Nevertheless, a short outlook at further applications is given in a later section.

4.4.1 Materials and methods

As mentioned before, the same data set as considered earlier is analyzed (see chapter 3). However, silent and missense mutations in the whole gene (and not only within the subset of protein domains) are included.

Input: The mutational opportunity space - the set of all possible point mutations away from the reference kinome sequence - is translated into availability tracks $\left\{\omega_{k,t}^{s(n),j}\right\}$ for each gene. The mutation data itself is translated into a separate observation track $\left\{O_{k,t}^{s(n),j}\right\}$. This set of tracks is the input to the HMM program.

Program: The HMM algorithms - forward-backward and Baum-Welch - are implemented as C++ programs. The GNU scientific library (GSL) version 1.15 [110] is extensively used in this implementation, especially for the maximization of the Q-function. After the Baum-Welch iteration converges to a local maximum of the observation likelihood, the robustness of the parameter estimates is investigated by MCMC sampling (Markov-Chain-Monte-Carlo sampling) [111, 112, 113]. In short, this is a stochastic method to approximate the posterior probability distribution for the parameter estimates. It can be thought of as a random walk in parameter space, where transitions take place according to the "energy function" - here the Baum-Welch Q-function. A transition to a state of lower "energy" (with difference $\Delta Q < 0$) is in principle allowed but suppressed by a Boltzmann-factor $\exp(\Delta Q) < 1$, whereas transitions with $\Delta Q \geq 0$ are always accepted. Thus the random walk will spend most of the time in regions of high probability and rarely sample the tails of the equilibrium distribution.

Output: The final output of the HMM program consists of (i) numerical estimates for the mutation and selection parameters (ii) samples of the posterior probability distribution of the parameters (iii) the list of kinase genes with their probability to be in one of the two states - cancer-neutral or cancer-selected.

4.4.2 Results

For completeness, both the germline and the somatic protein kinase cancer mutation data from [10] were separately subjected to a HMM analysis.

Germline mutations: As expected, the germline mutations show a strong signal for *negative* selection, i.e. a lower than random rate of missense mutations. The mean selection pressure over all 518 genes is $\sigma_g = -1.89 \pm 0.05$ (subscript g denoting the germline, see figure 4.2). A large proportion of the kinase genes (86%) have a higher probability to be under purifying selection in the germline than to be neutral (see figure 4.1).

Somatic mutations: The outcome is quite different for the set of somatic mutations in the kinome. Only 15% of the genes are more likely to be under selection in cancer than to be cancer-neutral (see figure 4.1). The mean selection pressure for selected genes is positive: $\sigma_s = 1.77 \pm 0.31$. Of special importance is the list of kinase genes sorted by their probability to be cancer-selected. In table 4.2, the top ten of that list is compared to the corresponding list in [10] (table 3 therein). The agreement is convincing.

Rank	Gene name	Prob($\sigma_s > 0$)	Gene (as in [10])
1.	TTN	1.00	TTN
2.	BRAF	0.98	BRAF
3.	ATM	0.96	ATM
4.	TAF1L	0.94	TAF1L
5.	ERN1	0.91	ERN1
6.	FGFR2	0.88	MAP2K4
7.	NTRK3	0.87	CHUK
8.	EPHA6	0.84	FGFR2
9.	MAP2K4	0.84	NTRK3
10.	MGC42105	0.83	MGC42105

Table 4.2: Comparison of the 10 genes most likely to be under positive selection in cancer as a result of the HMM method. The analogous table in [10] (table 3 therein) was derived without the use of HMM. This result is mainly a proof of principle for the HMM method.

The bare mutation rates $\{\mu_j\}_{j\in\mathcal{C}}$ found by the HMM program for each set are not very informative by themselves (after several projections). The bias in those HMM estimates can be compared to the naive channel bias found in the data sets (without taking the opportunity/availability of mutations into account) depicted in figure 3.1. The outcomes are nearly identical (see figure 4.3).

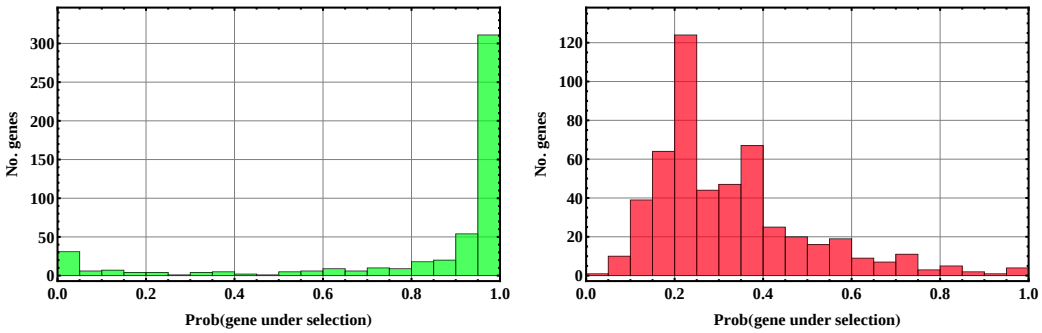

Figure 4.1: Histogram of the 518 protein kinase genes with respect to their probability to be under selection according to the Hidden Markov Model as explained in the text. For germline mutations, most genes (86%) show a higher probability to be under purifying (negative) selection (left). On the other hand, for somatic cancer mutations most genes show now signal of selection (right). The proportion of genes with $P > 0.5$ is 15%.

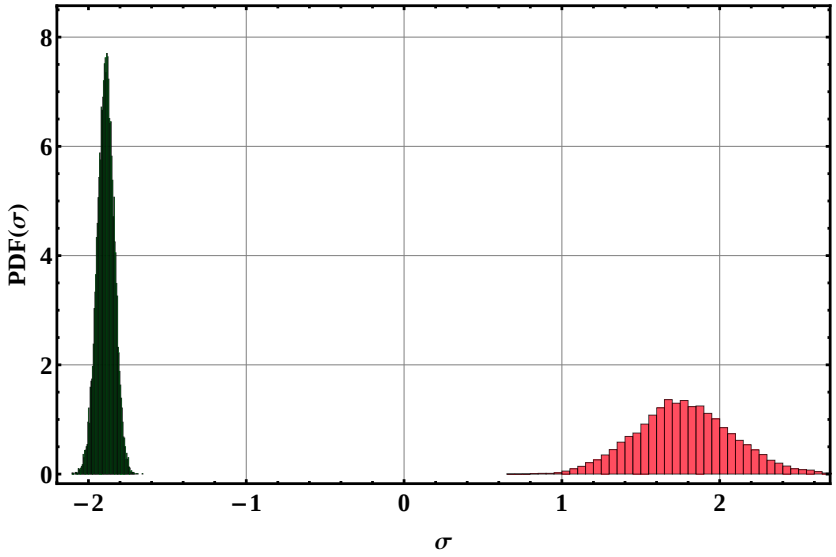

Figure 4.2: Approximation of the probability distribution of the mean selection pressures σ discovered by the HMM method in the protein kinases cancer data. Germline mutations (green) under selection are most compatible with $\sigma = -1.89 \pm 0.05$, somatic mutations under selection (red) with $\sigma = 1.77 \pm 0.25$. The distributions are approximated by Markov-Chain Monte-Carlo simulations (10^5 steps, of which every tenth was used).

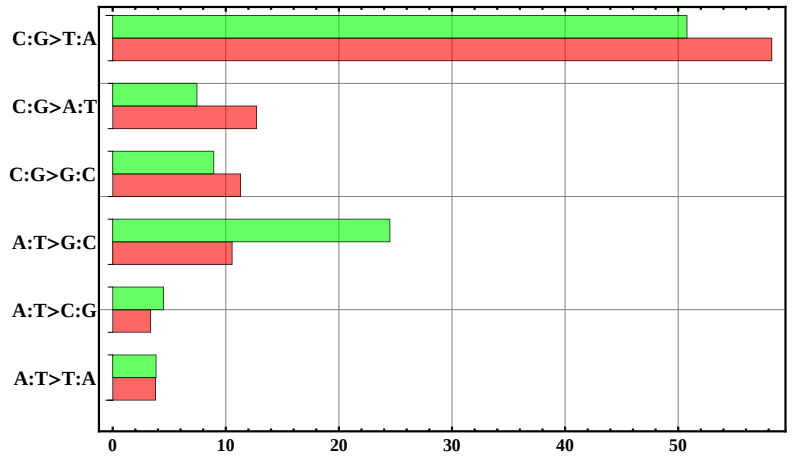

Figure 4.3: Mutation channel bias found by the HMM in both germline (green) and somatic (red) data sets. Actual mutation rates are normalized to percent.

4.5 Domain-level analysis

We are finally in a position to reap some rewards from the general and flexible set-up of the cancer-HMM. The analysis in the last section considered genes as units with defined cancer-selection status. The same can be repeated for domain families. However, for the considered kinome data set, this analysis is necessarily biased, for all genes in the kinome include by definition one of the two protein kinase domains. Nevertheless, it is worth investigating. Operationally, the only modification must be made in the input of the HMM: the position coordinate t now corresponds to an alignment column within a domain family profile. The observation and availability values (the O_t, ω_t) are projected over all kinome sites that fall onto that specific column. The rest of the program is completely identical to the gene-level analysis. The results can be seen in table 4.3 and figures 4.4 and 4.5. Reassuringly, the selection strengths found in the germline and somatic mutation set are similar to the gene-level analysis results ($\sigma_g = -1.87 \pm 0.07$ and $\sigma_s = 0.64 \pm 0.25$). In both sets, almost all domains are more consistent with selection (94% in the germline, 100% in the somatic set). As expected, the top ten domains under selection in cancer are lead by the two protein kinase domains. It is biologically rather interesting that most of the highly selected domains are involved in cell signaling processes. However, more biologically informative would be an analysis of cancer mutation sequencing data that is not preconditioned on a specific family of genes. Thus, not too much weight will be put on an interpretation of the findings, but rather on the potential of this type of analysis.

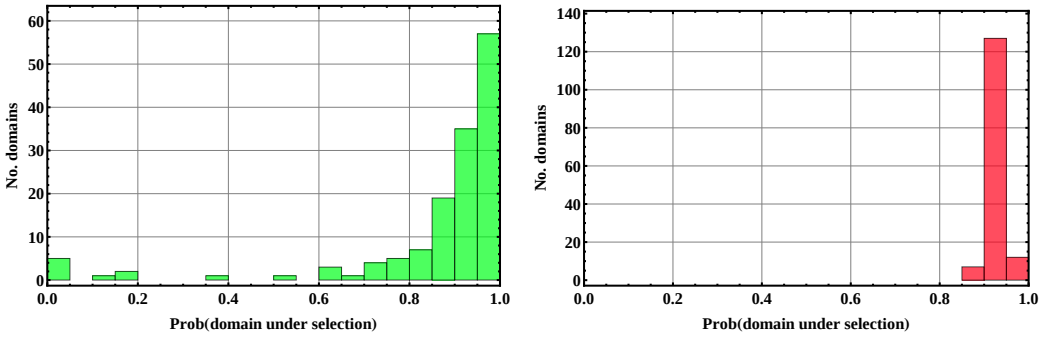

Figure 4.4: Histogram of the 146 protein domains found in the kinome with respect to their probability to be under selection in cancer. For both germline and cancer somatic mutations, most domains (94% and 100%, resp.) show a higher probability to be under selection.

Rank	Pfam ID	Domain name	Function	Prob($\sigma_s > 0$)
1.	PF07714	Tyrosine kinase	phosphorylation	1.00
2.	PF00069	Protein kinase	phosphorylation	1.00
3.	PF07697	7TM-HD receptor	N/A	0.99
4.	PF00041	Fibronectin type III	cell adhesion/growth	0.99
5.	PF12157	DUF3591	N/A	0.99
5.	PF01404	Ephrin receptor	cell signaling	0.98
6.	PF02259	focal adhesion targeting	cell signaling	0.98
7.	PF00439	Bromodomain	protein binding	0.97
8.	PF00169	Pleckstrin homology	intracellular signaling	0.97
9.	PF00629	MAM domain	extracellular receptor	0.96
10.	PF00241	Cofilin-ADF	actin binding	0.95

Table 4.3: Comparison of the 10 domains most likely to be under positive selection in cancer as a result of the HMM method. The domain name and function is taken from Pfam [9]. (DUF = domain of unknown function)

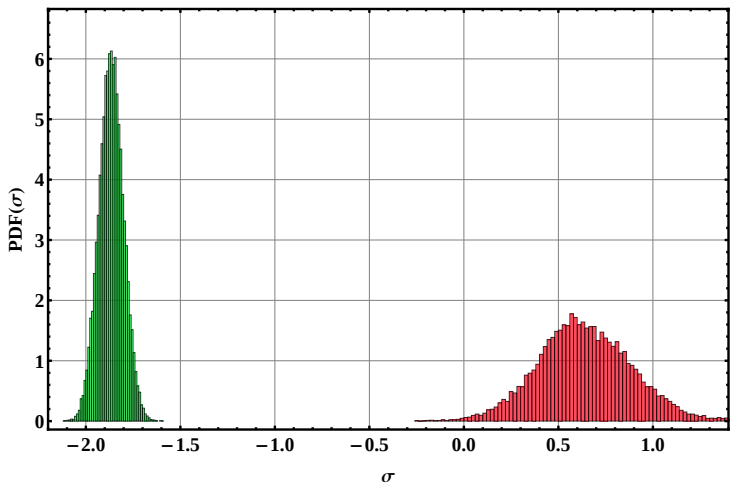

Figure 4.5: Same figure as 4.2 but for domains. Germline mutations (green) under selection are most compatible with $\sigma_g = -1.87 \pm 0.07$, somatic mutations under selection (red) with $\sigma_s = 0.64 \pm 0.25$.

4.5.1 Outlook: sub-domains under selection?

The gene- and domain-level HMM analysis was carried out with the restriction that the cancer-selection status is not allowed to change within a unit. If this restriction is to be lifted, the HMM needs to find optimal values for the transition

rates between the states as well. This increases the number of parameters of the HMM by two. For all genes and almost all domains in the data set, this is too much flexibility for the HMM to produce meaningful results. Even for the tyrosine kinase domain (PF07714), which has both the most domain instances found in the kinome and the most somatic mutations, the HMM results are at least questionable (e.g. they strongly depend on the initial guesses for the length scales). So this should be taken as a demonstration of the capability of the HMM ansatz.
The protein kinase domain consists of twelve sub-domains [60] and it is interesting to compare this annotation with the output of the HMM analysis. In [102], Torkamani *et al.* found that these sub-domains were enriched with disease-associated SNPs. The figure 4.6 shows the result of a HMM-analysis of the tyrosine kinase domain alone. The solid line shows the probability that a nucleotide at a certain domain position is in the cancer-selected state, i.e. $\gamma_i(\text{selected})$ (with $\sigma_s \approx 4.4$). The colored regions are the twelve sub-domains (sub-domains III and IV are taken as one and have the same color, sub-domains XI and XII are adjacent). Only sub-domains VII and VIII show a high probability for cancer-selection. Altogether, the correlation is not strong (The probability to be cancer-selected is on average about 29% higher within sub-domains than outside).

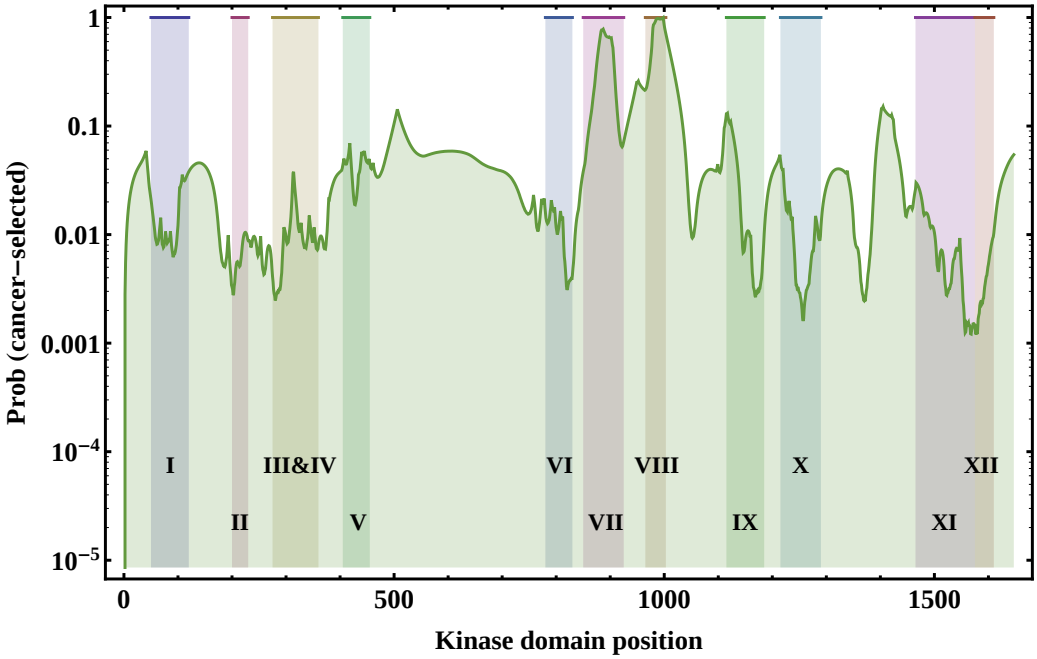

Figure 4.6: The cancer-selection probability for each nucleotide in the tyrosine kinase domain as given by an HMM-analysis of the cancer mutation data [10]. In sub-domains VII and VIII, there is a high probability for the presence of cancer-selection in the data.

4.6 Discussion and future directions

This chapter presented the HMM method as an efficient way to find signals of selection in cancer mutation data. The HMM is able to distinguish genes or domains under selection from neutral units. In the present study this was accomplished by computing for each unit the probability to be in a selected state. Ideally, one would like to find selection strengths e.g. for each gene separately. But to do this in a conclusive way, more extensive data sets are required. At this point, only mean selection pressures could be detected with confidence.

It is of immense biological interest to identify locations within genes or domains of immediate importance to cancer-evolution. To achieve this "spatial" resolution demands not only larger data sets, it also needs further development of the "state-transition" part of the HMM. In the HMM-analysis it is implicitly assumed that there is a typical length-scale associated to each cancer-selection state. This might not be the case in reality. In fact, it is also possible that cancer mutation hot-spots can only be identified in the folded form of the protein, where residues are close in real space but far apart in the coding sequence.

Regarding the implementation of the HMM program, one difficulty is that the Baum-Welch (or EM) algorithm returns only local maxima of observation likelihood. Thus, depending on the initial parameter estimates, one might miss the true global maximum and with it the optimal parameters. This deficiency can be addressed by Monte-Carlo methods such as simulated annealing [114] for continuous parameters [115]. This method tries to find the ground state by slowly "freezing" the system, i.e. allowing ever less random transitions that increase energy. This attractive idea comes with huge computational costs, especially for HMM. For every random trial move in parameter space the observation probability must be calculated anew, which amounts to a forward/backward evaluation. Further development of the software implementation needs to be done before simulated annealing is feasible for the present HMM. For other HMM models, it is already in use [116].

The HMM in its present form takes the non-homogeneous state of the reference sequence explicitly into account by means of the mutation-availability tracks ω_t. But there is more external information available that can be fed into the HMM: the germline fitness scoring effect Δs of mutations, described in the last chapters. In domains, where Δs is available, this scale could be used as explicit selection strength, i.e. $\sigma_s = -\Delta s_t$ (and a mean σ_s outside of domains). This set-up assumes that "cancer-beneficial" is equivalent to "germline-deleterious". It is interesting to see whether this model would be able to detect tumor-suppressor genes.

Chapter 5

Stochastic tunneling in a two locus model with recombination

"Your DNA may be destined to mingle with mine. Salutations!"

Richard Dawkins, *River out of Eden*, 1995

5.1 Introduction

In the preceding chapters, the understanding of a minimal model of evolution under the influence of three major evolutionary forces - mutation u, selection s and drift $\frac{1}{N}$ was extensively used. This model described the evolution of a population of size N whose members were either of (geno-)type A or B, the two possible alleles. Two two alleles carry different fitness values s_A and s_B, with difference $s := s_A - s_B$. Not only is a complete qualitative understanding of the behavior of this system and the relevant parameter regimes available, but importantly one can give analytical expressions for all relevant time scales and fixed state probabilities. It is known that for a small mutation rate $Nu \ll 1$, the population will be most of the time monomorphic, i.e. in a state of either all A-alleles or all B-alleles, and it will rarely switch to the other monomorphic state (see e.g. Rouzine's review [38]). In essence, the derivation of the germline fitness score in earlier chapters is based on the substitution rate $\Gamma_{B\to A}(u,\sigma)$, the typical time scale between successive switching events. This rate depends non-linearly on the scaled fitness difference $\sigma := N(s_A - s_B)$.

$$\Gamma_{B\to A}(u,\sigma) = \frac{u\,\sigma}{1-e^{-\sigma}} \tag{5.1}$$

Expressions like this are extremely important in order to estimate model parameters from direct observations. This chapter leaves behind the application of cancer

mutation scoring altogether. The primary goal here is to find a corresponding expression for the substitution rate of a more extended evolutionary model - one that includes two more forces: recombination and epistasis. Recombination is the exchange of genetic material of the two parents to form the genome of the child in sexually reproducing organisms. This exchange takes place in the crossover of homologous chromosomes during meiosis to produce the haploid gametes: sperm and egg cells. Epistasis is a term describing the non-additive fitness effect of mutations at different loci. This happens whenever the alleles at two loci are interdependent: the effect of a mutation at one locus depends on the allele at the other locus. The fact that sexual reproduction is ubiquitous in nature is a long standing topic of research [117]. Most explanations include epistatic interactions between loci.

There are different epistatic scenarios: under the assumption that most mutations decrease fitness, the combined effect of many such deleterious mutations can be lower (larger) than the sum of their effects, which is called antagonistic (synergistic) epistasis. Which one of the two forms of epistasis (antagonistic or synergistic) is more prevalent in nature is still matter of debate. Theoretical studies hint that in fact antagonistic interactions between deleterious mutations - i.e. a "buffering" of the effects - might be favored in evolution [118].

The form of epistasis considered here is that of "sign-epistasis" [12]. This term means that the sign of the fitness effect of a mutation depends on the genetic background. In a minimal set-up, this describes the process of compensatory mutations, where two fitness peaks are separated by a fitness valley. The first mutation away from one such local fitness maximum genotype always reduces fitness. But a mutation at the second locus (over-)compensates this effect. If the fitness of both peaks is equal, one is faced with neutral compensatory mutations [119, 120]. Of interest here is the more general case of an over-compensating effect of the second mutation. This is an interesting situation to analyze the process of adaption (fixation of the globally fittest allele) in the presence of evolutionary bottlenecks. Models of this kind are known to show the effect of "stochastic tunneling" [15], where the population shifts between peaks without populating the states in the fitness valley to a macroscopic extent.

The determination of evolutionary relevant time scales for adaption in these models is a topic of ongoing research. All of the above considerations can be realized in a model of evolution of genomes with two loci and two alleles each, i.e. the evolution of four competing genotypes. The fitness assignments to these genotypes are realized in a way to exhibit two local fitness maxima (e.g. wild type genotype and the genotype two mutations apart from the wild type). This model will be explained in detail in the next section. For infinite population sizes - when random fluctuation can be neglected - there have recently been advances to describe stationary states [14] and deterministic times until fixation of the double mutant

genotype [13]. One important observation is that in infinite populations fixation of the double mutant is impeded for recombination above a critical threshold [13]. This divergence of the fixation time is resolved as soon as finite populations are taken into account. A full stochastic treatment also exposes the effect of stochastic tunneling and other fixation bottlenecks. Recently, Weissman et al. [121] have found expressions for the fixation rate in the case of shallow fitness valleys and for $r < r_c$ (and for $r \gg r_c$). The present work aims to fill an important gap in the theoretical description of compensatory adaption dynamics: the scaling of the fixation rate for values $r = \mathcal{O}(r_c)$ and above for deep fitness valleys. It will be shown how recombinatorial reshuffling of genotypes leads to a phase transition at $r = r_c$ and how the fixation dynamics is influenced by stochastic bottlenecks. The main ideas and first results of this analysis were published in [122].

5.2 Formulation of the model

The minimal model to realize all these aspects and to include the two new forces mentioned above is that of a population of constant size N, where the individuals carry a genome of two loci with two alleles each. This means that there are four genotypes *ab*, *Ab*, *aB* and *AB*. Epistasis is the non-additive effect of mutations depending on the genetic background. For our minimal model, this translates to a non-linear fitness-landscape (see figure (5.1)). In fact, we study here the more special case where the global fitness maximum is separated from the wild type by a deep fitness valley. A first mutation away from the wild type *ab* - at any locus - decreases the fitness of an individual significantly. Only the mutation at the second locus over-compensates this effect and produces an individual of maximal fitness. A natural question to ask is then the following: what is the time scale of ultimate fixation of the double mutant starting from a population monomorphic in the wild type? The answer will be the analog of the substitution rate (5.1). Let us recapitulate the defining characteristics of the model:

- Minimal model of evolution that includes the five evolutionary forces:
  1. mutation u
  2. selection s
  3. genetic drift $\frac{1}{N}$
  4. recombination r
  5. epistasis
- "Genome size": two loci, two alleles each. There are four different genotypes: wild type *ab*, single mutants *aB* and *Ab* and double mutants *AB*.

- The fitness assignment to the four genotypes is such that the first mutation away from the wild type is strongly deleterious and the second mutation over-compensates this effect.
- The primary quantity of interest is the time scale on which a population starting in the wild type reaches the state of maximal fitness and fixes in the double mutant state.

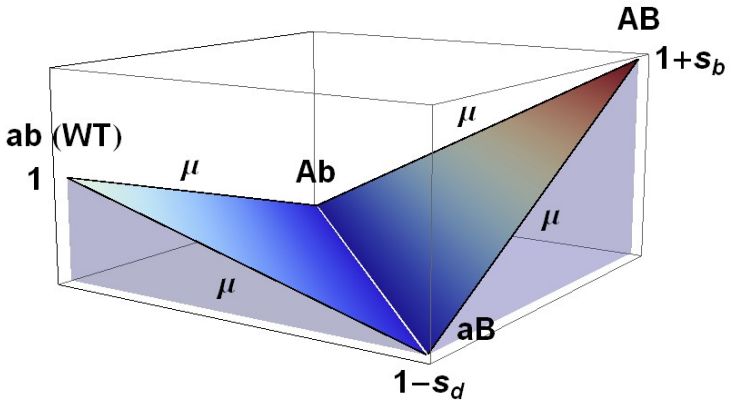

Figure 5.1: The fitness assignment for the four genotypes in the model of compensatory mutations. For a population starting out in the wild type *ab*, fixation in the global fitness maximum at all *AB* is delayed by the crossing of a deep fitness valley.

5.2.1 The twofold effect of recombination

Recombination means the exchange of genetic material between the two parents in sexually reproducing organisms to form the genome of the offspring. Qualitatively, the effect of recombination is that of a source of variation, quite similar to mutations. However, it can produce new variants much more rapidly by reshuffling of the present genotypes. As a thought experiment, imagine a population that is half wild type *ab* and half double mutant *AB* evolving sexually under the influence of recombination. Initially, half of all offspring will have inter-genotype parents and half of these children will be of mixed type. Thus, within a single generation a quarter of the population will be of mixed genotype, whereas mutation alone would produce only $\mathcal{O}(2Nu)$ of them.
In the present context, this effect of recombination as a source of variation comes in two flavors: (i) it opens up an additional channel to produce the favorable double mutants by combining the genetic material of two (different) single mutant

parents; (ii) much more strongly, recombination counteracts the fixation of double mutants by reshuffling the alleles in reproduction events with the wild type. We will rediscover these two aspects later in the mathematical formulation.

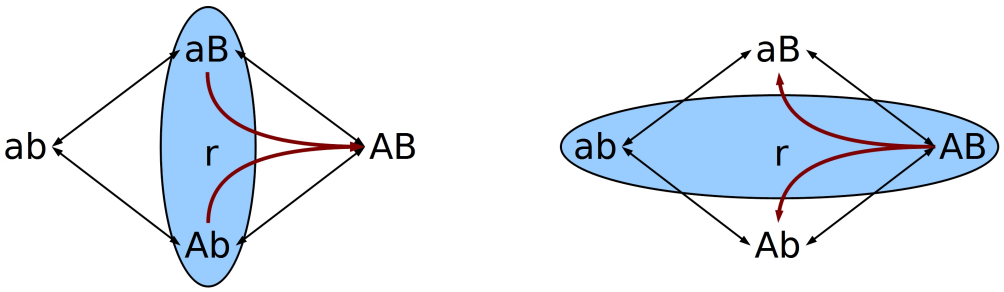

Figure 5.2: The two opposing effects of recombination described in the text. Left: recombination opens a new channel to produce double mutants by breeding two single mutants. Right: recombinatorial reshuffling of alleles makes fixation of the double mutant difficult.

5.2.2 Mathematical model: Moran birth-death process

In the field of population genetics, there are two well-established models to describe evolution under constant population size [39]: (i) the Wright-Fisher model of generation-wise population updates and (ii) the Moran model of individual birth-death processes with overlapping generations. Whereas the former is clearly preferable in simulations, the latter is used for the calculations in this part. Since the Moran model includes only nearest neighbor transitions in the form of birth and death events, it is much more analytically tractable than the long range Wright-Fisher model. In the appendix C, we show that in the biologically relevant parameter regime both models are in fact equivalent and either can be used for simulations (or calculations if need be).

The Moran model is essentially a protocol of evolution in discrete time steps. In every "turn" of the process, the following events take place:

1. One of the N individuals dies. Which one is decided randomly. The effect of selection is incorporated already at this point in the form of biased death probabilities $\{D_i\}$ for the different genotypes according to their fitness.

2. Two individuals are chosen at random (unbiased) to serve as parents, to reproduce and spawn a single offspring that fills the empty spot. Copies of the parents' genomes are the raw material to build the child's genome.

3. With probability r, the two copies exchange their alleles at the second locus. This approximates recombination by chromosomal crossover.

4. Of the two new genomes, one is chosen at random (unbiased) to be the actual offspring.

5. At each locus of the offspring, a mutation to the other allele can occur with probability u. Double mutation events are neglected. The total probability that the offspring is of genotype i, is denoted with B_i.

Just as was done in chapter two for the Kimura model, the above evolution protocol can be cast in the following continuous-time Moran-model Master equation.

$$\partial_t P(\boldsymbol{n},t) = \left[\sum_{i\neq j}\left(\mathbb{E}_i^- \mathbb{E}_j^+ - 1\right) B_i(\boldsymbol{n})\, D_j(\boldsymbol{n})\right] P(\boldsymbol{n},t) \tag{5.2}$$

$$(\boldsymbol{n})_i = n_i, \quad i,j \in \{ab,Ab,aB,AB\}, \quad \mathbb{E}_i^{\pm} f(\boldsymbol{n}) = f(\boldsymbol{n} \pm \boldsymbol{e}_i), \quad (\boldsymbol{e}_i)_j = \delta_{ij} \tag{5.3}$$

where $P(\boldsymbol{n},t)$ is again the probability to find n_i copies of genotype i in the population at time t. This compact form of the Master equation uses the shift operators $\mathbb{E}_i^{\pm}$ that shift the i-th component of the argument of any function $f(\boldsymbol{n})$ by plus or minus one. The effective birth and death rates $\{B_i, D_i\}$ incorporate all the evolutionary forces. The index marks the end result of an event, e.g. B_{ab} is the probability (per unit time) that the born offspring is of the wild type. According to the protocol above, this probability is given by a composition of sexual and asexual reproduction steps.

$$A_i := \text{Prob(offspring of type i in purely asexual reproduction, } r = 0) \tag{5.4}$$

$$S_i := \text{Prob(offspring of type i in purely sexual reproduction, } r = 1) \tag{5.5}$$

Because the sexual reproduction step is only in effect with a probability r, we have in total for the effective birth and death probabilities:

$$B_i(\boldsymbol{x}) := A_i\left((1-r)\boldsymbol{x} + r\boldsymbol{S}(\boldsymbol{x})\right), \quad D_i(\boldsymbol{x}) := \frac{m_i x_i}{\bar{m}}, \quad \text{with} \quad x_i := \frac{n_i}{N} \tag{5.6}$$

$$\text{death biases}: \quad m_{ab} := 1, \quad m_{aB} = m_{Ab} := 1 + s_d, \quad m_{AB} := 1 - s_b \tag{5.7}$$

$$\text{mean fitness}: \quad \bar{m} = 1 + s_d\left(x_{aB} + x_{Ab}\right) - s_b x_{AB} \tag{5.8}$$

The biases including selection are denoted with m_i ("mortalities"). We here list the purely asexual and sexual birth probabilities explicitly.

$$\begin{aligned} A_{ab} &= x_{ab}(1-2u) + (x_{aB}+x_{Ab})\,u & S_{ab} &= x_{ab}^2 + x_{ab}(x_{aB}+x_{Ab}) + x_{aB}x_{Ab} \\ A_{aB} &= x_{aB}(1-2u) + (x_{ab}+x_{AB})\,u & S_{aB} &= x_{aB}^2 + x_{aB}(x_{ab}+x_{AB}) + x_{ab}x_{AB} \\ A_{Ab} &= x_{Ab}(1-2u) + (x_{ab}+x_{AB})\,u & S_{Ab} &= x_{Ab}^2 + x_{Ab}(x_{ab}+x_{AB}) + x_{ab}x_{AB} \\ A_{AB} &= x_{AB}(1-2u) + (x_{aB}+x_{Ab})\,u & S_{AB} &= x_{AB}^2 + x_{AB}(x_{aB}+x_{Ab}) + x_{aB}x_{Ab} \end{aligned}$$

This leads to the following effective birth rate e.g. for the wild type.

$$B_{ab} = [(1-r)\,x_{ab} + r\,S_{ab}(\boldsymbol{x})]\,(1-2u) \;+ \sum_{i=aB,Ab} [(1-r)x_i + r\,S_i(\boldsymbol{x})]\,u \tag{5.9}$$

This should be read as a direct translation of the protocol (asexual or sexual reproduction, weighted by recombination r, followed by eventual mutation). Before proceeding, the following comments are in order.

Note 1: The above set-up explicitly guarantees a constant population size N. It would be equally valid to let birth and death events happen independently. The according Master equation would read as

$$\partial_t P(\boldsymbol{n},t) = \sum_i [(\mathbb{E}_i^- - 1)\,B_i(\boldsymbol{n}) + (\mathbb{E}_i^+ - 1)\,D_i(\boldsymbol{n})]\,P(\boldsymbol{n},t) \tag{5.10}$$

Here, the population size N would be constant on average only. This is ensured by the normalization of all relevant rates.

$$\sum_i A_i = \sum_i S_i = \sum_i B_i = \sum_i D_i = 1 \tag{5.11}$$

Note 2: It would also be equally valid to include the effect of selection in the birth rates and use death rates that are "flat", i.e. $D_i = x_i$. Also the exact order in the update protocol could be changed as well. In the limits regarded here, i.e. all rates small $u, s_d, s_b, r \ll 1$, all of these different implementations are equivalent in the sense that they yield the same limiting evolution equation for $N \to \infty$.

Note 3: The terms S_i describing the outcome of sexual reproduction are purely combinatorial. They can be written more compactly by introducing the coefficient of linkage disequilibrium $\mathrm{LD}(\boldsymbol{x})$

$$\mathrm{LD}(\boldsymbol{x}) := x_{ab}x_{AB} - x_{aB}x_{Ab} \quad \Rightarrow \quad S_i = x_i \pm \mathrm{LD}(\boldsymbol{x}) \tag{5.12}$$

where the plus sign holds for the single mutants and the minus sign for the other two genotypes.

Note 4: Although the model is formulated in discrete evolutionary "turns", it can be cast into the continuous-time Master equation (5.2) by the following observation. Let the turns take place not at evenly spaced discrete time points but rather in a random fashion, where the waiting time between consecutive turns is exponentially distributed. Then the change of the probability distribution is only evaluated at evenly spaced time points on a time scale, where there is one turn per unit time on average. In the limit of infinitesimally close sampling points, this will yield the above Master equation. Similarly, if all processes are decoupled and take place independently, every single one of the possible transitions has a waiting time that is exponentially distributed with its own mean (the B_i or D_i). This set-up will yield the alternative Moran Master equation (5.10). It is also at the heart of the Gillespie algorithm [123] that exploits the exponential distribution of waiting times to speed up Moran type simulations. In appendix C.4, we show this equivalence explicitly.

5.2.3 Parameter regimes

We now return to the main exposition. It is worth noting that it usually suffices to consider the evolutionary force parameters only to linear order within the transition rates. In fact, it will be the scaled parameters Nu, Ns etc. which decide the qualitatively distinct sectors of the model. In the one-locus/two-allele model, $Nu \ll 1$ or $Nu \gg 1$ puts the system in the substitution- or mutation-selection-balance regime, respectively, and the substitution rate (5.1) is a non-linear function of Ns [38]. So the interesting limit in the present case will be:

$$N \to \infty, \quad \text{with} \quad \mu := Nu,\ \sigma_b := Ns_b,\ \sigma_d := Ns_d,\ \rho := Nr \quad \text{constant} \tag{5.13}$$

Note: Of course, in reality we don't expect e.g. individual reproduction rates to scale with population size. Why should they? The above statement is just a prescription for taking a biologically meaningful limit.

In the face of this expansion, it should suffice to keep the transition rates only to leading order in $\frac{1}{N}$. The birth and death rates are then in that sense:

$$\begin{aligned}
B_{ab} &:= x_{ab} - 2u x_{ab} + u\left(x_{aB} + x_{Ab}\right) - r\,\mathrm{LD}(\boldsymbol{x}) \\
B_{aB} &:= x_{aB} + u\left(x_{ab} + x_{AB} - 2x_{aB}\right) + r\,\mathrm{LD}(\boldsymbol{x}) \\
B_{Ab} &:= x_{Ab} + u\left(x_{ab} + x_{AB} - 2x_{Ab}\right) + r\,\mathrm{LD}(\boldsymbol{x}) \\
B_{AB} &:= x_{AB} - 2u x_{AB} + u\left(x_{aB} + x_{Ab}\right) - r\,\mathrm{LD}(\boldsymbol{x})
\end{aligned} \tag{5.14}$$

$$\begin{aligned} D_{ab} &:= x_{ab} - s_d\, x_{ab}\left(x_{aB} + x_{Ab}\right) + s_b\, x_{ab}\, x_{AB} \\ D_{aB} &:= x_{aB} + s_d\, x_{aB}\left(x_{ab} + x_{AB}\right) + s_b\, x_{aB}\, x_{AB} \\ D_{Ab} &:= x_{Ab} + s_d\, x_{Ab}\left(x_{ab} + x_{AB}\right) + s_b\, x_{Ab}\, x_{AB} \\ D_{AB} &:= x_{AB} - s_d\, x_{AB}\left(x_{aB} + x_{Ab}\right) - s_b\, x_{AB}\left(1 - x_{AB}\right) \end{aligned} \tag{5.15}$$

with the linkage disequilibrium $\mathrm{LD}(\boldsymbol{x})$ defined in (5.12). Expressing the rates in terms of the scaled parameters, we have their expansion in $\frac{1}{N}$, which is called *canonical form* in [40].

$$B_i(\boldsymbol{x}) =: x_i + \tfrac{1}{N}\, b_i(\boldsymbol{x}) + \mathcal{O}\left(N^{-2}\right), \quad D_i(\boldsymbol{x}) =: x_i + \tfrac{1}{N}\, d_i(\boldsymbol{x}) + \mathcal{O}\left(N^{-2}\right) \tag{5.16}$$

In the above limit, the remaining four parameters span a huge parameter space. A full cartography of which is not the aim of this thesis. Instead, the focus is on a specific sector where substitution dynamics and fixation bottlenecks of various kinds can be found. Consequently, some of the important parameter combinations and their relative sizes are now fixed:

- $Nu = \mu \ll 1$: This is the relevant scale of mutation for all biological evolution models, except maybe for viral or cancer evolution [38]. In fact, the true mutation rate at a locus depends on its size. Per nucleotide mutation rates are small even for the previous examples.

- $\sigma_b, \sigma_d \gg 1$: The effect of the epistatic fitness landscape can only be detected if selection is sufficiently strong to overcome genetic drift.

- $\sigma_b \ll \sigma_d$: This means that the fitness valley is "deep". The qualitative behavior of the system is quite different for shallow and deep valleys [121]. In shallow valleys, small recombination can even increase the rate of fixation. We are here interested in the case of strong selection against single mutants. This scenario much more clearly exhibits the evolutionary bottlenecks on the route to fixation.

- $\frac{\mu}{\sigma_d} \ll 1$: Although the first condition rules out a mutation-selection balance for the deleterious single mutants, we still expect $\frac{\mu}{\sigma_d}$ to be the scale of the average frequency of single mutants.

- It should be clear, that the rate with which the first double mutants AB appear (denoted by γ) involves the combinations $\mu\, \frac{\mu}{\sigma_d}$ (compensatory mutation of a single mutant) and $\rho\, \frac{\mu}{\sigma_d}\, \frac{\mu}{\sigma_d}$ (recombination of two single mutant parents). The previous conditions make $\gamma \ll 1$ clearly the slowest time scale in the system.

- The parameter ρ will be used to tune the system through a phase transition at $\rho \sim \sigma_b$, as we will see later. At least we demand that $\rho \gg \mu$.

5.3 Expansion of the Master equation

At this point, the large N limit is taken and the Master equation (5.2) is expanded to leading order in $1/N$ to arrive at the corresponding Fokker-Planck equation. The usual routine is to start with the van Kampen linear noise approximation [40], i.e. setting $n_i = N\phi_i(\tau) + \sqrt{N}\xi_i$. This is a separation into a macroscopic part ϕ and small quadratic fluctuations ξ around it. The macroscopic part evolves according to the deterministic equation:

$$\partial_\tau \phi_i(\tau) = \left(\boldsymbol{\alpha}_{1,0}\right)_i(\boldsymbol{\phi}(\tau)), \quad \tau = \frac{t}{N} \tag{5.17}$$

where $\boldsymbol{\alpha}_{1,0}$ is the vector of first jump moments using the zeroth order term of the transition rates in their canonical form (5.16), see [40].

$$\left(\boldsymbol{\alpha}_{1,0}\right)_i(\boldsymbol{x}) := \left(B_i(\boldsymbol{x}) - D_i(\boldsymbol{x})\right)^{(0)} \equiv 0 \tag{5.18}$$

The disappearance of the macroscopic law renders the linear noise approximation invalid, because fluctuations ultimately grow to sizes much greater than $\sqrt{N}$ [40]. The correct expansion for Master equations of this *diffusion type* is in terms of the intensive frequency $\boldsymbol{x} = \frac{\boldsymbol{n}}{N}$ which yields the non-linear Fokker-Planck equation.

$$\partial_\tau P(\boldsymbol{x},\tau) = \left[-\partial_{x_i}\left(\boldsymbol{\alpha}_{1,1}\right)_i(\boldsymbol{x}) + \tfrac{1}{2}\,\partial_{x_i}\partial_{x_j}\left(\boldsymbol{\alpha}_{2,0}\right)_{ij}(\boldsymbol{x})\right] P(\boldsymbol{x},\tau), \quad \tau = \frac{t}{N^2} \tag{5.19}$$

In our case, the first and second jump moments given by the first and zeroth order terms in (5.16), respectively, are:

$$\left(\boldsymbol{\alpha}_{1,1}\right)_i(\boldsymbol{x}) = \left(B_i(\boldsymbol{x}) - D_i(\boldsymbol{x})\right)^{(1)} = b_i(\boldsymbol{x}) - d_i(\boldsymbol{x}) =: F_i(\boldsymbol{x}) \tag{5.20}$$

$$\begin{aligned}\left(\boldsymbol{\alpha}_{2,0}\right)_{ij}(\boldsymbol{x}) &= \left[\delta_{ij}\left(B_i(\boldsymbol{x}) + D_i(\boldsymbol{x})\right) - 2B_i(\boldsymbol{x})D_j(\boldsymbol{x})\right]^{(0)} \\ &= 2\left(x_i\,\delta_{ij} - x_i x_j\right) =: D_{ij}(\boldsymbol{x})\end{aligned} \tag{5.21}$$

where we have identified the drift terms F_i and diffusion terms D_{ij} of the Fokker-Planck equation.

Note: The above moments can be derived explicitly by expanding the shift operators in the Master equation (5.2) to second order and rearranging terms. For analytical functions f, we have by Taylor expansion:

$$\mathbb{E}^{\pm} f(n) = f(n \pm 1) = f(n) \pm f'(n)1^1 + \frac{1}{2}f''(n)1^2 + \cdots = \sum_{k=0}^{\infty}\frac{f^{(k)}(n)}{k!} = e^{\partial_n} f(n)$$

$$\mathbb{E}_i^{\pm} = e^{\pm\partial_{n_i}} = e^{\pm\frac{1}{N}\partial_{x_i}} = 1 \pm \frac{1}{N}\partial_{x_i} + \frac{1}{2N^2}\partial_{x_i}^2 + \ldots \tag{5.22}$$

5.3.1 The drift terms of the Fokker-Planck equation

Now that the Fokker-Planck equation (5.19) is established as the leading order in a systematic expansion in $\frac{1}{N}$, it is important to note that there is no macroscopic law in the sense of equation (5.17). All that one has at this moment is the following non-closed evolution equation for the mean frequencies that involves higher moments due to non-linear drift terms F_i (equation 5.20).

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\langle x_i\rangle = \frac{\mathrm{d}}{\mathrm{d}\tau}\int \mathrm{d}^4x\, x_i\, P(\boldsymbol{x},\tau) = \int \mathrm{d}^4x\, x_i \left(-\sum_j \partial_j F_j(\boldsymbol{x}) + \sum_{k,j}\partial_k\partial_j D_{kj}(\boldsymbol{x})\right) P(\boldsymbol{x},\tau)$$

$$\Rightarrow \quad \frac{\mathrm{d}}{\mathrm{d}\tau}\langle x_i\rangle = \langle F_i(\boldsymbol{x})\rangle + \text{boundary terms} \tag{5.23}$$

(In most cases, boundary terms can be neglected.) This equation for the mean is not really helpful. If however the drift terms themselves involve parameters that are large, we can carry out a secondary expansion. Remember, that the limit $N \to \infty$ is already carried out. However, one could take the limit e.g. $\sigma \to \infty$. In a way completely analogous to the first linear noise approximation, one could set $x_i = \phi_i(\tau) + \frac{1}{\sqrt{\sigma}}\xi_i$ and expand in powers of $\sqrt{\sigma}$. Within this approximation, one would recover a new "macroscopic law" that involves the drift terms mentioned before.

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\phi_i(\tau) = F_i(\boldsymbol{\phi}) \tag{5.24}$$

Note: In the one-locus/two-allele model this secondary linear noise approximation would only work in the limit $Nu = \mu \gg 1$, where the macroscopic law would exhibit a single stable fix point at $\frac{\mu}{\sigma}$ (mutation-selection balance) and fluctuations are small of the order $\frac{1}{\sqrt{\mu}}$. In the opposite case $\mu \ll 1$, this expansion would not work, since the stationary states would be at the boundaries $x = 0$ and $x = 1$, where a continuum approximation is not warranted.

Thus we see that the drift terms do play an important role and the location of their fixed points will influence the behavior of the system both qualitatively and quantitatively.

5.4 Stationary distribution of the fast variables

We now return to the problem of finding the fixation rate of double mutants. Initially, the population starts in the wild type, i.e. $x_{ab} = 1,\ x_{i\neq ab} = 0$. Before even the very first double mutant appears (on a time scale $1/\gamma$), a quasi-stationary distribution of single mutants is established on the much shorter time scale of $1/s_d$

[38]. For $s_d \gg s_b$, i.e deep fitness valleys, this is in fact the shortest time scale present and one can expect a true separation of time scales. This would mean that there is always an instantaneous quasi-stationary distribution for the single mutants that follows the evolution of the slower modes adiabatically. This adiabatic elimination is quite technical and performed in detail in appendix D on both the deterministic and the stochastic level of the model. It is strongly suggested to first follow the exposition in this main part to gain an intuition for the qualitative effects involved before verifying the validity of time scale separation in the appendix. We will here follow a more pragmatic route. The dynamics of the single mutants are analyzed while regarding the double mutant frequency x_{AB} as effectively *constant*, i.e. changing on a time scale too slow to notice. The wild type frequency can be eliminated due to the normalization $\sum_i x_i = 1$. If, furthermore, the frequency of single mutants remains small at all times (this will need to be checked *a posteriori*), the following two crucial assumptions can be made:

1. The two single mutant populations are statistically independent and symmetric in the sense that $\langle x_{aB}\rangle = \langle x_{Ab}\rangle$ and $\langle x_{aB}x_{Ab}\rangle \approx \langle x_{aB}\rangle\langle x_{Ab}\rangle$.

2. The Master equation describing their behavior can be linearized in $x_{aB,Ab}$.

Anticipating, that for $\mu \ll 1$ boundary effects might be important, we proceed cautiously and return to the birth-death Moran Master equation for fixed x_{AB}. For each single mutant species individually it is:

$$\partial_t P(n_i,t) = \left[(\mathbb{E}^+ - 1)\, f_i^-(\boldsymbol{x}) + (\mathbb{E}^- - 1)\, f_i^+(\boldsymbol{x})\right] P(n_i,t), \quad i = aB,Ab \tag{5.25}$$

$$f_i^+(\boldsymbol{x}) := B_i(\boldsymbol{x})\,(1 - D_i(\boldsymbol{x})), \quad f_i^-(\boldsymbol{x}) := D_i(\boldsymbol{x})\,(1 - B_i(\boldsymbol{x})) \tag{5.26}$$

The rates are expanded with $x_{ab} = 1 - \sum_{i\neq ab} x_i$ to first order in x_{aB}, x_{Ab} and to lowest order in $\frac{u}{s_d}$ and $\frac{u}{s_b}$. Setting $x_{AB} \to z$, one gets

$$f_i^-(n_i;z) =: R_-(z)\,n_i + \mathcal{O}\left(n_i^2\right), \quad f_i^+(n_i;z) =: \nu(z) + R_+(z)\,n_i + \mathcal{O}\left(n_i^2\right) \tag{5.27}$$

$$\nu(z) = u + rz(1-z), \quad R_\pm = \frac{1}{N}\left[1 + \mathcal{O}\left(u, s_b, s_d, r\right)\right] \tag{5.28}$$

$$\Delta R(z) := R_+(z) - R_-(z) = -\frac{1}{N}\left[s_d + z\left(s_b + r\right) - 4u\right] < 0 \tag{5.29}$$

The effective "mutation" rate $\nu(z)$ contains two contributions for the production of new single mutants: first the mutational channel $\propto u$ and second the recombinatorial reshuffling channel $\propto rz(1-z)$ (refer to the discussion in section 5.2.1). We see that its scaled pendant $N\nu(z)$ will ultimately exceed values of order one whenever $Nr = \rho \gg 1$ and $z(1-z) \gg \frac{1}{Nr}$, quite independently from the value of Nu itself. To repeat, for large recombination $Nr \gg 1$, the characteristics of the

single mutant dynamics will explore quite different qualitative sectors. Owing to the linearity of the transition rates, we can give the stationary distribution $P^s(n_i)$ analytically [40] in terms of its generating function.

$$G(\zeta,\tau) := \sum_{n_i=0}^{\infty} \zeta^n P(n_i,\tau) \quad \leftrightarrow \quad P(n_i,\tau) = \frac{1}{n_i!}\,\partial_\zeta^{(n_i)}\Big|_{\zeta=0} G(\zeta,\tau) \tag{5.30}$$

$$G(\zeta,t) = \left(\frac{|\Delta R|}{R_- - R_+\,\zeta - e^{-|\Delta R|t}\,R_+(1-\zeta)}\right)^{\frac{\nu}{R_+}} \xrightarrow{\;t\gg|\Delta R|^{-1}\;} G^s(\xi) := \left(\frac{|\Delta R|}{R_- - R_+\,\zeta}\right)^{\frac{\nu}{R_+}} \tag{5.31}$$

In passing, we note that the quasi-stationary distribution is indeed established on a time scale $\frac{1}{|\Delta R|} \approx \frac{1}{s_d}$. The mean value is calculated directly as:

$$\langle n_i\rangle^s = \partial_\zeta\big|_{\zeta=1} \ln G^s(\zeta) = \frac{\nu}{|\,\Delta R\,|} \tag{5.32}$$

The mean value of each single mutant distribution is thus found to be:

$$\boxed{\langle x_i\rangle^s = \frac{u + r z(1-z)}{s_d + z(s_b+r) - 4u}, \quad i = aB, Ab} \tag{5.33}$$

For consistency, it needs to be ensured that $\langle x_i\rangle^s \ll 1$ at all times (or for all values of $z \in [0,1]$), lest we leave the regime of the linear approximation. This can only be guaranteed, if $r, s_b \ll s_d$, i.e. for deep fitness valleys:

$$\partial_z \langle x_i\rangle^s \stackrel{!}{=} 0 \quad \Rightarrow \quad z_{\max} = \frac{\sqrt{s_d}}{\sqrt{s_d} + \sqrt{s_d + s_b + r}} + \mathcal{O}\left(\frac{u}{s}\right) \tag{5.34}$$

$$\Rightarrow \quad \langle x_i\rangle^s_{\max} = \frac{r}{(r+s_b)^2}\left(r + s_b + 2s_d - 2\sqrt{s_d(s_d+s_b+r)}\right) + \mathcal{O}\left(\frac{u}{s}\right) \tag{5.35}$$

$$= \begin{cases} \frac{r}{r+s_b}\left(1 - 2\sqrt{\frac{s_d}{r+s_b}}\right) + \mathcal{O}\left(\frac{s_d}{r,s_b}\right) + \mathcal{O}\left(\frac{u}{s}\right) & \text{shallow valley} \\ \frac{r}{4\,s_d}\left(1 - \frac{r+s_b}{2\,s_d}\right) + \frac{u}{s_d} + \mathcal{O}\left(\left(\frac{r,s_b}{s_d}\right)^3\right) & \text{deep valley} \end{cases}$$

It seems that the linear approximation to the Master equation by itself is consistent for deep valleys $s_d \gg s_b, r$, where it is also in accordance with the time scale separation. (It might be even valid for shallow valleys in the (rather special) case $s_d \ll r \ll s_b$. Later, we will see that $r = \mathcal{O}(s_b)$ is the region of interest.)

For completeness, let us look at the variance and second moment of the single mutant distribution

$$\mathrm{var}^s(n_i) := \langle n_i^2\rangle - (\langle n_i\rangle)^2 = \partial_\zeta^2\Big|_{\zeta=1} \ln G^s(\zeta) + \langle n_i\rangle^s = \frac{\nu(z)R_-}{\Delta R^2}$$

$$\boxed{\mathrm{var}^s(x_i) \approx \frac{1}{N}\frac{u+rz(1-z)}{(s_d+z(s_b+r))^2} = \frac{\langle x_i\rangle^2}{Nu+Nrz(1-z)}, \quad i=aB,Ab} \tag{5.36}$$

$$\boxed{\langle x_i^2\rangle^s = \frac{\nu\,(\nu+R_-)}{N^2\Delta R^2}, \quad i=aB,Ab} \tag{5.37}$$

So indeed, fluctuations are strong for $N\nu(z) = Nu + Nrz(1-z) \ll 1$. The quasi-stationary distribution itself can be given as well (see [40]).

$$P^s(n_i) = \frac{|\Delta R|^{\nu/R_+}}{R_-^{\nu/R_+ + n_i}} \prod_{j=0}^{n_i-1}\left(\frac{\nu+R_+ j}{1+j}\right) \approx (N\,|\,\Delta R\,|)^{\nu/R_+}\,\frac{\Gamma(N\nu+n_i)}{\Gamma(N\nu)\Gamma(n_i+1)} \tag{5.38}$$

5.5 The effective dynamics for double mutants

If the two main assumptions of time scale separation and statistical independence of the two single mutant species hold, one can set

$$x_i \to \langle x_i\rangle(z), \quad x_{aB}x_{Ab} \to \langle x_{aB}x_{Ab}\rangle(z) = (\langle x_i\rangle(z))^2, \quad i=aB,Ab,\ z=x_{AB} \tag{5.39}$$

The fast fluctuating modes of the system are replaced by their instantaneous mean value, which depends parametrically on the remaining slow mode: the frequency of double mutants. This step is essentially an adiabatic decoupling of the dynamics. The result is an effective description of the slow mode. The problem is now a one dimensional one and subject to further analysis.

5.5.1 The initial linear regime

Proceeding similarly to above, first the effective Master equation for the double mutants with transition rates to linear order in $z = x_{AB}$ will be derived. Beware, that this will capture the initial behavior of the double mutant species only. Ultimately, one needs to go beyond the linear regime to capture the full path to fixation.

$$\partial_t P(n,t) = \left[(\mathbb{E}^+ - 1)\,(\Gamma_- n) + (\mathbb{E}^- - 1)\,(\gamma + \Gamma_+ n)\right] P(n,t), \quad n = n_{AB}$$

where the terms $\Gamma_\pm$ and γ on the right hand side are defined as the coefficients of a linear expansion in $n = n_{AB}$ of the original transition rates. This is structurally

exactly the same linear Master equation from the last section, so the solution and moments are known. The parameters are given by replacing the x_i ($i = aB, Ab$) by their mean (see (5.39)) and expanding to $\mathcal{O}(z)$. One can then read off the effective parameters (to leading order in u).

$$\gamma \approx u\frac{2u}{s_d} + r\frac{u^2}{s_d^2} \overset{!}{\ll} \frac{1}{N} \tag{5.40}$$

$$N\Gamma_+ \approx 1 - r - 2u\left(1 - \frac{r(r+2s_d)}{s_d^2}\right) \tag{5.41}$$

$$N\Gamma_- \approx 1 - s_b - 2u \tag{5.42}$$

$$N\Delta\Gamma \approx (s_b - r) + 2u\frac{r(r+2s_d)}{s_d^2} \tag{5.43}$$

We finally recognize γ as the effective production rate of initial double mutants. As was anticipated in section 5.2.3, its two contributions describe the two channels of double mutant production: the compensatory mutation of a single mutant and the recombination of two single mutant parents. If γ is much smaller than one per generation, we are in a fluctuation dominated regime. The effective initial selection pressure on the double mutants is $\Delta\Gamma$. But contrary to the situation with the permanently diadvantaged single mutants, there is now a critical recombination value where this quantity changes sign:

$$\Gamma_+ = \Gamma_- \quad \Rightarrow \quad r = r_c := s_b + \frac{2u}{s_d}\frac{s_b}{s_d}(s_b + 2s_d) + \mathcal{O}\left(u^2\right) \tag{5.44}$$

To see what the immediate qualitative result of that transition is, we look at the rate on which the number of double mutants in the population reaches an arbitrary intermediate size $n_f \geq 1$. Since $\gamma \ll 1$, we can separate the rate for this particular process into the rate for the first arrival of a double mutant (with rate γ) and the subsequent growth of that subpopulation to a size n_f without going extinct on the way. For this second stage, we can ignore further production of double mutants (i.e. set $\gamma = 0$). The probability $\pi_R(n_i = 1)$ of escape [41] through $R = n_f$ (rather than through the extinction terminal at $L = 0$) starting at $n_i = 1$ is given by ([40], eq. XII.2.8, p. 300)

$$\text{Rate}(0 \to n_f) \approx \gamma\,\pi_{R=n_f}(n_i = 1) \tag{5.45}$$

$$\pi_R(n_i) := \left[1 + \sum_{k=L+1}^{n_i-1} \prod_{j=L+1}^{k} \frac{f^-(j)}{f^+(j)}\right] \Bigg/ \left[1 + \sum_{k=L+1}^{R-1} \prod_{j=L+1}^{k} \frac{f^-(j)}{f^+(j)}\right] \tag{5.46}$$

$$n_i = 1, L = 0, R = n_f, \quad f^+(n) := \gamma + \Gamma_+ n,\ f^-(n) := \Gamma_- n$$

$$\Rightarrow\ \pi_{R=n_f}(n_i = 1) = \frac{1 - \Gamma_-/\Gamma_+}{1 - (\Gamma_-/\Gamma_+)^{n_f}} = \frac{\Delta\Gamma/\Gamma_+}{1 - (1 - \Delta\Gamma/\Gamma_+)^{n_f}} \tag{5.47}$$

$$\approx \frac{\Delta\Gamma/\Gamma_+}{1 - e^{-n_f \Delta\Gamma/\Gamma_+}} \approx \begin{cases} \frac{1}{n_f} & n_f < \frac{\Gamma_+}{|\Delta\Gamma|} \sim \frac{1}{|s_b - r|} \\ \frac{\Delta\Gamma}{\Gamma_+} & n_f > \frac{\Gamma_+}{|\Delta\Gamma|},\ \Delta\Gamma > 0 \\ \frac{|\Delta\Gamma|}{\Gamma_+} e^{-n_f |\Delta\Gamma|/\Gamma_+} & n_f > \frac{\Gamma_+}{|\Delta\Gamma|},\ \Delta\Gamma < 0 \end{cases} \tag{5.48}$$

First, the young double mutant subpopulation grows *neutrally* (without the influence of any selection) up to a size of $n_f \approx \Gamma_+/\Delta\Gamma \sim |s_b - r|^{-1}$. The probability of growth to even larger values of n_f is either constant (in the case of a fitness advantage) or exponentially suppressed (see Figure 5.3). This is of course exactly the same phenomenon and calculation, that applies in the one-locus/two-allele model and ultimately leads to the all-important substitution rate equation (5.1). The corresponding frequency threshold for this initial neutral zone is denoted with z_n.

$$z_\mathrm{n} := \frac{1}{N|s_b - r|} \tag{5.49}$$

5.5.2 Fix points of the effective drift term

The calculation above is strictly limited to the linear regime of the double mutants. To assess fixation ($n \to N \Leftrightarrow z \to 1$), we need to look for encounters with non trivial fix points/saddle points of the full problem, i.e. values of $z = x_{AB}$ with $F_z(z) = 0$. $F_z(z)$ is here the drift term of the one-dimensional Fokker-Planck that results from the full Fokker-Planck equation (5.19) through the replacement (5.39).
Especially for $r > r_c \sim s_b$, we expect the presence of two stable fix points $z_0 \approx 0$, $z_1 \approx 1$ and an unstable fix point z_cr in between. However, the last section also shows that the size z_n of the initial "neutral zone" needs to be taken into account. We have thus the following scales to compare.

$$z_\mathrm{n} \approx \frac{1}{N|s_b - r|} \quad \text{vs.} \quad z_0 \approx \frac{\gamma}{r - s_b} \quad \text{vs.} \quad z_\mathrm{cr} \quad \text{vs.} \quad z_1 \approx 1 - \frac{2u}{s_d + s_b + r} \quad (r > r_c)$$

An approximation for the saddle point z_cr can be found by realizing that mutational processes only ensure that the boundaries at $z = 0$ and $z = 1$ are not absorbing. For all other matters, esp. the location of z_cr, we can set $u = 0$. Anticipating that the

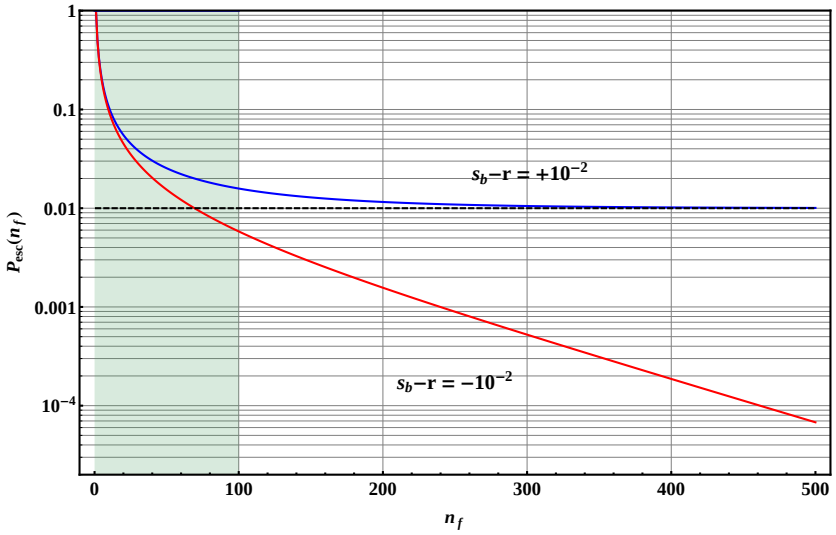

Figure 5.3: The probability of growth of a double mutant subpopulation to a size $n_f \geq 1$ (before going extinct at $n = 0$) starting at $n_i = 1$ according to equation (5.48) for $s_b - r = +10^{-2}$ (blue) and for $s_b - r = -10^{-2}$ (red). Within the "neutral zone" (shaded blue), the evolution of the subpopulation is effectively neutral. Note that this describes only the initial growth phase.

fixation dynamics for $r > r_c$ will be well captured by a Fokker-Planck equation, we now turn to the effective drift term for the double mutants therein (refer to equation (5.19) and following).

$$\partial_\tau P(z,\tau) = \left\{-\partial_z F_z(z) + \partial_z^2 z(1-z)\right\} P(z,\tau), \quad \tau = \frac{t}{N^2} = \frac{\text{gen.}}{N} \tag{5.50}$$

$$F_z(z)|_{\mu=0} = z(1-z)\left[(\sigma_b - \rho) + \frac{2\rho(\rho+\sigma_d)z}{\sigma_d + (\sigma_b+\rho)z} + \frac{\rho^3 z(1-z)}{(\sigma_d + (\sigma_b+\rho)z)^2}\right] \tag{5.51}$$

where we have re-introduced the scaled versions of the parameters (greek letters). To find z_{cr}, only a quadratic equation must be solved. The positive solution exists for $r > r_c^{(0)} = s_b$ and is for deep fitness valleys given by:

$$z_{\text{cr}} := \frac{\rho - \sigma_b}{2\rho}\left(1 - \frac{\rho^2 + \sigma_b^2}{2\sigma_d \rho}\right) + \mathcal{O}\left(\left(\frac{\rho, \sigma_b}{\sigma_d}\right)^2\right) \tag{5.52}$$

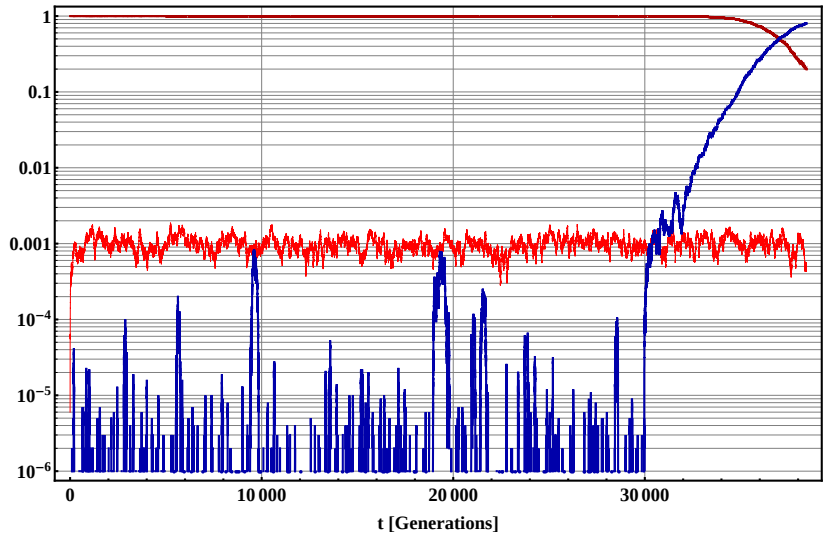

Figure 5.4: A typical fixation trajectory for the allele frequencies of wild type (dark red) single mutants (red) and double mutant (blue). The simulation parameters are $N = 10^6$, $u = 10^{-5}$, $s_b = 10^{-3}$, $s_d = 10^{-2}$ and $r = 0$. The escape of the double mutant is dominated by escape from the neutral zone at $z_n = \frac{1}{N|s_b - r|} = 10^{-3}$. The single mutant frequencies are close to the deterministic fix point at $\frac{u}{s_d} = 10^{-3}$

To capture the essential behavior we simplify the drift term by a simple polynomial with the same zeros, where we need to fix the prefactor α.

$$\left.\frac{F_z}{z(1-z)}\right|_{z_{cr}} \stackrel{!}{=} 0 \quad \Rightarrow \quad \frac{F_z}{z(1-z)} \approx 2\alpha\,(z - z_{cr}) \tag{5.53}$$

$$\left.\frac{F_z}{z(1-z)}\right|_{z=0} = \sigma_b - \rho \stackrel{!}{=} -2\alpha z_{cr} \quad \Rightarrow \quad \alpha := \frac{\rho - \sigma_b}{2 z_{cr}} \tag{5.54}$$

If recombination is high enough ($r > r_c \approx s_b$), the AB clone that has successfully reached a size greater than $z_n \sim \frac{1}{N|r - s_b|}$ is (depending on the population size) faced with climbing a hill with the tip at $z_{cr} \sim \frac{r - s_b}{2r}$. The probability of success for this secondary escape can be evaluated using the Fokker-Planck approximation of that process (with $u = 0$). However, this is not the standard problem of Kramer's escape, because the escape does not start in a local potential minimum but rather in the flank of the potential hill.

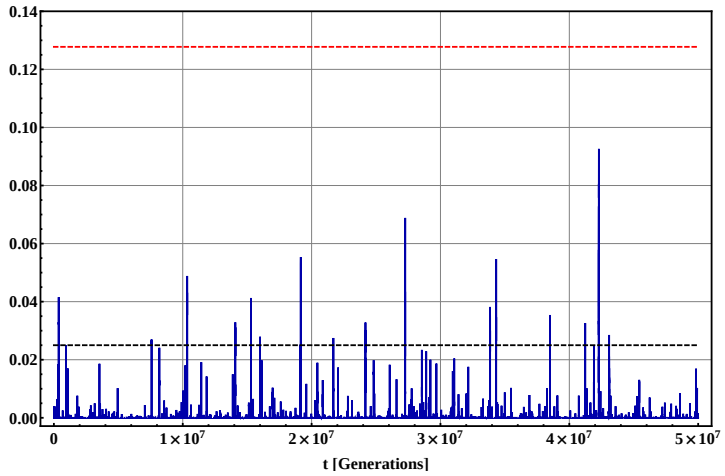

Figure 5.5: A typical fixation trajectory for the allele frequency of double mutants (blue). The simulation parameters are $N = 10^5$, $u = 10^{-5}$, $s_b = 10^{-3}$, $s_d = 10^{-2}$ and $r = 1.4 \cdot 10^{-3} > r_c$. The escape of the double mutant is not dominated by escape from the neutral zone at $z_n = \frac{1}{N|s_b - r|} \approx 0.025$, but escape over the saddle point at $z_{cr} \approx 0.13$. The trajectory leaves the neutral zone several times before the final fixation succeeds (not shown).

5.5.3 Classification of the fixation dynamics

We are now in a position to classify different fixation scenarios, depending on the strength of recombination. This list is meant to give an overview over the different qualitative regimes of the system. We will later concentrate on just a few cases of interest. (All times are measured in generations.)

1. $r < s_b$: There is only the stable fix point at $z_1 \approx 1$ (fixation).

   (a) For $N\gamma \gg 1$, the fixation time is determined by the deterministic equation of motion: $T_{\text{fix}} \propto \int_0^1 \frac{dz}{F_z(z)}$.

   (b) For $N\gamma \ll 1$, we need to wait until one of the double mutants eventually grows to a clone of size larger than $z_n \sim \frac{1}{N|s_b - r|}$, i.e. out of the neutral zone, after which it then quickly fixates. This means $\Gamma_{\text{fix}} \sim N\gamma(s_b - r)$.

2. $r > s_b$: There are three fix points of the drift term: $z_0 < z_{cr} < z_1$

   (a) $N\gamma \gg 1$, i.e. $z_n < z_0$: The size z_n is reached in deterministic time, i.e. there is no fluctuation barrier. The clone quickly settles in the "local minimum" z_0. The fixation is then dominated by Kramer's escape over the saddle point z_{cr}. The effective drift taking the fix points at the

boundaries into account is:

$$F_z \propto (z-z_0)(z-z_{\rm cr})(z-z_1) \tag{5.55}$$

$$z_0 \approx \frac{N\gamma}{\rho-\sigma_b},\ z_{\rm cr} \approx \frac{\rho-\sigma_b}{2\rho},\ z_1 \approx 1-\frac{N\gamma}{\rho+\sigma_b}-\frac{2\mu}{\rho+\sigma_b+\sigma_d} \tag{5.56}$$

The escape rate is then the standard Kramer's escape rate [124]:

$$\Gamma_{\rm fix} \approx \frac{\sqrt{F_z'(z_0)\,|\,F_z'(z_{\rm cr})\,|}}{2\pi}\sqrt{\frac{z_0(1-z_0)}{z_{\rm cr}(1-z_{\rm cr})}}\,e^{\int_{z_0}^{z_{\rm cr}} {\rm d}z\ \frac{F_z(z)}{z(1-z)}}$$

This will not be further elaborated here, for the validity of this formula requires extremely large population sizes.

(b) $N\gamma \ll 1$, i.e $z_0 < z_{\rm n}$ with $z_{\rm n} < z_{\rm cr}$: Once the clone is over the fluctuation barrier at $z_{\rm n}$ it finds itself in the uphill flank of the potential hill. Given the parameter regime that was specified at the beginning of this chapter, this will be the typical fixation scenario for $r > s_b$

(c) $N\gamma \ll 1$, i.e. $z_0 < z_{\rm n}$ with $z_{\rm cr} < z_{\rm n}$: Once the clone is over the fluctuation barrier it would find itself in the downhill flank of the potential hill. The fixation would be dominated solely by escape over the neutral barrier $z_{\rm n}$. However, this situation almost never happens for reasons we will see shortly.

The remainder of this work will be devoted to the small mutation scenarios (1.b) and (2.b) above. The presence of an instable fix point at $z_{\rm cr}$, but also the fact that $z_{\rm n} \sim \frac{1}{N|r-s_b|}$ grows large at $r \sim s_b$ forces us to go beyond the linear approximation to the Master equation to find the fixation rate $\Gamma_{\rm fix}$.

5.5.4 Estimation of the escape probability $P_{\rm esc}$

The strategy to find the fixation rate is essentially similar to the considerations above. For $N\gamma \ll 1$, we can split a fixation event into the rare arrival of a first double mutant into the population and its subsequent growth and escape to fixation. This means that we must find the conditional probability of escape $P_{\rm esc}$ of that first seed. Here it will be calculated with the use of the Fokker-Planck approximation to the Master equation immediately for $z \geq \frac{1}{N}$. This can be expected to be a valid approximation if the important processes take place away from the boundary at $z = 0$. Encouraging is the fact that the same strategy leads to the correct substitution rate (5.1) for the one-locus/two-allele model, as well. The analogous calculation of the splitting probability for the Fokker-Planck equation can be found in

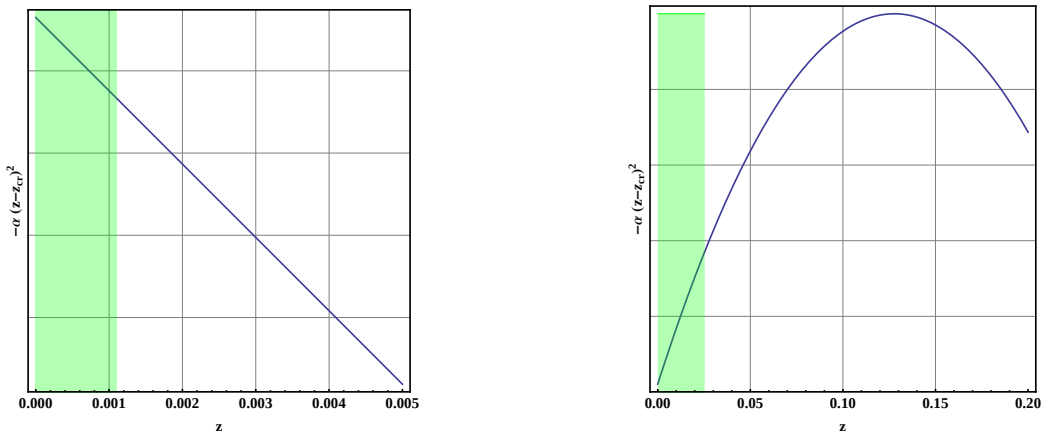

Figure 5.6: The effective potential $-\alpha\,(z-z_{\rm cr})^2$ (blue) for the dynamics of the double mutants. Below the critical recombination size (left), the potential outside of the neutral zone (green) allows instant fixation. Above criticality, a successful double mutant trajectory needs to climb the potential hill, after it has left the neutral zone (right).

[41]. First, we restate the effective one-dimensional Fokker-Planck equation for the double mutant frequency $z = n_{AB}/N$, which was derived in the last sections.

$$\partial_\tau P(z,\tau) = \left\{-\partial_z F_z(z) + \partial_z^2 z(1-z)\right\} P(z,\tau), \quad \tau = \frac{t}{N^2} = \frac{\text{gen.}}{N} \tag{5.57}$$

$$F_z(z)|_{\gamma=0} \approx 2\,\alpha\; z(1-z)\,(z-z_{\rm cr}) \tag{5.58}$$

$$z_{\rm cr} \approx \frac{\rho-\sigma_b}{2\rho}\left(1-\frac{\rho^2+\sigma_b^2}{2\rho\,\sigma_d}\right), \quad \alpha := \frac{\rho-\sigma_b}{2\,z_{\rm cr}} \approx \rho + \frac{\sigma_b^2}{\sigma_d} \tag{5.59}$$

Given this particular Fokker-Planck equation with *drift* term $F_z(z)$ an *diffusion* term $z(1-z)$, one can approximate the total rate of double mutant fixation $\Gamma_{\rm fix}$ in exactly the same way as was done at the beginning of chapter two for the simple one-locus/two-allele model: as a product of the arrival rate of new double mutants ($N\gamma$) and their probability of subsequent escape to fixation ($P_{\rm esc}$).

$$\Gamma_{\rm fix} \approx N\gamma\, P_{\rm esc}, \quad \text{with} \quad P_{\rm esc} := \lim_{z\to 1} \pi_{\rm esc}(z) \tag{5.60}$$

$$\pi_{\rm esc}(z) := \pi_{R=z}(1/N) = \frac{\int_0^{1/N} {\rm d}z'\ \psi(z')}{\int_0^{z} {\rm d}z'\ \psi(z')} \tag{5.61}$$

$$\text{and} \quad \psi(z) := \exp\left(-\int_0^z {\rm d}z' \left.\frac{F_z(z')}{z'(1-z')}\right|_{\gamma=0} \right) \propto \exp\left(-\alpha\,(z-z_{\rm cr})^2\right) \tag{5.62}$$

The integrals in the fraction for $\pi_{\rm esc}(z)$ can be expressed in terms of the error-function:

$$\pi_{\rm esc}(z) = \frac{{\rm Erf}\left(\sqrt{\alpha}\left(\frac{1}{N}-z_{\rm cr}\right)\right)+{\rm Erf}\left(\sqrt{\alpha}\,z_{\rm cr}\right)}{{\rm Erf}\left(\sqrt{\alpha}\,(z-z_{\rm cr})\right)+{\rm Erf}\left(\sqrt{\alpha}\,z_{\rm cr}\right)} \tag{5.63}$$

$${\rm Erf}(x) := \frac{2}{\sqrt{\pi}}\int_0^x {\rm d}t\ e^{-t^2} \approx \begin{cases} \frac{2}{\sqrt{\pi}}\,x, & |x| \ll 1 \\ 1-\frac{e^{-x^2}}{\sqrt{\pi}\,x}, & x \gg +1 \\ -1+\frac{e^{-x^2}}{\sqrt{\pi}\,|x|}, & x \ll -1 \end{cases} \tag{5.64}$$

For $z_{\rm cr} > 0$, the denominator of $\pi_{\rm esc}(z)$ varies significantly only in the neighborhood of $z = z_{\rm cr}$ with a width of $\frac{1}{\sqrt{\alpha}}$. Thus, as long as $z_{\rm cr} + \frac{1}{\sqrt{\alpha}} < 1$, we can safely set the z-dependent part in $\pi_{\rm esc}(z)$ to its limiting value, which is unity. Note, that for $r = \mathcal{O}(s_b)$ we have $\alpha = \mathcal{O}(Nr) \gg 1$. The remaining quantity to be considered is the (scaled) distance of $z_{\rm cr}$ to $z = 0$. This is in fact the parameter that tunes the system through a sort of phase transition of the fixation dynamics.

By taking the leading order contribution to the numerator of $\pi_{\rm esc}$ for $\sqrt{\alpha}/N = \mathcal{O}\left(\sqrt{r/N}\right) \ll 1$, we have in principle the desired result for the escape probability:

$$P_{\rm esc} = \frac{{\rm Erf}\left(\frac{\sqrt{\alpha}}{N}-A\right)+{\rm Erf}(A)}{1+{\rm Erf}(A)} \approx \frac{\frac{2}{N}\sqrt{\frac{\alpha}{\pi}}\,e^{-A^2}}{1+{\rm Erf}(A)}, \quad \text{with} \quad A := \sqrt{\alpha}\,z_{\rm cr} \tag{5.65}$$

To generate more explicit results involving the original parameters, we can use the three asymptotic expansions for the error-function to obtain limiting exponential expressions for the escape probability.

$$P_{\rm esc} \approx \begin{cases} \frac{2\sqrt{\alpha}}{N}\,|A| & A \ll -1 \\ \frac{2}{N}\sqrt{\frac{\alpha}{\pi}}\,e^{-\frac{2}{\sqrt{\pi}}A-\left(1-\frac{2}{\pi}\right)A^2} & |A| \ll 1 \\ \frac{1}{N}\sqrt{\frac{\alpha}{\pi}}\,e^{-A^2} & A \gg +1 \end{cases} \tag{5.66}$$

In the intermediate regime ($|A| \ll 1$) the following logarithmic expansion was used.

$$\ln\left(1+\mathrm{Erf}(A)\right) = \mathrm{Erf}(A) - \frac{1}{2}\mathrm{Erf}^2(A) + \cdots = \frac{2}{\sqrt{\pi}}A - \frac{2}{\pi}A^2 + \mathcal{O}\left(A^3\right) \tag{5.67}$$

Altogether and in terms of the original parameters, the central result for the fixation rate below, at and above critical recombination strength can now be stated.

$$\boxed{\Gamma_{\mathrm{fix}} \approx \begin{cases} N\gamma\,|r-s_b| & A < -1 \\ \gamma\sqrt{\frac{4\alpha}{\pi}}\, e^{-\frac{2}{\sqrt{\pi}}\sqrt{\alpha}z_{\mathrm{cr}} - \left(1-\frac{2}{\pi}\right)\alpha z_{cr}^2} & |A| < 1 \\ \gamma\sqrt{\frac{\alpha}{\pi}}\, e^{-\alpha z_{cr}^2} & A > +1 \end{cases}} \tag{5.68}$$

$$\text{with} \quad z_{\mathrm{cr}} \approx \frac{r-s_b}{2r}\left(1-\frac{r^2+s_b^2}{2s_d r}\right), \quad \alpha = \frac{N(r-s_b)}{2z_{\mathrm{cr}}} \approx N\left(r+\frac{s_b^2}{s_d}\right) \tag{5.69}$$

$$\text{and} \quad A = \sqrt{\alpha}z_{\mathrm{cr}} \approx \sqrt{\frac{N}{4r}}(r-s_b), \quad \gamma = \frac{2u^2}{s_d} + r\frac{u^2}{s_d^2} \tag{5.70}$$

The conditions for the three regimes are also expressed in the original parameters.

$$A \lessgtr \mp 1 \quad \Rightarrow \quad Nr \lessgtr Ns_b + 2 \mp 2\sqrt{Ns_b} \tag{5.71}$$

For small recombination $r < s_b$, the earlier conjecture is confirmed, i.e. that escape from the "neutral zone" dominates the fixation rate. For recombination well above criticality, the fixation rate depends exponentially on the combination $A = \sqrt{\alpha}z_{\mathrm{cr}}$. Especially for $A \gg 1$, fixation is exponentially suppressed. For recombination large enough, fixation is - for all practical purposes - blocked and the global fitness maximum of the system is never achieved.

5.5.5 Estimation of the escape barrier z_{esc}

The last section established the mean rate of fixation of the double mutants for all relevant sizes of recombination. Another meaningful and measurable quantity is the critical size that a double mutant clone needs to overcome to be ultimately successful, i.e. the size of the escape barrier z_{esc}. Previously, we identified z_{n} - the size of the "neutral zone" - as one such fluctuation dominated barrier.
In simulations, the escape barrier can be evaluated by measuring the *largest unsuccessful* clone for many independent trajectories. Mathematically, this value can be defined in two ways: (i) as the scale on which $\pi_{\mathrm{esc}}(z)$ (5.63) reaches its constant limiting value; and (ii) as the value of z, from where conditional escape to $z = 0$ is as likely as escape to the other terminal at $z = 1$. This last condition leads

to a definite analytical expression for z, which will be denoted in the following by z_{split}. This is expressed again through the conditional probability $\pi_{L,R}(z)$ of escape through the left or right end of the interval $[L,R]$ starting from z.

$$\pi_{L=0}(z) \stackrel{!}{=} \pi_{R=1}(z) \;\Leftrightarrow\; \int_z^1 \mathrm{d}z'\ \psi(z') \stackrel{!}{=} \int_0^z \mathrm{d}z'\ \psi(z') \tag{5.72}$$

$$\Rightarrow \quad 2\,\mathrm{Erf}\left(\sqrt{\alpha}(z-z_{\text{cr}})\right) = \mathrm{Erf}\left(\sqrt{\alpha}(1-z_{\text{cr}})\right) - \mathrm{Erf}\left(\sqrt{\alpha}z_{\text{cr}}\right) \tag{5.73}$$

$$\boxed{z_{\text{split}} \approx z_{\text{cr}} + \frac{1}{\sqrt{\alpha}}\,\mathrm{Erf}^{-1}\left[\frac{1}{2}\left(1-\mathrm{Erf}\left(\sqrt{\alpha}\,z_{\text{cr}}\right)\right)\right]} \tag{5.74}$$

The three scales for $A = \sqrt{\alpha}\,z_{\text{cr}}$ need to be considered separately to arrive at more useful expressions for z_{split}.

Estimation of z_{esc} well below criticality: $A = \sqrt{\alpha}\,z_{\text{cr}} \ll -1$

To find the scale, on which $\pi_{\text{esc}}(z)$ becomes constant, we set $z =: x\,|z_{\text{cr}}|$ and expand the denominator in (5.63) in $A \ll -1$.

$$\pi_{\text{esc}}^{-1} \propto \mathrm{Erf}\left((1+x)\,|A|\right) - \mathrm{Erf}\left(|A|\right) \overset{A\ll -1}{\propto} \frac{e^{-A^2}}{|A|} - \frac{e^{-(1+x)^2A^2}}{(1+x)\,|A|} = \frac{e^{-A^2}}{|A|}\left[1 - \frac{e^{-A^2x(2+x)}}{(1+x)}\right]$$

This approaches its limiting value when the size of the x-dependent exponent in the term in square brackets above is larger than unity.

$$A^2x(2+x) > 1 \;\Leftrightarrow\; x > \sqrt{1-\frac{1}{A^2}} - 1 \;\Leftrightarrow\; z > \frac{\sqrt{\alpha\,z_{\text{cr}}^2+1} - \sqrt{\alpha\,z_{\text{cr}}^2}}{\sqrt{\alpha}} \tag{5.75}$$

$$\boxed{z_{\text{esc}} \approx \frac{1}{N\,|r-s_b|}\left(1 - \frac{r}{N(r-s_b)^2}\right), \quad A \ll -1} \tag{5.76}$$

Now this is very close to the initial guess $z_{\text{esc}} \sim z_{\text{n}}$. On the other hand, equation (5.74) for z_{split} gives a slightly different result.

$$\boxed{z_{\text{split}} \approx \frac{\ln(2)}{N\,|r-s_b|}, \quad A \ll -1} \tag{5.77}$$

This seeming discrepancy between the two values for the escape barrier can already be found in the simple one-locus/two-allele model, which shares the essential characteristics of stochastic tunneling with the present model for low recombination ($A \ll -1$). Neglecting mutation for the conditional escape process and

using the appropriate Fokker-Planck equation for the probability distribution of the mutant allele frequency x

$$\partial_\tau P(x,\tau) = \left[-\partial_x \sigma x(1-x) + \partial_x^2 x(1-x)\right] P(x,\tau), \quad \sigma = Ns \gg 1 \tag{5.78}$$

we find the equivalent expressions for the escape barrier x_{esc} in this simpler model.

$$\psi(x) = e^{-Nsx} \;\Rightarrow\; \pi_{\text{esc}}(x) = \frac{\int_0^{\frac{1}{N}} \mathrm{d}y\ \psi(y)}{\int_0^{x} \mathrm{d}y\ \psi(y)} = \frac{1-e^{-s}}{1-e^{-Nsx}} \;\Rightarrow\; x_{\text{esc}} = \frac{1}{Ns} \tag{5.79}$$

For the value x_{split} of the frequency, from where conditional escape to the left terminal at $L=0$ is as likely as escape to the right end at $R=1$, we get:

$$1 \stackrel{!}{=} \frac{\pi_{L=0}(x)}{\pi_{R=1}(x)} = \frac{\int_x^{1} \mathrm{d}y\ \psi(y)}{\int_0^{x} \mathrm{d}y\ \psi(y)} = \frac{e^{-Nsx}-e^{-Ns}}{1-e^{-Nsx}} \quad \Rightarrow \quad x_{\text{split}} = \frac{1}{Ns}\ln\left(\frac{2}{1+e^{-Ns}}\right) \tag{5.80}$$

In essence, the discrepancy between z_{esc} and z_{split} stems from the fact that the actual escape barrier for a trajectory is itself a random variable with its own distribution. In the present case, the clones can initially grow neutrally ($\pi_{\text{esc}}(z) \sim \frac{1}{Nz}$) and fluctuations are strong. It is then not sensible to compare scales such as mean and median of the underlying escape distribution.

Estimation of z_{esc} well above criticality: $A = \sqrt{\alpha}\, z_{\text{cr}} \gg 1$

In this limit, the escape dynamics is completely determined by the presence of the saddle point z_{cr}. We set $z =: x z_{\text{cr}}$ and show that $x = 1$ is the point where π_{esc} establishes its limiting value:

$$\pi_{\text{esc}} \propto \frac{e^{-A^2}}{\text{Erf}\left((x-1)A\right) + \text{Erf}(A)} \overset{A\gg 1}{\approx} \frac{e^{A^2(1-x)^2}(1-x)}{\frac{e^{-A^2}}{\sqrt{\pi}A} - e^{A^2(x-1)^2} e^{A^2}\left(x-1+|x-1|\right)} \tag{5.81}$$

$$= \begin{cases} \frac{e^{-A^2 x(2-x)}(1-x)}{\sqrt{\pi}A} & x<1 \Leftrightarrow z<z_{\text{cr}} \\ \\ \frac{1}{2}e^{-A^2} & x>1 \Leftrightarrow z>z_{\text{cr}} \end{cases} \quad \Rightarrow \quad \boxed{z_{\text{esc}} \approx z_{\text{cr}}} \tag{5.82}$$

This ansatz leads to the approximation of the escape barrier at the critical value z_{cr} itself, just as we expected. The alternative scale z_{split} yields:

$$z_{\text{split}} \approx z_{\text{cr}} + \frac{e^{-\alpha z_{\text{cr}}^2}}{4\alpha z_{\text{cr}}} - \frac{e^{-\alpha(1-z_{\text{cr}})^2}}{4\alpha(1-z_{\text{cr}})} \tag{5.83}$$

which is indeed very close to $z_{\rm cr}$. Thus, in this limit the two previously different values almost coincide, which is consistent with a peaked and symmetric escape barrier distribution.

Estimation of $z_{\rm esc}$ close to criticality: $|A| = \sqrt{\alpha}\,|z_{\rm cr}| < 1$

In this regime, the saddle point $z_{\rm cr}$ just emerges and the drift term in the Fokker-Plank equation changes its sign. It is this regime, where the quadratic part of the drift becomes comparable to the linear part. Of the two estimates of the escape barrier, $z_{\rm split}$ is the easier one to evaluate and expand for $|A| \ll 1$. The following approximation is arrived at by expanding the definition of $\ln z_{\rm split}$ for $A \ll 1$.

$$z_{\rm split} \approx \frac{\kappa}{\sqrt{\alpha}} \exp\left(\frac{2 - e^{\kappa^2}}{2\,\kappa}\sqrt{\alpha}\, z_{\rm cr}\right), \quad \text{with} \quad \kappa := \mathrm{Erf}^{-1}\left(\tfrac{1}{2}\right) \approx 0.476936 \qquad (5.84)$$

5.5.6 Simulation results

It is now time to put the results of the last sections to the test. Before presenting the simulation results, let us recapitulate the assumptions that went into the main result for the fixation rate equations (5.65) and (5.68):

- There is a separation of time scales. The distribution of single mutants is adiabatically coupled to the slower motion of the double mutants.
- The frequency of double mutants remains small at all times to ensure the validity of the linear approximation to their Master equation.
- We saw that a deep fitness valley, i.e. $s_d \gg s_b$ was necessary to realize the first two points.
- The frequency of single mutants could thus be set to their instantaneous mean value. Moreover - due to their low numbers - we could neglect correlations between them.
- Mutation u needs to be small enough to have at least $N\gamma \ll 1$ (this is in fact a weak condition). This made it possible to separate the fixation problem to conditional runs for escape of rarely occurring double mutant "pioneers".

The following plots show measurements of both the mean time to fixation (the condition is $x_{AB} > 80\%$) and the mean escape barrier (defined as the largest unsuccessful AB clone per run). The simulations were carried out using the full four-dimensional Wright-Fisher analog to the Moran model described in the text (see appendix C), i.e. multinomial sampling of the next round generation using

half the size of the original parameters and counting two Wright-Fisher turns as N Moran turns. We show here three measurements for $Nu \ll 1$, $Nu = 1$ and $Nu \gg 1$ to make clear that only $N\gamma \ll 1$ is the important condition on mutation size.
The simulations were performed using the computing infrastructure provided by the regional computing center of the university of Cologne. The cluster architecture allowed to simulate the stochastic evolution trajectories in parallel (under openMP [125]), thus massively accelerating the measurements.

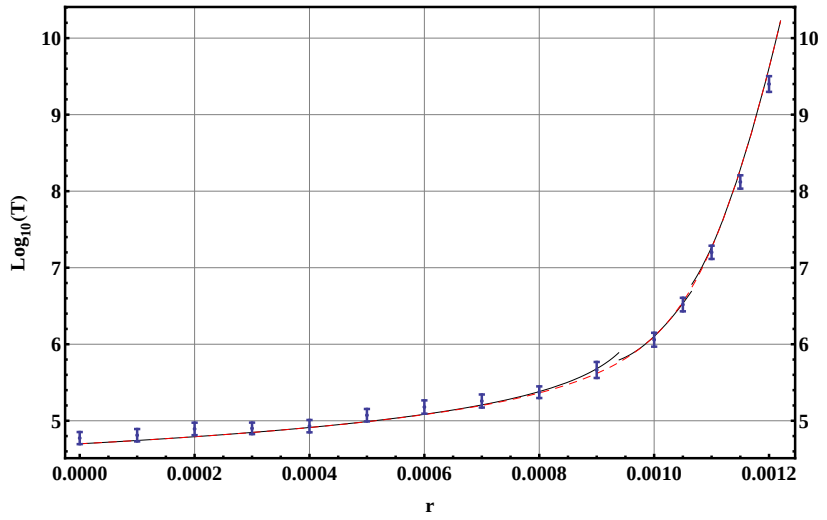

Figure 5.7: Measurement of the mean time to fixation of the *AB* double mutants species, starting from an all wild type population. For each value of recombination, 100 fixation trajectories were measured. The red dotted line shows the full fixation rate according to (5.65), the black line the exponential approximations in the three regimes (5.68). Parameters in this plot: $N = 10^6$, $u = 10^{-5}$, $s_d = 10^{-2}$, $s_b = 10^{-3}$.

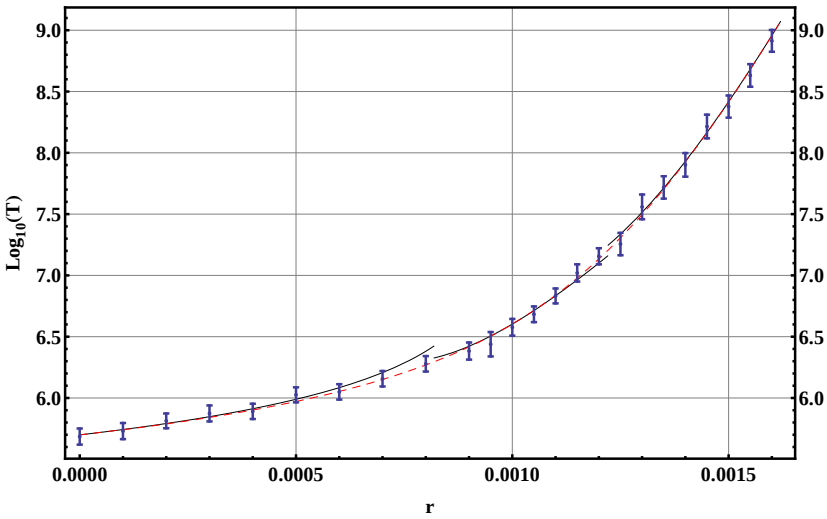

Figure 5.8: Same plot as before, but with the parameters: $N = 10^5$, $u = 10^{-5}$.

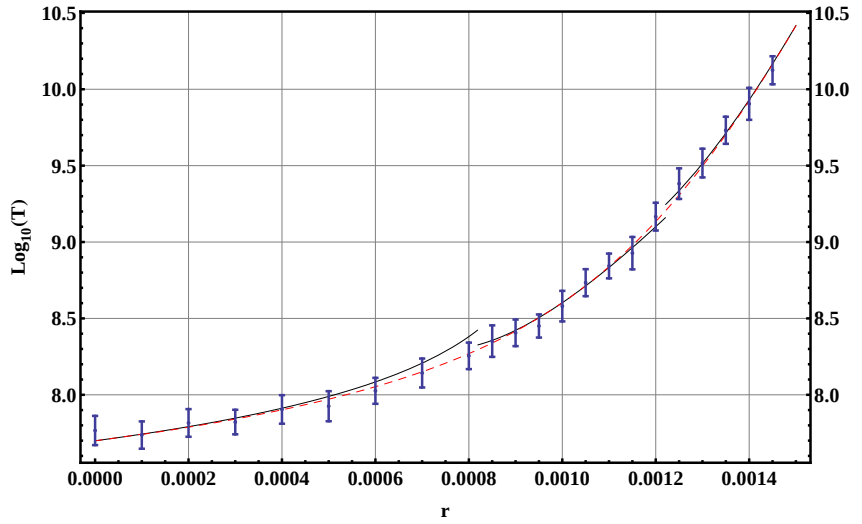

Figure 5.9: Same plot as before, but with the parameters: $N = 10^5$, $u = 10^{-6}$.

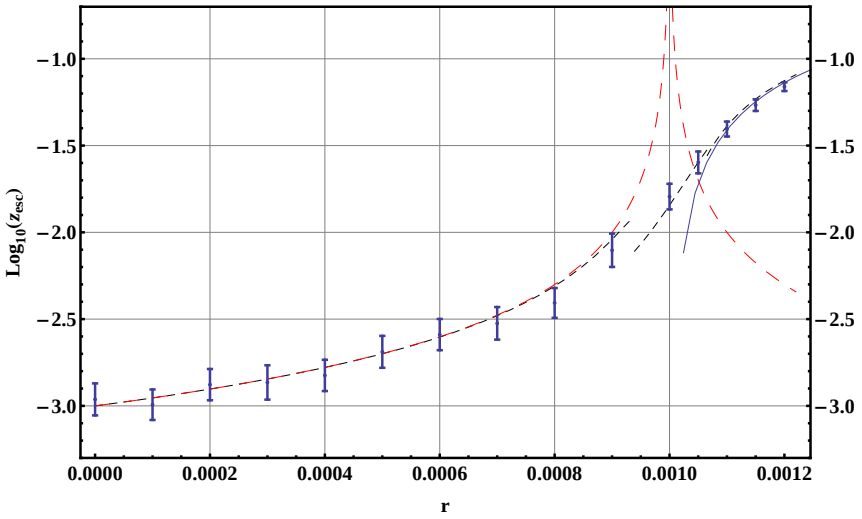

Figure 5.10: This one and the two following plots show the mean escape barrier for the same measurements. The black dotted line shows in the three recombination regimes the approximations that are most suitable, equations (5.76), (5.82) and (5.84). The red dotted line is the size of the neutral zone $z_n \sim \frac{1}{N|s_b - r|}$, whereas the blue line shows the position of the saddle point at z_{cr} (found numerically as fix point to the four dimensional drift term). As before, we have here $Nu = 10$.

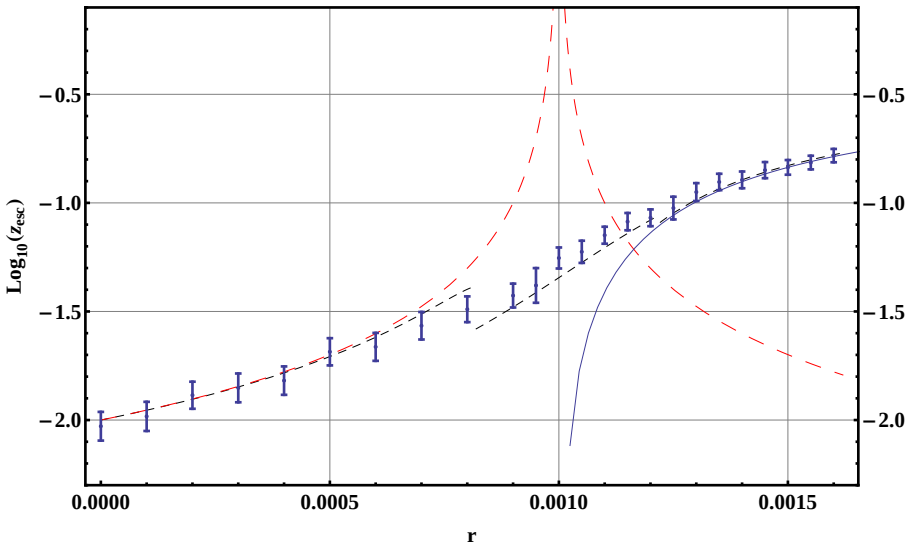

Figure 5.11: Same plot as before, but with $Nu = 1$.

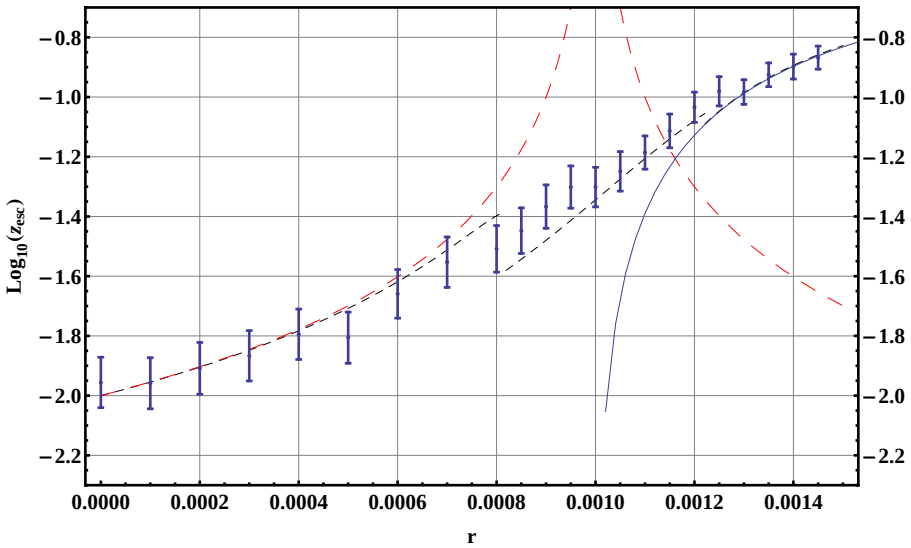

Figure 5.12: Same plot as before, but with $Nu = 0.1$.

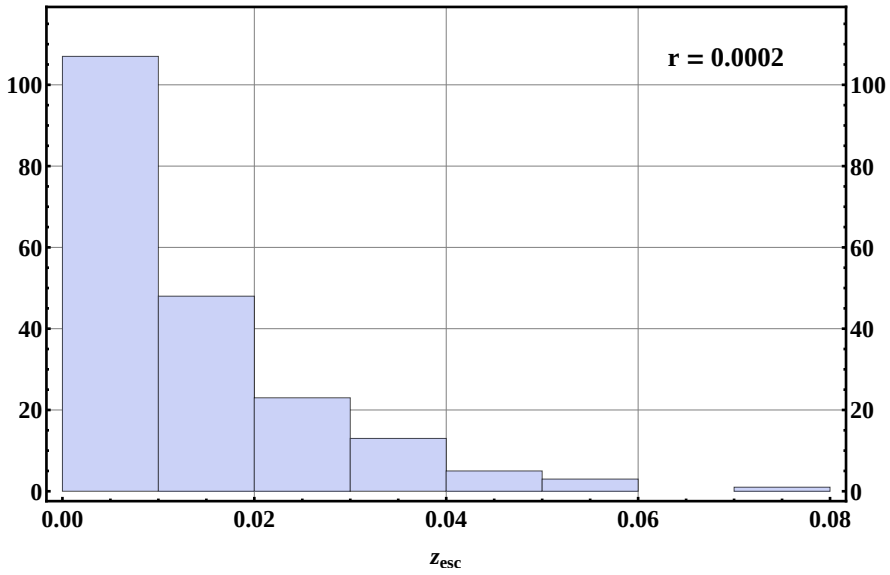

Figure 5.13: Histogram of the largest unsuccessful *AB* clone for a specific value of recombination below criticality ($Nu = 1$). The distribution is very broad, such that a discrepancy between e.g. mean and median is to be expected (see text).

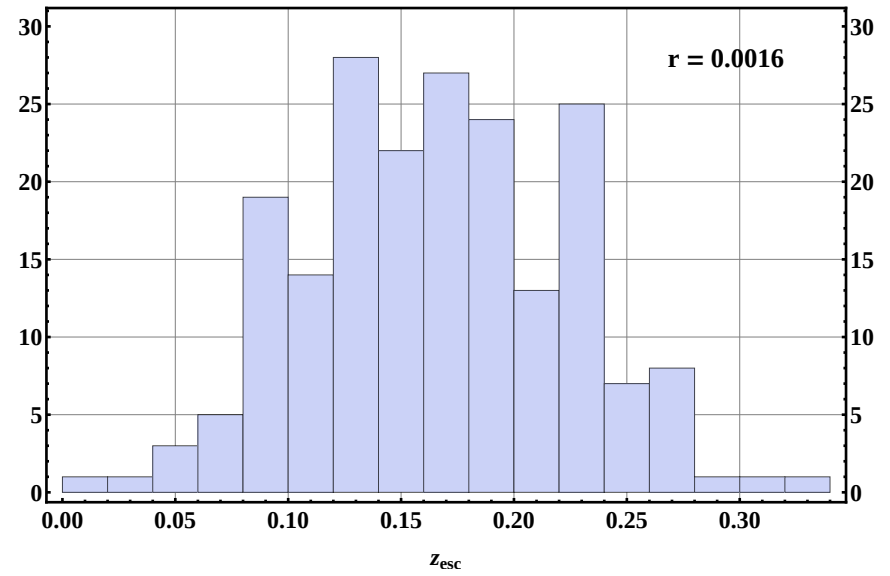

Figure 5.14: Same histogram as before, but now for recombination well above criticality (still with $Nu = 1$). The distribution is much more symmetric around its mean.

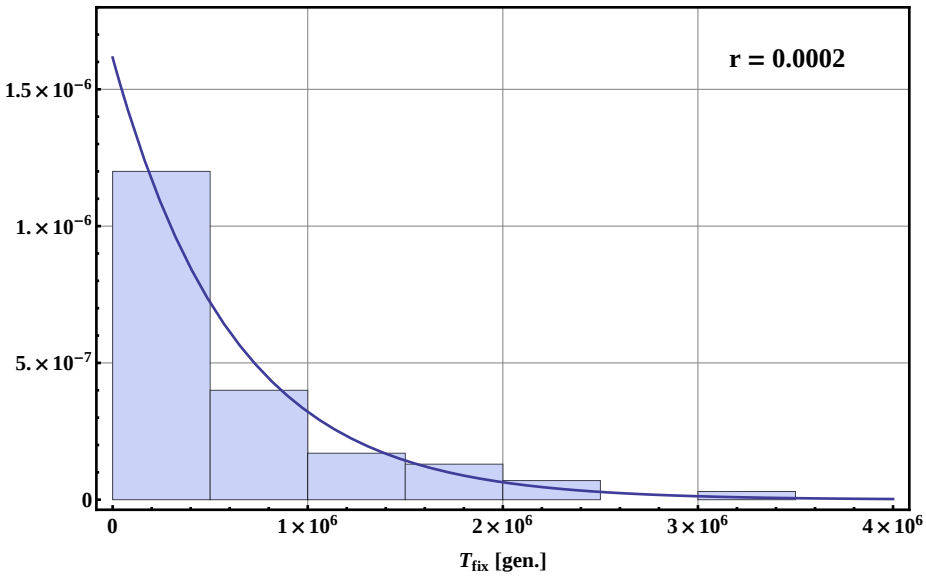

Figure 5.15: Distribution of the fixation time T_{fix} below criticality ($Nu = 1$). The distribution is essentially exponential (solid line), due to fixation being a very rare event.

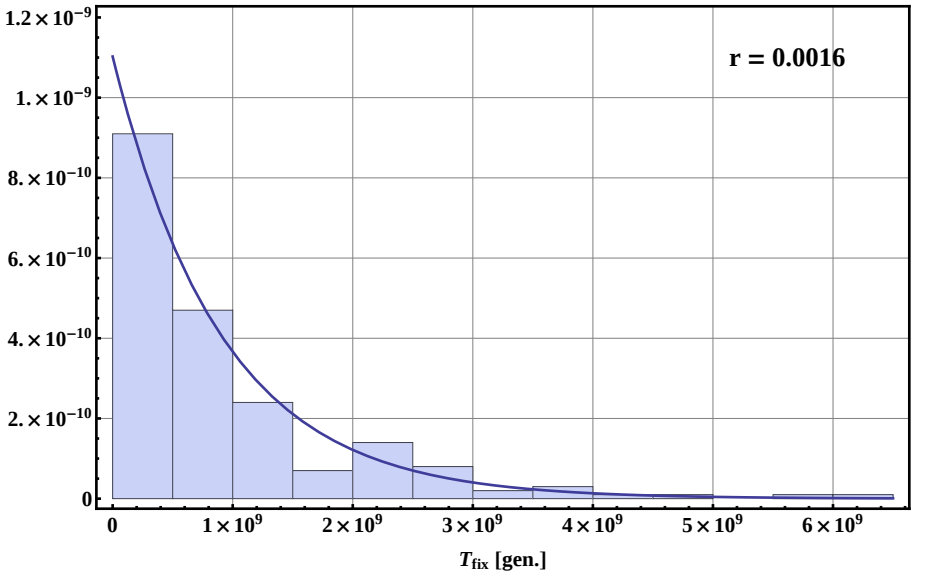

Figure 5.16: Same distribution as before, but now for recombination well above criticality (still with $Nu = 1$). The distribution is still exponential (solid line), fixation even more rare.

5.6 Discussion

Precise theoretical predictions for measurable quantities are of immediate importance for the interpretation of experimental data. Minimal models play a special role in deriving these predictions. these try to capture the essential influences of concepts such as evolutionary forces while still being eligible for analytical treatment. In the first few chapters, we have seen how the understanding of a minimal model of evolution under selection, mutation and drift can help to set up biologically meaningful scales to quantify the functional impact of mutations in cancer. In this chapter, an attempt was made to replicate this consequent analysis for a more extended evolutionary model. Due to the large number of potential qualitative regimes of the model, the focus was here put on the particular situation of adaption via over-compensatory mutations in the presence of recombination. The derivation of relevant time scales was performed along the standard lines of the analysis of stochastic processes: the formulation of a microscopic model, the qualified limit of large population size, the consideration of macroscopic laws and stationary states, the identification of fast and slow modes of motion and elimination of the fast to arrive at effective dynamics of the slow. In the end, the derivation of escape (fixation) rates was conceptually very similar to the treatment of the more basic Kimura model of evolution of genomes with just one locus with two alleles.

There are still some open points left for further research. First, the modeling of the evolution of populations of constant size might not always be the most realistic one. New qualitative effects are to be expected in expanding or contracting populations. It would be interesting to observe, whether recombination maintains its adaption-blocking effect for varying population sizes. In the context of evolutionary cooperation, the effects of non-constant population sizes have recently gained attraction [126]. Especially for the mathematical modeling of cancer evolution - which is of course asexual - a constant population size is certainly not maintainable.

It was found in this work that the treatment of compensatory evolution via adiabatic elimination of fast variables is only consistently possible for deep fitness valleys. In the intermediate and opposite regime, other approximation schemes are needed and have successfully been applied [121]. There may be parameter regimes, where a full high-dimensional treatment of the stochastic fixation dynamics is unavoidable.

Ultimately, the predictions from this analysis should find their way to the practice of interpreting observed genetic data. However, the large number of parameters and the non-local nature of the model make it necessary to identify measurable quantities that can be used to estimate recombination sizes or effects of epistasis.

Chapter 6

Summary

The theme of this work is the concept of minimal models of evolution. By expressing the theory of evolution in a mathematical language, one can find concrete expressions for measurable quantities - such as fixed state probabilities - in terms of few model parameters - such as mutation strength u, selection coefficient s etc. These formulas are of direct utility to explain the variation that is seen in genetic sequence data. One can in particular find the most likely set of parameters, *given* some concrete data. The first part of this thesis demonstrated, how the fitness effects of mutations with respect to a simple evolutionary model can be inferred from protein domain sequence alignments. Driven by the need to estimate parameters also in the light of few data, the methods of Bayesian inference [7] were consequently applied. Augmented with the guiding principle to choose always the most conservative model to incorporate all information at any given point in time - the maximum entropy and minimum discrimination information principles - this analysis resulted in a definite instruction set for the estimation of germline fitness effects. This set includes both a prescription for a consistent choice of pseudo-counts to account for missing data, and for point estimates derived from the Bayesian posterior distribution. It was demonstrated, that the choice of priors - at the beginning of the inference program - is actually closely linked to the way that point estimates are chosen at the end.

This scheme to estimate germline fitness related scores was then applied to mutations found in cancer cells. Conceptually, it is not at all clear, why this scale should be related to the evolutionary process of cancer at all. Both germline evolution and cancer evolution are clearly distinct biological mechanisms. It was therefore of principal interest to test the utility of the germline fitness scoring scheme for the task of finding cancer driver genes. For the studied data set of cancer mutations in protein kinase genes, the analysis found a strong tendency of mutations found in genes with cancer implication to be germline deleterious. This

is especially true for tumor suppressor genes, which are "deactivated" in cancer cells. This result is of immediate importance to find new cancer genes in present and future cancer screening studies.

The Hidden Markov Model (HMM) method [11] described in chapter four used a complementary ansatz to find genes of importance to cancer progression. It is clearly a change of perspective: from the effect a mutation *would* (supposedly) have in germline evolution to the effect it *does* have in cancer evolution. Since there is not yet a mathematical theory of general cancer evolution itself, the signals of selection can be measured only indirectly. Genes, whose altered state is of relevance to the tumor progression, exhibit an increased rate of missense substitutions, higher than the level that would be expected by chance alone. This higher rate can be measured and is a scale for *positive* selection pressures on genes in cancer evolution. The probabilistic nature of HMM allows for a statistically meaningful analysis of mutation data that returns not only the genes most likely to be under cancer-selection, but also provides estimates for the size of selection. Moreover, it was demonstrated how - in the presence of more extended data sets - the HMM would in principle be able to identify specific sequence sites (such as protein kinase sub-domains) that are under positive selection. This could potentially contribute to the understanding of the biological processes that govern cancer progression in more detail. Internally, the HMM presented in this work uses the very same predictions from minimal evolution models to connect the observed data with model parameters such as selection strength.

Mutation, selection and genetic drift are not the only forces of evolution, although these three suffice to set up a first basic evolution model. The last part of this work considered the next level of generalization in the minimal design of evolution: a two-locus/two-allele model including additionally the effects of recombination and epistasis, i.e. the interaction between different loci. This generalization increases not only the degrees of freedom (from two to four competing genotypes) and the number of model parameters and their combinations, but also the number of qualitatively different selection scenarios considerably. Depending on the exact relative scale of all the parameters, the model displays distinct modes of adaption. Arguably the most interesting situation is that of sign-epistatsis [12], also covered in this work, where a first mutation away from the wild type at any of the two loci is strongly deleterious, but is over-compensated by a mutation at the second locus. By construction, this is an example for a strong non-additive interaction between the loci, i.e. epistasis. It is easy to visualize that the crossing of such a fitness valley poses a bottleneck for the adaption of a population to the globally fittest state. In sexually reproducing finite populations, recombination can increase the strength of this barrier up to a size where fixation can only be achieved by rare and

extremely strong fluctuations. This process is often termed as “stochastic tunneling” [15]. A lot of theoretical work describing the different sectors of the model has already been contributed by several authors, covering both the deterministic [14, 13] and the stochastic sector of the theory for neutral [119, 120] and non-neutral compensatory mutations [121, 122, 127]. This work tried to fill a gap in that it exposes the importance of time scale separation for the case of deep fitness valleys. In this regime, it is shown how the high-dimensional dynamics in genotype frequency space can be separated into a slow and a fast mode of adaption. The effective one-dimensional dynamics of super-fit double mutants allows for a derivation of the typical time scales needed to cross the fitness valley by stochastic tunneling. Importantly, this can be done even for recombination strengths beyond the point where deterministic theories predict a divergence of fixation time. Moreover, the critical threshold for the frequency of a growing double mutant subpopulation is derived that needs to be overcome in order to fixate successfully. This “escape barrier” shows a transition from a regime, where initial demographic fluctuations dominate to a regime, where fixation depends on the crossing of a saddle point. The predictions are compared to numerical simulations and their validity within the specified parameter regime is demonstrated. This study will hopefully contribute to the complete characterization of this important next-order minimal evolution model and bring it closer to a practical utility for the analysis of real sequence data.

Appendix A

Bayesian inference: multinomial sampling

A.1 The Dirichlet prior

In this appendix, some of the calculation in the context of inference of multinomial weights are presented. In Bayesian inference, the Dirichlet distribution is the standard distribution to encode prior information for problems of multinomial sampling [47].

$$\mathrm{Dir}(\boldsymbol{\theta}\,|\,\boldsymbol{\alpha}) := \frac{\Gamma(A)}{\prod_{i=1}^{n}\Gamma(\alpha_i)}\prod_{i=1}^{n}\theta^{\alpha_i-1} =: \frac{1}{\mathrm{Beta}(\boldsymbol{\alpha})}\prod_{i=1}^{n}\theta^{\alpha_i-1} \tag{A.1}$$

$$\boldsymbol{\alpha}\in\mathbb{R}^n_{\geq 0},\quad A := \sum_{i=1}^{n}\alpha_i > 0,\quad \sum_{i=1}^{n}\theta_i = 1 \tag{A.2}$$

where in the first line, the beta function $\mathrm{Beta}(\boldsymbol{\alpha})$ is defined. The mean and logarithmic mean values are given by

$$\forall\, i = 1\ldots n:\quad \langle\theta_i\rangle = \frac{\alpha_i}{A},\quad \langle\ln\theta_i\rangle = \psi(\alpha_i) - \psi(A) \tag{A.3}$$

where $\psi(x) := \frac{d}{dx}\ln\Gamma(x)$ is the digamma function. The reason for this choice of prior is that the Dirichlet distribution is conjugate to the multinomial in the sense that the posterior distribution for a sample $\{k_i\}_{i=1\ldots n}$ is again of Dirichlet type. The region of integration is always the simplex .

$$\Delta^{n-1} := \left\{\boldsymbol{\theta}\in\mathbb{R}^n \,\middle|\, \sum_{i=1}^{n}\theta_i = 1, \forall_i:\ \theta_i \geq 1\right\} \tag{A.4}$$

To compute the posterior distribution for a particular sample $\boldsymbol{k}$ drawn from the multinomial model, we set:

$$P(\theta\,|\,\boldsymbol{k}, I_0) = \frac{P(\boldsymbol{k}|\,\theta)\,P(\boldsymbol{\theta}\,|\,I_0)}{P(\boldsymbol{k}|\,I_0)}, \quad P(\boldsymbol{\theta}\,|\,I_0) = \mathrm{Dir}(\boldsymbol{\theta}\,|\,\boldsymbol{\alpha}), \quad \sum_{i=1}^{n} k_i =: K \tag{A.5}$$

The normalization factor $P(\boldsymbol{k}|\,I_0)$ is computed using the normalization of the Dirichlet distribution for arbitrary parameters:

$$P(\boldsymbol{k}|\,I_0) = \int_{\Delta^{n-1}} \mathrm{d}^n\theta\; P(\boldsymbol{k}|\,\boldsymbol{\theta})\,P(\boldsymbol{\theta}\,|\,I_0) \tag{A.6}$$

$$= \int_{\Delta^{n-1}} \mathrm{d}^n\theta\; \frac{K!\,\Gamma(A)}{\prod_{i=1}^{n}(k_i!\,\Gamma(\alpha_i))} \prod_{i=1}^{n} \theta_i^{k_i+\alpha_i-1} \tag{A.7}$$

$$= \frac{\Gamma(A)\,K!}{\Gamma(A+K)} \prod_{i=1}^{n} \frac{\Gamma(\alpha_i+k_i)}{\Gamma(\alpha_i)\,k_i!} \tag{A.8}$$

Inserting this factor and the definitions immediately leads to the result

$$P(\theta\,|\,\boldsymbol{k}, I_0) = \mathrm{Dir}(\boldsymbol{\theta}\,|\,\boldsymbol{\alpha}+\boldsymbol{k}). \tag{A.9}$$

A.2 Kullback-Leibler divergences

The Kullback-Leibler divergence [52] is a non-symmetric, positive definite measure for the difference between two probability distributions $p(x)\,dx$ and $q(x)\,dx$. It is defined as:

$$D_{\mathrm{KL}}(p\,|\,q) := \int \mathrm{d}x\; p(x) \ln\frac{p(x)}{q(x)} \geq 0, \quad D_{\mathrm{KL}}(p\,|\,q) = 0 \;\Leftrightarrow\; p = q \tag{A.10}$$

The Kullback-Leibler divergence between a Dirichlet distribution and a uniform distribution on the simplex Δ^{n-1} is given by

$$D_{\mathrm{KL}}(\mathrm{Dir}(\boldsymbol{\theta}\,|\,\alpha)\,|\,\chi_{\Delta^{n-1}}(\boldsymbol{\theta})) = \int_{\Delta^{n-1}} \mathrm{d}^n\theta\; \mathrm{Dir}(\boldsymbol{\theta}\,|\,\boldsymbol{\alpha})\; \ln\frac{\mathrm{Dir}(\boldsymbol{\theta}\,|\,\boldsymbol{\alpha})}{\Gamma(n)} \tag{A.11}$$

$$= -\ln\mathrm{Beta}(\boldsymbol{\alpha}) - \ln\Gamma(n) + \sum_{i=1}^{n}(\alpha_i-1)\,\langle\ln\theta_i\rangle \tag{A.12}$$

$$= -\ln\mathrm{Beta}(\boldsymbol{\alpha}) - \ln\Gamma(n) + \sum_{i=1}^{n}(\alpha_i-1)\,(\psi(\alpha_i)-\psi(A)) \tag{A.13}$$

$$= -\ln\mathrm{Beta}(\boldsymbol{\alpha}) - \ln\Gamma(n) + \sum_{i=1}^{n}(\alpha_i-1)\,\psi(\alpha_i) - (A-n)\,\psi(A) \tag{A.14}$$

$$= (n-A)\,\psi(A) + \ln\frac{\Gamma(A)}{\Gamma(n)} - \sum_{i=1}^{n}[\ln\Gamma(\alpha_i) - (\alpha_i-1)\,\psi(\alpha_i)] \tag{A.15}$$

The Kullback-Leibler divergence between two different Dirichlet distributions is analogously calculated to be

$$D_{\mathrm{KL}}\left(\mathrm{Dir}(\boldsymbol{\theta}\,|\,\boldsymbol{\alpha})\,|\mathrm{Dir}(\boldsymbol{\theta}\,|\,\boldsymbol{\beta})\right) = \int_{\Delta^{n-1}} \mathrm{d}^n\theta\ \mathrm{Dir}(\boldsymbol{\theta}\,|\,\boldsymbol{\alpha})\ \ln\frac{\mathrm{Dir}(\boldsymbol{\theta}\,|\,\boldsymbol{\alpha})}{\mathrm{Dir}(\boldsymbol{\theta}\,|\,\boldsymbol{\beta})} \tag{A.16}$$

$$= \ln\frac{\mathrm{Beta}(\boldsymbol{\beta})}{\mathrm{Beta}(\boldsymbol{\alpha})} + \sum_{i=1}^{n}(\alpha_i-\beta_i)\langle\ln\theta_i\rangle = \ln\frac{\mathrm{Beta}(\boldsymbol{\beta})}{\mathrm{Beta}(\boldsymbol{\alpha})} + \sum_{i=1}^{n}(\alpha_i-\beta_i)(\psi(\alpha_i)-\psi(A)) \tag{A.17}$$

$$= \ln\frac{\mathrm{Beta}(\boldsymbol{\beta})}{\mathrm{Beta}(\boldsymbol{\alpha})} - (A-B)\psi(A) + \sum_{i=1}^{n}(\alpha_i-\beta_i)\psi(\alpha_i), \quad A=\sum_i\alpha_i,\ B=\sum_i\beta_i \tag{A.18}$$

A.3 Loss functions and minimal loss

Loss functions $L(\boldsymbol{\theta}',\boldsymbol{\theta})$ are auxiliary and sometimes subjective quantities that are supposed to provide guidance in the decision step of parameter estimation problems [7], i.e. which point estimate $\hat{\boldsymbol{\theta}}$ to choose for the unknown parameter $\boldsymbol{\theta}$ in the face of new information. Given a posterior distribution, the mean loss is generally defined as (here for multinomial sampling)

$$\langle L\rangle(\boldsymbol{\theta}) = \int_{\Delta^{n-1}} \mathrm{d}^n\theta\ L(\boldsymbol{\theta},\boldsymbol{\theta}')\ P\left(\boldsymbol{\theta}'\,\middle|\,\boldsymbol{k}, I_0\right) \tag{A.19}$$

The guidance consists of the prescription to minimize this mean loss in the estimate under the constraints posed by $(\boldsymbol{k}, I_0)$:

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}} \langle L\rangle(\boldsymbol{\theta}) \tag{A.20}$$

Possible loss functions are square and absolute distance to the true value [7] or even a delta function.

$$\text{e.g.}\quad L_0(\boldsymbol{\theta}',\boldsymbol{\theta}) = \delta(\boldsymbol{\theta}-\boldsymbol{\theta}') \tag{A.21}$$

$$\text{or}\quad L_1(\boldsymbol{\theta},\boldsymbol{\theta}') = \sum_{i=1}^{n}|\theta_i-\theta_i'| \tag{A.22}$$

$$\text{or}\quad L_2(\boldsymbol{\theta}',\boldsymbol{\theta}) = \sum_{i=1}^{n}(\theta_i-\theta_i')^2 \tag{A.23}$$

The loss function used in this work is the discrimination information [46, 51]:

$$L_{\mathrm{DI}}(\boldsymbol{\theta},\boldsymbol{\theta}') := D_{\mathrm{KL}}\left(P(\boldsymbol{k}|\,\boldsymbol{\theta})\,\middle|P\left(\boldsymbol{k}|\,\boldsymbol{\theta}'\right)\right), \quad P(\boldsymbol{k}|\,\boldsymbol{\theta}) = \frac{K!}{\prod_{i=1}^{n}k_i!}\prod_{i=1}^{n}\theta_i^{k_i} \tag{A.24}$$

It is easily evaluated by inserting all the definitions:

$$L_{\mathrm{DI}}(\boldsymbol{\theta},\boldsymbol{\theta}') = D_{\mathrm{KL}}\left(P(\boldsymbol{k}|\,\boldsymbol{\theta})\,\middle|\,P\left(\boldsymbol{k}|\,\boldsymbol{\theta}'\right)\right) = \sum_{\boldsymbol{k}} P(\boldsymbol{k}|\,\boldsymbol{\theta})\,\ln\frac{P(\boldsymbol{k}|\,\boldsymbol{\theta})}{P\left(\boldsymbol{k}|\,\boldsymbol{\theta}'\right)} \tag{A.25}$$

$$= \sum_{\boldsymbol{k}} P(\boldsymbol{k}|\,\boldsymbol{\theta})\,\sum_{i=1}^{n} k_i\,\ln\frac{\theta_i}{\theta_i'} = \sum_{i=1}^{n}\langle k_i\rangle_{\boldsymbol{\theta}}\ln\frac{\theta_i}{\theta_i'} = K\sum_{i=1}^{n}\theta_i\ln\frac{\theta_i}{\theta_i'} \tag{A.26}$$

The idea of MinDI is to choose the one $\hat{\boldsymbol{\theta}}$ for which the sampling distribution is "closest" to the true one, i.e. for which *further* sampling would give the least information gain. Since the size of this putative subsequent sample is irrelevant for this question (K is a prefactor above), we can set $K = 1$. Of course we do not know the true value $\boldsymbol{\theta}$ and thus the true sampling distribution. Therefore, the loss function above is averaged over the posterior distribution, which includes all information we have at this point.

$$\langle L_{\mathrm{DI}}\rangle(\boldsymbol{\theta}) = \int_{\Delta^{n-1}} \mathrm{d}^{n-1}\theta'\,\sum_{i=1}^{n}\theta_i\ln\frac{\theta_i}{\theta_i'}\,P\left(\boldsymbol{\theta}'\,\middle|\,\boldsymbol{k}, I_0\right) \tag{A.27}$$

$$= \sum_{i=1}^{n}\theta_i\left(\ln\theta_i - \langle\ln\theta_i'\rangle\right) = \sum_{i=1}^{n}\theta_i\left(\ln\theta_i - \psi(\alpha_i + k_i) + \psi(A+K)\right) \tag{A.28}$$

To find the minimum of $\langle L_{\mathrm{DI}}\rangle$ w.r.t. $\boldsymbol{\theta}$ on the simplex Δ^{n-1}, we employ Lagrange multipliers to guarantee the normalization:

$$0 \stackrel{!}{=} \partial_{\theta_i}\left(\langle L_{\mathrm{DI}}\rangle(\boldsymbol{\theta}) + \lambda\left(1 - \sum_{i=1}^{n}\theta_i\right)\right) \tag{A.29}$$

$$= \ln\theta_i - \psi(\alpha_i + k_i) + \psi(A+K) + 1 - \lambda \tag{A.30}$$

$$\Rightarrow\ \theta_i = \exp\left(\psi(\alpha_i + k_i) - \psi(A+K) - 1 + \lambda\right) \tag{A.31}$$

Using the normalization condition, this leads to the final result that

$$\boxed{\hat{\theta}_i = \frac{e^{\psi(\alpha_i + k_i)}}{\sum_{j=1}^{n} e^{\psi(\alpha_j + k_j)}}} \tag{A.32}$$

Reinserting this value in the mean loss gives a scale for the quality of this estimate.

$$\langle L_{\mathrm{DI}}\rangle\left(\hat{\boldsymbol{\theta}}\right) = \sum_{i=1}^{n}\frac{e^{\psi(\alpha_i + k_i)}}{\sum_{j=1}^{n} e^{\psi(\alpha_j + k_j)}}\left(\psi(\alpha_i + k_i) - \ln\sum_{j=1}^{n} e^{\psi(\alpha_j + k_j)} - \psi(\alpha_i + k_i) + \psi(A+K)\right)$$

$$= \psi(A+K) - \ln\left(\sum_{j=1}^{n} e^{\psi(\alpha_j + k_j)}\right) \tag{A.33}$$

Appendix B

The Viterbi algorithm

For completeness only, we here give a description of the Viterbi algorithm to find the globally most likely state sequence for a HMM [11].

In section 4.2.2 on the forward/backward algorithm, we found the best current state sojourn probability $P(q_t = S_i | O, \theta)$. One could be tempted to set the most likely state sequence Q^* simply by choosing the local state of maximum probability, i.e.

$$\forall t = 1 \ldots T: \quad q_t^* = \underset{i}{\operatorname{argmax}} P(q_t = S_i | O, \theta) \tag{B.1}$$

But there is a conceptual problem here: consider a set-up where not all transitions between hidden states are possible. The above locally best path will in general return a path that is *impossible* to realize. To actually find the most likely state sequence in the space of all *allowed* paths, the following auxiliary quantity is defined:

$$\delta_t(i) := \max_{q_1 \cdots q_{t-1}} P(q_1, q_2, \ldots, q_t = S_i, O_1, \ldots, O_t \mid \theta) \tag{B.2}$$

This is the most likely state sequence ending in the state i at time t and with partial observation $O_1 \ldots O_t$. The following Viterbi algorithm finds iteratively maximum likelihood state sequences for *all* end states simultaneously. Once the end point at $t = T$ is reached, the one global maximum likelihood sequence is singled out by backtracking. The (internal) variable ψ is used to memorize all the intermediate best paths.

Viterbi Algorithm

1. $\delta_1(i) = \pi_i b_i(O_1), \quad \psi_1(i) = 0, \quad \forall i = 1, \ldots, N$

2. $$\left.\begin{aligned} \delta_{t+1}(j) &= \left[\max_{1 \le i \le N} \left(\delta_t(i) a_{ij}\right)\right] b_j(O_{t+1}) \\ \psi_{t+1}(j) &= \operatorname*{argmax}_{1 \le i \le N} \left[\delta_t(i) a_{ij}\right] \end{aligned}\right\} \forall i = 1, \ldots, N, \quad \forall t = 1, \ldots, T-1$$

3. $P^* := P(Q^*, O \mid \theta) = \max_{1 \le i \le N} \delta_T(i), \quad q_T^* = \operatorname*{argmax}_{1 \le i \le N} \delta_T(i)$

4. $q_t^* = \psi_{t+1}(q_{t+1}^*), \ \forall t = T-1, \ldots, 1$

The state sequence q^* is the most likely path of the system given the observation and a particular set of parameters. In implementations of this algorithm, one needs to work with the logarithm of all these very small quantities to avoid underflow errors.

Appendix C

Wright-Fisher vs. Moran models

C.1 Expansion of the Wright-Fisher Master equation

One of the standard microscopic evolutionary models under constant population size N is the generation based Wright-Fisher process. It is a discrete Markov process and the transition probabilities (per unit time) for a one dimensional Wright-Fisher Master equation are binomial.

$$\partial_t P(n,t) = \sum_{n'} W(n \mid n') P(n',t) - W(n' \mid n) P(n,t), \quad n = 1 \ldots N \tag{C.1}$$

$$W(n' \mid n) := \binom{N}{n'} \left(x + \frac{1}{N} f(x)\right)^{n'} \left(1 - x - \frac{1}{N} f(x)\right)^{N-n'}, \quad x := \frac{n}{N} \tag{C.2}$$

The function $f(x)$ encapsulates the evolutionary forces present. It is assumed to be of order 1 (or larger, but not of order N). This form anticipates that only the scaled versions of the different evolutionary parameters (describing mutation u, selection s, recombination r etc.) are relevant for the characteristics of the model: $Nu = \mu$, $Ns = \sigma$ and $Nr = \rho$ etc. Then $f(x)$ will be the leading order correction to the *neutral* evolution process. What we want to achieve here is a controlled large population size limit. The standard routine is van Kampen's system size expansion [40]. But there is no way to express the above rates in the required *canonical form* [40] (i.e. an expansion in orders of $1/N$, where each order is then only a function of the initial state - here x - and the jump length - here $n' - n$). But we can still employ the Kramers-Moyal expansion [40, 128, 129] and look at it

order by order to get a controlled expansion.

$$\partial_t P(x,t) = \sum_{l=1}^{\infty} \frac{(-1)^l}{l!} \partial_x^{(l)} \left[a_l(x) P(x,t) \right], \quad a_l(x) := \sum_{\Delta x} \Delta x^l \, W(\Delta x; x), \quad x = \frac{n}{N} \tag{C.3}$$

The transition rates $W(n' \mid n)$ are here expressed in terms of origin $x = n/N$ and jump size $\Delta x = x' - x$. To get the scaling of the jump moments $a_l(x)$ with $1/N$, we look at the moment generating function for the overall transition rate $W(\Delta x; x)$ ($\varepsilon := 1/N$)

$$M(s) := \sum_{\Delta x} e^{s\Delta x} \, W(\Delta x; x) = e^{s(1-x)} \left[x + \varepsilon f + e^{-s\varepsilon} (1 - x - \varepsilon f) \right]^{1/\varepsilon} \tag{C.4}$$

$$= e^{s(1-x)} \exp\left[1/\varepsilon \log\left(x + \varepsilon f + e^{-s\varepsilon} (1 - x - \varepsilon f) \right) \right]$$

$$= \exp\left[\varepsilon \left(f s + \frac{1}{2} x(1-x) s^2 \right) + \mathcal{O}\left(\varepsilon^2\right) \right] \tag{C.5}$$

$$a_1(x) = \partial_s^{(1)} \Big|_{s=0} M(s) = \varepsilon \, f(x) + \mathcal{O}\left(\varepsilon^2\right) \tag{C.6}$$

$$a_2(x) = \partial_s^{(2)} \Big|_{s=0} M(s) = \varepsilon \, x(1-x) + \mathcal{O}\left(\varepsilon^2\right) \tag{C.7}$$

$$a_{l \geq 3}(x) = \partial_s^{(l)} \Big|_{s=0} M(s) = \mathcal{O}\left(\varepsilon^2\right) \tag{C.8}$$

With this scaling, the lowest order in $\varepsilon = 1/N$ yields the typical diffusion approximation (note that time is now most appropriately measured in N generations)

$$N \partial_t P(x,t) = \left(-\partial_x f(x) + \frac{1}{2} \partial_x^2 x(1-x) \right) P(x,t) \tag{C.9}$$

Likewise, for a higher dimensional process with k-nomial sampling we get the equivalent result (using $\sum_i x_i = 1$, $\sum_i f_i = 0$ to ensure constant population size):

$$W(\boldsymbol{n}' \mid \boldsymbol{n}) = N! \prod_{i=1}^{k} \frac{(x_i + \varepsilon f_i)^{n_i'}}{n_i'!} \tag{C.10}$$

$$M(\boldsymbol{s}) = 1 + \frac{\varepsilon}{2} \left(\sum_i \left(s_i^2 x_i + 2 s_i f_i \right) - \sum_{i,j} s_i s_j x_i x_j \right) + \mathcal{O}\left(\varepsilon^2\right) \tag{C.11}$$

$$N \, \partial_t P(\boldsymbol{x}, t) = \left[-\partial_i f_i(\boldsymbol{x}) + \frac{1}{2} \partial_i \partial_j \left[x_i \delta_{ij} - x_i x_j \right] \right] P(\boldsymbol{x}, t) \tag{C.12}$$

The particular form of the force $f(x)$ is left undetermined. Later, it will be compared to the outcome of the system size expansion of the Moran process. Keep in mind, that we are interested in the WF process for a very pragmatic reason, i.e. in order to simulate the evolution process as fast as possible. In a concrete situation, we will have to choose the force $f(x)$ according to the Moran model at hand. This connection will be the aim of the next section.

C.2 Expansion of the Moran Master equation

In this section, the systematic system size expansion for very general Moran processes will be derived in arbitrary dimension. Importantly, the relationship to the WF model above will be found. The macroscopic state variable is denoted by $\boldsymbol{n} \in \mathbb{N}^k$ with $\sum_i n_i = N$, and the intensive rescaled variable $\boldsymbol{x} = \boldsymbol{n}/N =: \varepsilon \boldsymbol{n}$. From the state $\boldsymbol{n}$, there are in a single birth-death event $k(k-1)$ possible transitions, whose rates are denoted by

$$W_{ij}(\boldsymbol{n}' \mid \boldsymbol{n}) = W_{ij}(\Delta \boldsymbol{n}; \boldsymbol{x}) = W_{ij}(\boldsymbol{x}) \propto \delta_{n_i', n_i+1}\, \delta_{n_j', n_j-1} \left(1 - \delta_{ij}\right)$$

The Moran model Master equation for this process on a simplex is given by:

$$\partial_t P(\boldsymbol{x}, t) = \sum_{i \neq j} \left(\mathbb{E}_i^- \mathbb{E}_j^+ - 1\right) W_{ij}(\boldsymbol{x})\; P(\boldsymbol{x}, t) \tag{C.13}$$

The canonical form [40] of the transition rates assumes the form:

$$W_{ij}(\boldsymbol{x}) =: f(N) \left[\Phi_{ij}^{(0)}(\boldsymbol{x}) + \varepsilon\, \Phi_{ij}^{(1)}(\boldsymbol{x}) + \mathcal{O}\left(\varepsilon^2\right)\right] \tag{C.14}$$

with a global prefactor $f(N)$. The basic ansatz of the system size expansion is to set $\boldsymbol{n} = N\boldsymbol{\phi}(t) + \sqrt{N}\boldsymbol{\xi}_t$ and then to look at the stochastic process $\boldsymbol{\xi}_t$ with its distribution $\Pi(\boldsymbol{\xi}, t)$. The partial derivatives are evaluated as:

$$\frac{\partial \Pi}{\partial \xi_i} = \sqrt{N} \frac{\partial P}{\partial n_i} \tag{C.15}$$

$$\frac{\partial^2 \Pi}{\partial \xi_i \partial \xi_j} = \left(\sqrt{N}\right)^2 \frac{\partial^2 P}{\partial n_i \partial n_j} \tag{C.16}$$

$$\frac{\partial P}{\partial t} = \frac{\partial \Pi}{\partial t} - \sqrt{N}\, \dot{\phi}_i \frac{\partial \Pi}{\partial \xi_i} \tag{C.17}$$

Also the shift operators can be expanded in ε:

$$\mathbb{E}_i^{\pm} = \exp\left(\pm \partial_{n_i}\right) = 1 \pm \sqrt{\varepsilon} \frac{\partial}{\partial \xi_i} + \frac{1}{2} \varepsilon \frac{\partial^2}{\partial \xi_i^2} + \mathcal{O}\left(\varepsilon^{3/2}\right) \tag{C.18}$$

And finally the above ansatz must be inserted in the leading order term of the transition rate in canonical form:

$$\Phi_{ij}^{(0)}(\boldsymbol{x}) = \Phi_{ij}^{(0)}\left(\boldsymbol{\phi}(t) + \sqrt{\varepsilon}\boldsymbol{\xi}\right) = \Phi_{ij}^{(0)}\left(\boldsymbol{\phi}(t)\right) + \sqrt{\varepsilon}\, \xi_k \frac{\partial \Phi_{ij}^{(0)}}{\partial \xi_k} + \mathcal{O}(\varepsilon) \tag{C.19}$$

All these bits are now inserted into the original Master equation (C.13) and ordered by their order in N (or ε) ($\partial_i = \partial/\partial\xi_i$).

$$\partial_t \Pi - \sqrt{N} \sum_i \dot{\phi}_i(t)\, \partial_i \Pi \tag{C.20}$$

$$= f(N) \sum_{k\neq l} \left[\left(1 - \frac{1}{\sqrt{N}}\partial_k + \frac{1}{2N}\partial_k^2 \right) \left(1 + \frac{1}{\sqrt{N}}\partial_l + \frac{1}{2N}\partial_l^2 \right) - 1 \right] \tag{C.21}$$

$$\left[\Phi_{kl}^{(0)}(\boldsymbol{\phi}(t)) + \frac{1}{\sqrt{N}} \xi_j\, \partial_j \Phi_{kl}^{(0)}(\boldsymbol{\phi}(t)) + \frac{1}{N} \Phi_{kl}^{(1)}(\boldsymbol{\phi}(t)) + \mathcal{O}\left(N^{-3/2}\right) \right] \Pi$$

In the appropriate time scale and to lowest order in $\sqrt{N}$ this yields with

$$\tau := \frac{f(N)\, t}{N} \tag{C.22}$$

$$\mathcal{O}\left(N^{-1/2}\right): \quad \sum_i \partial_\tau \phi_i\, \partial_i \Pi \stackrel{!}{=} \sum_{i,j} \left(\Phi_{ij}^{(0)}(\boldsymbol{\phi}(\tau)) - \Phi_{ji}^{(0)}(\boldsymbol{\phi}(\tau)) \right) \partial_i \Pi \tag{C.23}$$

This equation is satisfied trivially, if we chose for the macroscopic part $\boldsymbol{\phi}(\tau)$

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\phi_i = \sum_j \Phi_{ij}^{(0)}(\boldsymbol{\phi}(\tau)) - \Phi_{ji}^{(0)}(\boldsymbol{\phi}(\tau)) = \left[\alpha_{1,0}(\boldsymbol{\phi}(\tau)) \right]_i \tag{C.24}$$

which is the macroscopic law [40]. Remember that the generalized jump moments were given by ($\lambda = 0, 1, 2, \ldots$)

$$\left[\alpha_{1,\lambda}(\boldsymbol{x}) \right]_i := \sum_{\Delta\boldsymbol{n}} \Delta n_i\, \Phi^{(\lambda)}(\Delta\boldsymbol{n}; \boldsymbol{x}) = \sum_j \Phi_{ij}^{(\lambda)}(\boldsymbol{x}) - \Phi_{ji}^{(\lambda)}(\boldsymbol{x}) \tag{C.25}$$

$$\left[\alpha_{2,\lambda}(\boldsymbol{x}) \right]_{ij} := \sum_{\Delta\boldsymbol{n}} \Delta n_i \Delta n_j\, \Phi^{(\lambda)}(\Delta\boldsymbol{n}; \boldsymbol{x}) \tag{C.26}$$

$$= \delta_{i,j} \sum_k \left(\Phi_{ik}^{(\lambda)}(\boldsymbol{x}) + \Phi_{ki}^{(\lambda)}(\boldsymbol{x}) \right) - (1 - \delta_{ij}) \left(\Phi_{ij}^{(\lambda)}(\boldsymbol{x}) + \Phi_{ji}^{(\lambda)}(\boldsymbol{x}) \right) \tag{C.27}$$

We are ready to look at the next order in N and recover the linear noise approximation (LNA):

$$\partial_\tau \Pi(\xi, \tau) = \sum_{k\neq l} \left\{ \sum_j \partial_j \Phi_{kl}^{(0)}(\boldsymbol{\phi}) \left(\partial_l - \partial_k \right) \xi_j + \frac{1}{2} \Phi_{kl}^{(0)}(\boldsymbol{\phi}) \left(\partial_k^2 + \partial_l^2 - 2\partial_l \partial_k \right) \right\} \Pi(\xi, \tau)$$

$$= \left\{ - \sum_{i,j} \left[\partial_i \left(\alpha_{1,0}(\boldsymbol{\phi}) \right)_j \right] \partial_j \circ \xi_j + \frac{1}{2} \sum_{i,j} \left(\alpha_{2,0}(\boldsymbol{\phi}) \right)_{ij} \partial_i \partial_j \right\} \Pi(\xi, \tau) \tag{C.28}$$

This is the proper expansion, as long as $\alpha_{1,0}$ is not exactly zero [40]. In that case, we are faced with a genuine diffusion process [40]. And in the Moran model of evolution, this is mostly the limit one is interested in. We now express the transition rates W_{ij} in the Master equation (C.13) in terms of birth rates B_i and death rates D_i. The concrete form of the birth and death rates is left unspecified. In any case, one is interested in a limit $N \to \infty$ of the process, where the scaled parameters describing selection $Ns =: \sigma$ and mutation $Nu =: \mu$ etc. are held constant. The leading order term of an expansion of the transition rates always describes *neutral* evolution (e.g. with $u = 0,\ s = 0$) and the next order includes all other forces. This leads to the following canonical form.

$$W_{ij}(\boldsymbol{x}) = B_i(\boldsymbol{x})\, D_j(\boldsymbol{x}), \quad \sum_i B_i = \sum_i D_i = \sum_i x_i = 1 \tag{C.29}$$

$$B_i(\boldsymbol{x}) =: x_i + \frac{1}{N} b_i(\boldsymbol{x}) + \mathcal{O}\left(\frac{1}{N^2}\right), \quad b_i(\boldsymbol{x}) = \mathcal{O}(\mu, \sigma, \text{etc.}), \quad d_i(\boldsymbol{x}) \text{ ditto} \tag{C.30}$$

This yields for the standard first and second jump moments:

$$\left[a_1(\boldsymbol{x})\right]_i := \sum_{\Delta \boldsymbol{n}} \Delta n_i\, W(\Delta \boldsymbol{n}; \boldsymbol{x}) = \sum_{j, j \neq i} \left(B_i D_j - D_i B_j\right) = B_i(\boldsymbol{x}) - D_i(\boldsymbol{x}) \tag{C.31}$$

$$= \frac{1}{N}\left(b_i(\boldsymbol{x}) - d_i(\boldsymbol{x})\right) + \mathcal{O}\left(\frac{1}{N^2}\right) \tag{C.32}$$

$$\stackrel{!}{=} f(N)\left[\left[\alpha_{1,0}(\boldsymbol{x})\right]_i + \frac{1}{N}\left[\alpha_{1,1}(\boldsymbol{x})\right]_i + \mathcal{O}\left(\frac{1}{N^2}\right)\right] \tag{C.33}$$

$$\Rightarrow \quad \alpha_{1,0}(\boldsymbol{x}) \equiv 0, \quad \alpha_{1,1}(\boldsymbol{x}) = \boldsymbol{b}(\boldsymbol{x}) - \boldsymbol{d}(\boldsymbol{x}) \tag{C.34}$$

where in the last line the zeroth and first order terms of an expansion of the first jump moment in $1/N$ are read off. Likewise for the second moment we get:

$$\left[a_2(\boldsymbol{x})\right]_{ij} := \sum_{\Delta \boldsymbol{n}} \Delta n_i \Delta n_j\, W(\Delta \boldsymbol{n}; \boldsymbol{x}) \tag{C.35}$$

$$= \delta_{ij} \sum_{k, k \neq i} \left(B_i D_k + B_k D_i\right) - \left(1 - \delta_{ij}\right)\left(B_i D_j + B_j D_i\right)$$

$$= \delta_{ij}\left(B_i + D_i - 2 B_i D_i\right) - \left(1 - \delta_{ij}\right)\left(B_i D_j + B_j D_i\right) \tag{C.36}$$

$$= \delta_{ij}\, 2 x_i - 2 x_i x_j + \mathcal{O}\left(\frac{1}{N}\right) \tag{C.37}$$

$$\stackrel{!}{=} f(N)\left[\left[\alpha_{2,0}(\boldsymbol{x})\right]_{ij} + \frac{1}{N}\left[\alpha_{2,1}(\boldsymbol{x})\right]_{ij} + \mathcal{O}\left(\frac{1}{N^2}\right)\right] \tag{C.38}$$

$$\Rightarrow \quad \left[\alpha_{2,0}(\boldsymbol{x})\right]_{ij} = 2 x_i\left(\delta_{ij} - x_j\right) \tag{C.39}$$

The global pre-factor can be set as $f(N) = 1$. What we need to employ in case of a diffusion process - with vanishing macroscopic law: $\alpha_{1,0} = 0$ - is the (now

controlled) Kramers-Moyal expansion (see equation (C.3)):

$$\partial_t P(n,t) = \left[-\partial_{n_i} \left[a_1(x)\right]_i + \frac{1}{2}\partial_{n_i}\partial_{n_j} \left[a_2(x)\right]_{ij} + \ldots \right] P(n,t) \tag{C.40}$$

where control of the expansion is achieved by switching from the extensive variable $\boldsymbol{n}$ to the intensive $\boldsymbol{x} = \boldsymbol{n}/N$ and inserting the above expansion of the jump moments $a_k(\boldsymbol{x})$ to lowest order in $1/N$ $\left(\text{with the short notation } \partial_i := \frac{\partial}{\partial x_i}\right)$:

$$N^2\,\partial_t P(\boldsymbol{x},t) = \left\{ -\partial_i \left[\alpha_{1,1}(\boldsymbol{x})\right]_i + \frac{1}{2}\partial_i\partial_j \left[\alpha_{2,0}(\boldsymbol{x})\right]_{ij} + \mathcal{O}\left(\frac{1}{N}\right) \right\} P(\boldsymbol{x},t) \tag{C.41}$$

$$= \left\{ -\partial_i \left(b_i(\boldsymbol{x}) - d_i(\boldsymbol{x})\right) + \partial_i\partial_j x_i \left(\delta_{ij} - x_j\right) + \mathcal{O}\left(\frac{1}{N}\right) \right\} P(\boldsymbol{x},t) \tag{C.42}$$

C.3 Comparison of the Wright-Fisher and Moran expansions

To see the connection between the two complementary models of evolution under constant population size we compare the end results of the last two sections, equations (C.12) and (C.42), i.e. the respective diffusion limit Fokker-Planck equations.

$$N\,\partial_{t'} P(\boldsymbol{x},t') = \left\{ -\partial_i f_i(\boldsymbol{x}) + \frac{1}{2}\partial_i\partial_j x_i \left(\delta_{ij} - x_j\right) \right\} P(\boldsymbol{x},t') \qquad \text{(Wright-Fisher)}$$

$$N^2\,\partial_t P(\boldsymbol{x},t) = \left\{ -\partial_i \left(b_i(\boldsymbol{x}) - d_i(\boldsymbol{x})\right) + \partial_i\partial_j x_i \left(\delta_{ij} - x_j\right) \right\} P(\boldsymbol{x},t) \qquad \text{(Moran)}$$

We deliberately primed the time scale of the Wright-Fisher process at this point and left the forces f_i undetermined so far, so as to adjust them now at will. The whole reason for the derivation in this chapter is to be able to work with a *single* Fokker-Planck equation, which describes both models. In that sense, they are microscopically different but mesoscopically equivalent. We finally achieve equality of the two equations for the particular choice:

$$\boxed{f_i(\boldsymbol{x}) := \frac{1}{2}\left(b_i(\boldsymbol{x}) - d_i(\boldsymbol{x})\right) =: \frac{1}{2}F_i(\boldsymbol{x}), \quad t' := 2\,\frac{t}{N}} \tag{C.43}$$

This is the correspondence: one Moran generation, i.e. N birth-death events, is equivalent to *two* Wright-Fisher multinomial update steps, but importantly at *half* the Moran drift force $F(\boldsymbol{x}) = \text{Birth}(\boldsymbol{x}) - \text{Death}(\boldsymbol{x})$. This equivalence holds for any evolutionary stochastic model in the diffusion limit that can be cast in either of the two formulations.

C.4 Discrete vs. continuous time Master equations

This small section covers a subtlety in the formulation of e.g. the Moran model of evolution: the relationship between discrete time and continuous time Master equations [130]. Essentially, for every stochastic Markov model that is formulated in "turns", i.e. events that take place in discrete, evenly spaced points in time, there is a corresponding continuous time model. Correspondence means here, that the solutions to both models coincide at all times [130]. To exemplify the argument, we will consider a simple, one-dimensional one-step process with the general Master equation

$$\partial_t P_n(t) = \left\{ \left(\mathbb{E}^+ - 1\right) f_-(n) + \left(\mathbb{E}^- - 1\right) f_+(n) \right\} P_n(t) \tag{C.44}$$

where $f_\pm(n)$ is the rate for a transition $n \to n \pm 1$. We show along the lines of [130] that this continuous time Master equation is connected to a discrete time random walk.

$$\left(\boldsymbol{P}_t\right)_n := \text{Prob}\left(\text{ system in state n after t steps}\right) \tag{C.45}$$

$$\boldsymbol{P}_{t+1} \stackrel{!}{=} M\,\boldsymbol{P}_t \quad \Rightarrow \quad \boldsymbol{P}_{t+T} = M^T\,\boldsymbol{P}_t \tag{C.46}$$

$$\text{with} \quad M_{n'n} := \pi_+(n)\,\delta_{n',n+1} + \pi_-(n)\,\delta_{n',n-1} + \left(1 - \pi_+(n) - \pi_-(n)\right)\,\delta_{n',n} \tag{C.47}$$

where the $\pi_\pm(n)$ are the *probabilities* for a transition $n \to n \pm 1$ in one step. Now suppose that the time Δt between consecutive jumps is itself a random variable with an exponential distribution $\psi(\Delta t) = \frac{1}{\tau}\exp(-\Delta t/\tau)$. The probability for exactly k transitions in T time steps is then given by $\phi_k(T) = \left(\frac{T}{\tau}\right)^k \frac{1}{k!}\, e^{-T/\tau}$. This yields the following derivation

$$\boldsymbol{P}_{t+\Delta t} = \sum_{k=0}^{\infty} \phi_k(\Delta t)\, M^k\, \boldsymbol{P}_t = e^{\frac{\Delta t}{\tau}(M-1)}\,\boldsymbol{P}_t \xrightarrow{\Delta t \to 0} \partial_t \boldsymbol{P}(t) = \frac{M-1}{\tau}\,\boldsymbol{P}(t) \tag{C.48}$$

where $M - 1$ is understood to be the matrix M minus the identity matrix. This last equation will be equivalent to the sought-after birth-death Master equation (C.44) if we choose

$$\frac{\pi_\pm(n)}{\tau} \stackrel{!}{=} f_\pm(n) \tag{C.49}$$

which means that the *rates* $f_\pm(n)$ appearing in the continuous time Master equation are in fact the *probabilities per unit time* for the transitions, where the *unit* of time is naturally the mean time τ between consecutive exponentially distributed transition events. The continuous time Master equation of e.g. the Moran model implicitly assumes this distribution of jump events.

Appendix D

Elimination of fast variables in the two-locus model with recombination

In this chapter, the elimination of the fast fluctuating variables - the single mutant frequencies x_{aB} and x_{Ab} - in the two-locus model of chapter 5 is carried out explicitly. First, the focus will be put on the deterministic level of the system and the small parameter for the time scale separation will be identified to be the size of the single mutants itself. The proper expansion of the deterministic equations of motion goes along the lines of [131]. Guided by these results, the adiabatic elimination is repeated on the stochastic level of the theory using the projector method [132], that was originally developed in the field of quantum statistics by Nakajima [133] and Zwanzig [134]. The main result is that the coupling of the fast and slow modes is linear and the system falls into the "silent slave" category [132], i.e. there is no feedback of the fluctuations of the fast modes towards the dynamics of the slow modes. In this regime, the "naive" deterministic elimination method is valid even on the stochastic level, which simply solves the fast dynamics first keeping the slow part as a constant parameter and then reinserts the result in the appropriate equations to arrive at an effective slow dynamics. This is the justification for the simplifications made in the main text.

D.1 Deterministic elimination

To find evidence for a time scale separation, the starting point is to investigate the "deterministic level" of the system. To repeat the important steps of the derivation made in chapter 5: the expansion of the Master equation for $N \to \infty$ (with Nu etc. held constant) revealed that there is no macroscopic law in the sense of the linear noise approximation (LNA) [40]. The proper expansion leads immediately to a Fokker-Planck equation (5.19). Only the presence of additional large param-

eters (such as Ns_d and Ns_b) allows for a secondary expansion (of LNA type) and awards the drift terms with the role of macroscopic (or deterministic) laws. For the very concrete two-locus/two-allele model considered in the main text, these deterministic equations of motion are:

$$x_{ab} \to x, \quad x_{aB} \to y_1, \quad x_{Ab} \to y_2, \quad x_{AB} \to z, \quad \sum_i x_i = 1 \tag{D.1}$$

$$\sigma_b := Ns_b, \quad \sigma_d := Ns_d, \quad \rho := Nr, \quad \mu := Nu \tag{D.2}$$

$$\frac{\mathrm{d}y_i}{\mathrm{d}t} = \rho\,(xz - y_1 y_2) - \sigma_d y_i\,(1 - y_1 - y_2) - \sigma_b y_i z + \mu\,(x + z - 2y_i), \quad i = 1,2 \tag{D.3}$$

$$\frac{\mathrm{d}z}{\mathrm{d}t} = \sigma_d z\,(y_1 + y_2) + \sigma_b z(1 - z) - \rho\,(xz - y_1 y_2) - \mu\,(2z - y_1 - y_2) \tag{D.4}$$

The symmetry of the two single mutant species suggests to introduce the new variables $y := y_1 + y_2$ and $\eta := y_1 - y_2$. With these, the equations of motion read as

$$\frac{\mathrm{d}\eta}{\mathrm{d}t} = -\sigma_d \eta \left(1 - y + \frac{\sigma_b}{\sigma_d} z + \frac{2\mu}{\sigma_d}\right) \tag{D.5}$$

$$\frac{\mathrm{d}y}{\mathrm{d}t} = -\sigma_d y(1 - y) - \sigma_b yz + 2\rho z - \frac{\rho}{2}(y + 2z)^2 + \frac{\rho}{2}\eta^2 + 2\mu\,(1 - 2y) \tag{D.6}$$

$$\frac{\mathrm{d}z}{\mathrm{d}t} = \sigma_d yz + \sigma_b z(1 - z) - \rho z + \frac{\rho}{4}(y + 2z)^2 - \frac{\rho}{4}\eta^2 + \mu(y - 2z) \tag{D.7}$$

The terms in those equations are ordered by the size of the evolutionary forces involved according to the parameter regime specified in section 5.2.3. In any case, from the last equation above it can be seen that the dynamics of the double mutant frequency (z) takes place on slower time scale than the other two variables only if $y \ll 1$ at all times. In fact, we anticipate that $y \leq \mathcal{O}(\sigma_b/\sigma_d) \ll 1$. Furthermore, we are most interested in values of recombination strength ρ in the neighborhood of σ_b, i.e. in the reach of criticality. These two observations motivate the following substitutions:

$$\rho =: R\sigma_b,\ R \leq \mathcal{O}(1),\ \varepsilon := \frac{\sigma_b}{\sigma_d} \ll 1,\ y =: \varepsilon\tilde{y},\ \eta =: \varepsilon\tilde{\eta},\ \frac{\mathrm{d}}{\mathrm{d}t} =: \sigma_b \frac{\mathrm{d}}{\mathrm{d}\tau} = \varepsilon\sigma_d \frac{\mathrm{d}}{\mathrm{d}\tau} \tag{D.8}$$

This leads to the following equations of motion:

$$\varepsilon \frac{\mathrm{d}\tilde{\eta}}{\mathrm{d}\tau} = -\tilde{\eta}\left(1 - \varepsilon\tilde{y} + \varepsilon z + \frac{2\mu}{\sigma_b}\varepsilon\right) \tag{D.9}$$

$$\varepsilon \frac{\mathrm{d}\tilde{y}}{\mathrm{d}\tau} = -\tilde{y}(1 - \varepsilon\tilde{y}) - \varepsilon\tilde{y}z + 2Rz - \frac{R}{2}(\varepsilon\tilde{y} + 2z)^2 + \frac{R}{2}\varepsilon^2\tilde{\eta}^2 + \frac{2\mu}{\sigma_b}(1 - 2\varepsilon\tilde{y}) \tag{D.10}$$

$$\varepsilon \frac{\mathrm{d}z}{\mathrm{d}\tau} = \varepsilon\tilde{y}z + \varepsilon z(1 - z) - R\varepsilon z + \frac{R}{4}\varepsilon(\varepsilon\tilde{y} + 2z)^2 - \frac{R}{4}\varepsilon^3\tilde{\eta}^2 + \varepsilon\frac{\mu}{\sigma_b}(\varepsilon\tilde{y} - 2z) \tag{D.11}$$

The time scale separation between $(\tilde{\eta}, \tilde{y})$ on one hand and z on the other is now explicit (assuming $\tilde{y} \leq \mathcal{O}(1)$). The next step is to expand the fast modes in orders of ε and look at the resulting equations of motion order by order.

$$\tilde{\eta} = \tilde{\eta}^{(0)} + \varepsilon\,\tilde{\eta}^{(1)} + \ldots \quad \text{and} \quad \tilde{y} = \tilde{y}^{(0)} + \varepsilon\,\tilde{y}^{(1)} + \ldots \tag{D.12}$$

$$\mathcal{O}\left(\varepsilon^0\right): \quad 0 = -\tilde{\eta}^{(0)} \tag{D.13}$$

$$0 = -\tilde{y}^{(0)} + 2Rz(1-z) + \frac{2\mu}{\sigma_b} \tag{D.14}$$

This order gives $\tilde{y}^{(0)}$ as a function of z and we recover the lowest order expansion of the single mutant frequency in equation (5.33) on page 113. We also note that indeed $\tilde{y}^{(0)}(z) \leq \mathcal{O}(1)$ for $\mu \ll \sigma_b$. The next order in ε yields:

$$\mathcal{O}\left(\varepsilon^1\right): \quad \frac{\mathrm{d}\tilde{\eta}^{(0)}}{\mathrm{d}\tau} = -\tilde{\eta}^{(1)} + \tilde{\eta}^{(0)}\left(\tilde{y}^{(0)} - z - \frac{2\mu}{\sigma_b}\right) \quad \Rightarrow \quad \tilde{\eta}^{(1)} = 0 \tag{D.15}$$

$$\frac{\mathrm{d}\tilde{y}^{(0)}}{\mathrm{d}\tau} = -\tilde{y}^{(1)} + \tilde{y}^{(0)}\left(\tilde{y}^{(0)} - z - 2Rz - \frac{4\mu}{\sigma_b}\right) \tag{D.16}$$

$$\frac{\mathrm{d}z}{\mathrm{d}\tau} = \tilde{y}^{(0)} z + z(1-z) - Rz(1-z) - \frac{2\mu}{\sigma_b} z \tag{D.17}$$

It is now apparent that $\tilde{\eta}$ disappears in all orders of ε. The second equation determines $\tilde{y}^{(1)}$ as a function of z (after inserting $\tilde{y}^{(0)}$ and the equation of motion for z). This will not be further pursued here. Importantly, the last equation above is the effective equation of motion for the slow variable z:

$$\frac{\mathrm{d}z}{\mathrm{d}\tau} = 2Rz(1-z)\left(z - \frac{R-1}{2R}\right) \tag{D.18}$$

Converting back to the original quantities, we thus recover the effective drift term equation (5.58) used in the main text on page 121, involving the saddle point z_{cr} (to leading order):

$$\frac{\mathrm{d}z}{\mathrm{d}\tau} = F_z(z) = 2\rho\, z(1-z)\left(z - \frac{\rho - \sigma_b}{2\rho}\right) \tag{D.19}$$

D.2 Stochastic elimination

The results from the last section are now used to guide the adiabatic elimination of the fast modes in the full stochastic context. The original three dimensional Fokker-Planck equation (5.19) on page 110 (for y_1, y_2 and z) is first converted

to the new variables (η, y, z) introduced in the last section. The resulting three-dimensional Fokker-Planck equation (FPE) reads as:

$$\begin{aligned}\partial_t P(y,\eta,z,t) \;=\; &\Big\{-\partial_y F_y - \partial_\eta F_\eta - \partial_z F_z + \partial_y^2 y(1-y) + \partial_z^2 z(1-z) + \partial_\eta^2\left(y-\eta^2\right)\\ &+\partial_y\partial_\eta\, 2\eta(1-y) - \frac{1}{2}\partial_y\partial_z z(3y+\eta) - \frac{1}{2}\partial_\eta\partial_z z(y+3\eta)\Big\} P(y,\eta,z,t)\end{aligned} \tag{D.20}$$

with the drift terms $F_{y,\eta,z}$ given by the right hand sides of equations (D.6) and (D.7). Now, the substitutions from (D.8) are inserted. Anticipating that we are interested in an expansion in the small parameter $\varepsilon = \sigma_b/\sigma_d$, the following Fokker-Plank operator $\hat{L}$ is given to sub-leading order in ε.

$$\partial_\tau P(\tilde{y},\tilde{\eta},z,\tau) =: \hat{L} P(\tilde{y},\tilde{\eta},z,\tau) \tag{D.21}$$

$$\begin{aligned}\hat{L} := \frac{1}{\varepsilon}\Big\{&-\partial_{\tilde{y}}\left(-\tilde{y} + 2Rz(1-z) + \frac{2\mu}{\sigma_b}\right) + \partial_{\tilde{\eta}}\,\tilde{\eta} + \partial_{\tilde{y}}^2\frac{\tilde{y}}{\sigma_b} + \partial_{\tilde{\eta}}^2\frac{\tilde{y}}{\sigma_b} + \partial_{\tilde{y}}\partial_{\tilde{\eta}}\frac{2\tilde{\eta}}{\sigma_b}\Big\}\\ &+\Big\{-\partial_{\tilde{y}}\,\tilde{y}\left(\tilde{y} - z(1+2R) - \frac{4\mu}{\sigma_b}\right) - \partial_z\left(z(1-z)(1-R) + \tilde{y}z - \frac{2\mu}{\sigma_b}z\right)\\ &-\partial_{\tilde{\eta}}\,\tilde{\eta}\left(-\tilde{y} + z + \frac{2\mu}{\sigma_b}\right) - \partial_{\tilde{y}}^2\frac{\tilde{y}^2}{\sigma_b} - \partial_{\tilde{\eta}}^2\frac{\tilde{\eta}^2}{\sigma_b} + \partial_z^2\frac{z(1-z)}{\sigma_b} - \partial_{\tilde{y}}\partial_z\frac{z(3\tilde{y}+\tilde{\eta})}{2\sigma_b}\\ &-\partial_{\tilde{\eta}}\partial_z\frac{z(y+3\eta)}{2\sigma_b} + \partial_{\tilde{y}}\partial_{\tilde{\eta}}\frac{2\tilde{\eta}\tilde{y}}{\sigma_b}\Big\} + \mathcal{O}(\varepsilon)\end{aligned} \tag{D.22}$$

The contributions to above FPE are already sorted by their order in ε, with a dominant part describing the evolution of the fast single mutant frequency sum and difference $(\tilde{y}, \tilde{\eta})$ (first line). The drift terms of that leading order operator are *linear* in $\tilde{y}$ and $\tilde{\eta}$. In a linear noise expansion, the leading order macroscopic part of a stochastic trajectory is fixed to the (z-dependent) size $2\tilde{\beta} := 2Rz(1-z) + \frac{2\mu}{\sigma_b}$ and 0, respectively. But these values are exactly the leading order terms of the deterministic elimination in the last section. This suggests to shift the sum variable once more:

$$(\tilde{y},\tilde{\eta},z) \to \left(v := \tilde{y} - 2\tilde{\beta}(z), \tilde{\eta}, z\right) \;\Rightarrow\; \left(\partial_{\tilde{y}}, \partial_{\tilde{\eta}}, \partial_z\right) \to \left(\partial_v,\; \partial_{\tilde{\eta}},\; -2\tilde{\beta}'(z)\partial_v + \partial_z\right) \tag{D.23}$$

$$\tilde{\beta}(z) := \frac{\beta(z)}{\sigma_b} := \frac{\rho z(1-z) + \mu}{\sigma_b} \tag{D.24}$$

$$\partial_\tau P(v,\tilde{\eta},z,\tau) = \hat{L} P(v,\tilde{\eta},z,\tau), \quad \text{with} \quad \hat{L} = \frac{1}{\varepsilon}\hat{L}_1 + \hat{L}_2 + \hat{L}_3 + \mathcal{O}(\varepsilon) \tag{D.25}$$

$$\hat{L}_1 = -\partial_v\,(-v) - \partial_{\tilde{\eta}}\,(-\tilde{\eta}) + \left(\partial_v^2 + \partial_{\tilde{\eta}}^2\right)\frac{1}{\sigma_b}\left(v + 2\tilde{\beta}(z)\right) + \partial_v\partial_{\tilde{\eta}}\,\frac{2\,\tilde{\eta}}{\sigma_b} \tag{D.26}$$

$$\hat{L}_2 = -\partial_z\,v\,z - \partial_v\,(\dots) - \partial_{\tilde{\eta}}\,(\dots) + \partial_v^2\,(\dots) + \partial_{\tilde{\eta}}^2\,(\dots) - \partial_v\,\partial_z\,(\dots) - \partial_{\tilde{\eta}}\,\partial_z\,(\dots) \tag{D.27}$$

$$\hat{L}_3 = -\partial_z\,z(1-z)\,(1-R-2Rz) + \partial_z^2\,\frac{z(1-z)}{\sigma_b} \tag{D.28}$$

The ellipses in $\hat{L}_2$ are simple functions of the three variables, whose specific form is not important for reasons to become clear momentarily. The main idea of the projector method is to find the stationary solution to $\hat{L}_1$, i.e. $\hat{L}_1\,p_s(v,\tilde{\eta}) = 0$, and use that distribution to construct a projector $\mathcal{P}$ on the null space of $\hat{L}_1$ [132].

$$\hat{L}_1\,p_s(v,\tilde{\eta}) = 0 \quad \Rightarrow \quad \mathcal{P}\,f(v,\tilde{\eta},z) := p_s(v,\tilde{\eta})\int dv'\,d\tilde{\eta}'\;f(v',\tilde{\eta}',z) \tag{D.29}$$

The present fast operator $\hat{L}_1$ affords for a simple potential solution for its stationary distribution [41].

$$p_s(v,\tilde{\eta}) = \mathcal{Z}^{-1}\,e^{-\phi(v,\tilde{\eta})}, \quad \phi = \sigma_b\,v + (1-\sigma_b\,\tilde{\beta})\,\ln\left(\left(v+2\tilde{\beta}\right)^2 - \tilde{\eta}^2\right) \tag{D.30}$$

$$\Rightarrow \quad \boxed{p_s(y,\eta) = \mathcal{Z}^{-1}\,e^{-\sigma_d\,y}\left(y^2-\eta^2\right)^{\beta(z)-1}} \tag{D.31}$$

The normalization constant $\mathcal{Z}$ is evaluated as

$$\mathcal{Z} := \int_0^1 \mathrm{d}y\int_{-y}^{y}\mathrm{d}\eta\;e^{-\sigma_d\,y}\left(y^2-\eta^2\right)^{\beta(z)-1} \tag{D.32}$$

$$= \sqrt{\pi}\,\frac{\Gamma(\beta)}{\sigma_d^{2\beta}}\,\frac{\Gamma(2\beta)-\Gamma(2\beta,\sigma_d)}{\Gamma(\beta+\frac{1}{2})} \tag{D.33}$$

$$\Gamma(a,b) := \int_b^\infty \mathrm{d}t\;t^{a-1}\,e^{-t}, \quad \Gamma(a) = \Gamma(a,0) \tag{D.34}$$

This is the quasi-stationary distribution of the fast variables (y,η). Most importantly, we note that re-introducing the original single-mutant frequencies leads to a distribution that is the product of two identical distributions:

$$x_{aB} = \frac{1}{2}\,(y+\eta)\,, \quad x_{Ab} = \frac{1}{2}\,(y-\eta) \tag{D.35}$$

$$p_s(y,\eta) \to p_s(x_{Ab},x_{aB}) =: q_s(x_{Ab})\cdot q_s(x_{aB}), \quad q_s(x) \propto e^{-\sigma_d\,x}\,x^{\beta(z)-1} \tag{D.36}$$

$$\Rightarrow \quad \langle x_{Ab}\,x_{aB}\rangle_s = \langle x_{Ab}\rangle_s\langle x_{aB}\rangle_s \tag{D.37}$$

This is the rigorous argument for the assumption made in the main text, that under time scale separation the two single mutant frequencies are uncorrelated to lowest order. The mean and variance of the fast modes can be readily given:

$$\langle y\rangle_s = \frac{\Gamma(2\beta+1)-\Gamma(2\beta+1,\sigma_d)}{\sigma_d\left(\Gamma(2\beta)-\Gamma(2\beta,\sigma_d)\right)} = \frac{2\beta}{\sigma_d}\left(1-e^{-\sigma_d}\right)+\mathcal{O}\left(\frac{\beta^2}{\sigma_d^2}\right) \tag{D.38}$$

$$\mathrm{var}_s(y) := \langle y^2\rangle_s - \langle y\rangle_s^2 = \frac{2\beta}{\sigma_d^2}\left[1-\Gamma(2,\sigma_d)\right]+\mathcal{O}\left(\frac{\beta^2}{\sigma_d^2}\right) \tag{D.39}$$

$$\langle \eta\rangle_s = 0, \quad \mathrm{var}_s(\eta) = \mathrm{var}_s(y) \tag{D.40}$$

For completeness, one can also average out the difference of single mutant frequencies η to arrive at the marginal distribution for the total y:

$$p_s(y) := \int_{-y}^{y} \mathrm{d}\eta\ p_s(y,\eta) = e^{-\sigma_d y} y^{2\beta(z)-1}\,\frac{\sigma_d^{2\beta}}{\Gamma(2\beta)-\Gamma(2\beta,\sigma_d)} \tag{D.41}$$

Proceeding with the elimination routine according to [132], we note that the operators $\hat{L}_i$ above are organized in a way to guarantee the three central relations

$$\mathcal{P}\hat{L}_1 = \hat{L}_1\mathcal{P} = 0, \quad \mathcal{P}\hat{L}_2\mathcal{P} = 0, \quad \mathcal{P}\hat{L}_3 = \hat{L}_3\mathcal{P} \tag{D.42}$$

The first and third relation holds by definition of $\mathcal{P}$ and $\hat{L}_3$ The second relation is ensured by both $\langle v\rangle_s = \langle\tilde{\eta}\rangle = 0$.

$$\left[\mathcal{P}\hat{L}_2\mathcal{P}\right] f(v,\eta,z) = p_s(v,\eta)\int \mathrm{d}v'\,\mathrm{d}\eta'\ \hat{L}_2\left(p_s(v',\eta')\,\hat{f}(z)\right) \tag{D.43}$$

$$\left[\mathcal{P}\left(\partial_z v z\right)\mathcal{P}\right] f(v,\eta,z) \propto \langle v\rangle_s = 0 \tag{D.44}$$

Terms in $\hat{L}_2$ that start with ∂_v or $\partial_{\tilde{\eta}}$ vanish upon being acted on by $\mathcal{P}$ [41], if boundary terms can be dropped. The projector $\mathcal{P}$ is now used is to find solutions to the FPE in the null space of $\hat{L}_1$, i.e. where the fast mode is in a quasi-stationary state and can be averaged out:

$$\begin{aligned} P(v,\tilde{\eta},z,\tau) \to p(v,\tilde{\eta},z,\tau) &:= \mathcal{P}P(v,\tilde{\eta},z,\tau) \\ = p_s(v,\tilde{\eta})\int \mathrm{d}v'\,\mathrm{d}\tilde{\eta}'\ P(v',\tilde{\eta}',z,\tau) &=: p_s(v,\tilde{\eta})\,\hat{P}(z,\tau) \end{aligned} \tag{D.45}$$

Without going into too much detail, which can be found in [132], the effective FPE describing the evolution of the reduced part $\hat{P}(z,\tau)$ is in the present case given by:

$$\partial_\tau \hat{P}(z,\tau) = \lim_{\varepsilon\to 0}\left(\hat{L}_3 - \varepsilon\,\mathcal{P}\hat{L}_2\hat{L}_1^{-1}\hat{L}_2\right)\hat{P}(z,\tau) = \hat{L}_3\hat{P}(z,\tau) \tag{D.46}$$

Since $\hat{L}_2$ does not scale with ε, there is no change in the FPE for the slow mode due to the fluctuations of the fast mode. This is the "silent slave" limit of [132]. In this limit, the "naive" adiabatic elimination performed out in the main text is valid, which completes the argument of the derivation of the double mutant fixation rate.

Bibliography

[1] T. Dobzhansky. Nothing in biology makes sense except in the light of evolution. *The American Biology Teacher*, 35:125–129, 1973.

[2] International Human Genome Sequencing Consortium, E. Lander, et al. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[3] J.N. Hirschhorn and M.J. Daly. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6(2):95–108, 2005.

[4] The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447:661–678, 2007.

[5] International Cancer Genome Consortium. International network of cancer genome projects. *Nature*, 464(7291):993–8, 2010.

[6] A. Fischer, C. Greenman, and V. Mustonen. Germline fitness based scoring of cancer mutations. *Genetics*, 188:383–393, 2011.

[7] E.T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

[8] A.M. Moses and R. Durbin. Inferring selection on amino acid preference in protein domains. *Molecular biology and evolution*, 26(3):527–36, 2009.

[9] R.D. Finn et al. The pfam protein families database. *Nucleic acids research*, 38(Database issue):D211–22, 2010.

[10] C. Greenman et al. Patterns of somatic mutation in human cancer genomes. *Nature*, 446:153–8, 2007.

[11] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77:257–285, 1989.

[12] D.M. Weinreich, L. Chao, and R.A. Watson. Sign epistasis and genetic constraint on evolutionary trajectories. *Evolution*, 59:1165–1174, 2005.

[13] K. Jain. Time to fixation in the presence of recombination. *Theoretical Population Biology*, 77(1):23 – 31, 2010.

[14] S.C. Park and J. Krug. Bistability in two-locus models with selection, mutation, and recombination. *J.Math.Biol.*, 2010.

[15] Y. Iwasa, F. Michor, and M.A. Nowak. Stochastic tunnels in evolutionary dynamics. *Genetics*, 166:1571–1579, 2004.

[16] S. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.

[17] M.R. Stratton. Exploring the genomes of cancer cells: Progress and promise. *Science*, 331:1553–1558, 2011.

[18] M.R. Stratton, P.J. Campbell, and A. Futreal. The cancer genome. *Nature*, 458:719–724, 2009.

[19] K. W. Vogelstein, B.and Kinzler. The multistep nature of cancer. *Trends in Genetics*, 9:138–141, 1993.

[20] P.C. Nowell. The clonal evolution of tumor cell populations. *Science*, 194:22–28, 1976.

[21] V. Mustonen and M. Lässig. From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaption. *Trends in Genetics*, 25:111–119, 2009.

[22] S. E. Scherer. *A Short Guide to the Human Genome*. Cold Spring Harbor Laboratory, 2008.

[23] E.S. Lander. Initial impact of the sequencing of the human genome. *Nature*, 470:187–197, 2011.

[24] W. Lee, P. Yue, and Z. Zhang. Analytical methods for inferring functional effects of single base pair substitutions in human cancers. *Human Genetics*, 126(4):481–98, 2009.

[25] A. Torkamani, G. Verkhivker, and N. J. Schork. Cancer driver mutations in protein kinase genes. *Cancer Letters*, 281(2):117–127, 2009.

[26] P.C. Ng and S. Henikoff. Sift: Predicting amino acid changes that affect protein function. *Nucleic acids research*, 31(13):3812–4, 2003.

[27] C.D. Greenman et al. Statistical analysis of pathogenicity of somatic mutations in cancer. *Genetics*, 173:2187–98, 2006.

[28] J. S. Kaminker, Y. Zhang, C. Watanabe, and Z. Zhang. Canpredict: a computational tool for predicting cancer-associated missense mutations. *Nucleic acids research*, 35(Web Server issue):W595–8, 2007.

[29] A. Torkamani and N. J. Schork. Prediction of cancer driver mutations in protein kinases. *Cancer Research*, 68(6):1675–82, 2008.

[30] H. Carter et al. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Research*, 69(16):6660–6667, 2009.

[31] D. Gilis and M. Rooman. Popmusic, an algorithm for predicting protein mutant stability changes. application to prion proteins. *Protein Engineering Design and Selection*, 13:849–856, 2000.

[32] S. Henikoff and J.G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA*, 89:10915–10919, 1992.

[33] M. Gribskov, A.D. McLachlan, and D. Eisenberg. Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci USA*, 84:4355–4358, 1987.

[34] R.L. Tatusov, S.F. Altschul, and E.V. Koonin. Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. *Proc Natl Acad Sci USA*, 91:12091–12095, 1994.

[35] M. Kimura. Stochastic processes and distribution of gene frequencies under natural selection. *Cold Spring Harb Symp Quant Biol*, 20:33–53, 1955.

[36] M. Kimura. Diffusion models in population genetics. *Journal of Applied Probability*, 1(2):177–232, 1964.

[37] V. Mustonen and M. Lässig. Evolutionary population genetics of promoters: Predicting binding sites and functional phylogenies. *Proceedings of the National Academy of Sciences*, 102(44):15936–41, 2005.

[38] I.M. Rouzine, A. Rodrigo, and J.M. Coffin. Transition between stochastic evolution and deterministic evolution in the presence of selection: General theory and application to virology. *Microbiology and Molecular Biology Reviews*, 65(1):151–185, 2001.

[39] W.J. Ewens. *Mathematical Population Genetics*. Springer, 2004.

[40] N.G. van Kampen. *Stochastic Processes in Physics and Chemestry*. North Holland Personal Library, 2007.

[41] C.W. Gardiner. *Stochastic methods: a handbook for the natural and social sciences*. Springer, 2009.

[42] M. Bulmer. The selection-mutation-drift theory of synonymous codon usage. *Genetics*, 129(3):897–907, 1991.

[43] A.L. Halpern and W. J. Bruno. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Molecular biology and evolution*, 15(7):910–7, 1998.

[44] B. Gaveau, M. Moreau, and J. Toth. Variational nonequilibrium thermodynamics of reaction-diffusion systems. i.the information potential. *Journal of Chemical Physics*, 111:7736–7747, 1999.

[45] E. Jaynes. Information theory and statistical mechanics i, ii. *Physical Review*, 106, 108:620–630, 171–190, 1957.

[46] S. Kullback. *Information theory and statistics*. Dover, 2 edition, 1997.

[47] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press, April 1998.

[48] J.G. Henikoff and S. Henikoff. Using substitution probabilities to improve position-specific scoring matrices. *Bioinformatics*, 12(2):135–134, 1996.

[49] C.E. Lawrence, S.F. Altschul, M.S. Boguski, J.S. Liu, A.F. Neuwald, and J.C. Wootton. Detecting subtle sequence signals: A gibbs sampling strategy for multiple alignment. *Science*, 262(5131):208–214, 1993.

[50] T. Seidenfeld. *Entropy and uncertainty*. Reidel, 1987.

[51] E. Soofi. Information theoretic regression methods. *Advances in Econometrics*, 12:25–83, 1997.

[52] S. Kullback and R.A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22:79–86, 1951.

[53] D.V. Gokhale and S. Kullback. The minimum discrimination information approach in analyzing categorical data. *Communications in Statistics - Theory and Methods*, 7:987–1005, 1978.

[54] E. Soofi. Principal information theoretic approaches. *Journal of the American Statistical Association*, 95:1349–1353, 2000.

[55] E. Soofi and J.J. Retzer. Information indices: unification and applications. *Journal of Econometrics*, 107:17–40, 2002.

[56] H. Alzer. Sharp inequalities for the digamma and polygamma functions. *Forum Math.*, 16:181–221, 2004.

[57] J. Cao, D. Niu, and F. Qi. Convexities of some functions involving the polygamma functions. *Applied Mathematics E-Notes*, 8:53–57, 2008.

[58] J.L. Jensen. Sur les fonctions convexes et les inégualités entre les valeurs moyennes. *Acta Mathematica*, 30:175–193, 1906.

[59] P.G. Higgs and T.K. Attwood. *Bioinformatics and Molecular Evolution*. Blackwell, 2004.

[60] G. Manning, D.B. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam. The protein kinase complement of the human genome. *Science*, 298(5600):1912–1934, 2002.

[61] S.R. Eddy. Profile hidden markov models. *Bioinformatics Review*, 14:755–763, 1998.

[62] S.R. Eddy. "hmmer". http://hmmer.org.

[63] S. Henikoff, J.C. Wallace, and J.P. Brown. Finding protein similarities with nucleotide sequence databases. *Methods Enzymol*, 183:111–132, 1990.

[64] S. Henikoff and J.G. Henikoff. Automated assembly of protein blocks for database searching. *Nucleic acids research*, 19:6565–6572, 1991.

[65] S. Henikoff and J.G. Henikoff. Position-based sequence weights. *Journal of Molecular Biology*, 243:574–578, 1994.

[66] J. Berg, S. Willmann, and M. Lässig. Adaptive evolution of transcription factor binding sites. *BMC Evolutionary Biology*, 4(42):1, 2004.

[67] S.A. Sawyer and D.L. Hartl. Population genetics of polymorphism and divergence. *Genetics*, 132(4):1161–76, 1992.

[68] L.M.F. Merlo, J. W. Pepper, B. J. Reid, and C.C. Maley. Cancer as an evolutionary and ecological process. *Nature Reviews Cancer*, 6(12):924–35, 2006.

[69] C. Attolini and F. Michor. Evolutionary theory of cancer. *Ann. N.Y. Acad. Sci.*, 1168:23–51, 2009.

[70] H.H.Q. Heng et al. The evolutionary mechanism of cancer. *Journal of Cellular Biochemistry*, 109:1072–1084, 2010.

[71] M. Gerstung and N. Beerenwinkel. Waiting time models of cancer progression. 2009. arXiv:0807.3638v2 [q-bio.PE].

[72] N. Beerenwinkel et al. Genetic progression and the waiting time to cancer. *PLoS Computational Biology*, 3:2239–2246, 2007.

[73] J. Foo, K. Leder, and F. Michor. Stochastic dynamics of cancer initiation. *Phys.Biol.*, 8:015002, 2011.

[74] Y. Zhang et al. Molecular evolutionary analysis of cancer cell lines. *Mol Cancer Ther*, 9(2):279–91, 2010.

[75] I. Bozica et al. Accumulation of driver and passenger mutations during tumor progression. *PNAS*, 107:18545–18550, 2010.

[76] R. Durrett et al. Evolutionary dynamics of tumor progression with random fitness values. *Theor.Pop.Biol.*, 78:54–66, 2010.

[77] Z. Yang, S. Ro, and B. Rannala. Likelihood models of somatic mutation and codon substitution in cancer genes. *Genetics*, 165(2):695–705, 2003.

[78] A.F. Rubin and P. Green. Mutation patterns in cancer genomes. *Proc Natl Acad Sci USA*, 106(51):21766–70, 2009.

[79] L. Li et al. Discovering cancer genes by integrating network and functional properties. *BMC medical genomics*, 2:61, 2009.

[80] W. Lee, Y. Zhang, K. Mukhyala, R. A. Lazarus, and Z. Zhang. Bi-directional sift predicts a subset of activating mutations. *PLoS ONE*, 4(12):e8311, 2009.

[81] V. E. Ramensky, P. Bork, and S. R. Sunyaev. Human non-synonymous snps: server and survey. *Nucleic acids research*, 30(17):3894–900, 2002.

[82] Y. Bromberg and B. Rost. Snap: predict effect of non-synonymous polymorphisms on function. *Nucleic acids research*, 35:3823–3835, 2007.

[83] D. M. Jordan, V. E. Ramensky, and S. R. Sunyaev. Human allelic variation: perspective from protein function, structure, and evolution. *Current Opinion in Structural Biology*, 20(3):342–50, 2010.

[84] P. Lahiry, A. Torkamani, N. J. Schork, and R. A. Hegele. Kinase mutations in human disease: interpreting genotype–phenotype relationships. *Nature Reviews Genetics*, 11(1):60–74, 2010.

[85] K.D: Pruitt, T. Tatusova, and D. R. Maglott. Ncbi reference sequences (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*, 35(Database issue):D61–5, 2007.

[86] Z. Kan et al. Diverse somatic mutation patterns and pathway alterations in human cancers. *Nature*, 466(7308):869–73, 2010.

[87] P. Yue et al. Inferring the functional effects of mutation through clusters of mutations in homologous proteins. *Hum Mutat*, 31(3):264–71, 2010.

[88] L.D. Wood et al. The genomic landscapes of human breast and colorectal cancers. *Science*, 318:1108–1113, 2007.

[89] A.J. Whitmarsh and R.J. Davis. Role of mitogen-activated protein kinase kinase 4 in cancer. *Oncogene*, 26:3172–3184, 2007.

[90] R.J. Clifford, M.N. Edmonson, C. Nguyen, and K.H. Buetow. Large-scale analysis of non-synonymous coding region single nucleotide polymorphisms. *Bioinformatics*, 20(7):1006–14, 2004.

[91] S.J. Baker et al. Chromosome 17 deletions and p53 gene mutations in colorectal carcinomas. *Science*, 244:217–221, 1989.

[92] S.A. Forbes et al. Cosmic: mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic acids research*, 39:D945–D950, 2010.

[93] I. Varela et al. Exome sequencing identifies frequent mutation of the swi/snf complex gene pbrm1 in renal carcinoma. *Nature*, 469:539–542, 2011.

[94] E. Papaemmanuil et al. Somatic sf3b1 mutation in myelodysplasia with ring sideroblasts. *New England Journal of Medicine*, 365:1384–1395, 2011.

[95] D. Talavera, M.S. Taylor, and J.M. Thornton. The (non)malignancy of cancerous amino acidic substitutions. *Proteins*, 78(3):518–29, 2010.

[96] M. Mort et al. In silico functional profiling of human disease-associated and polymorphic amino acid substitutions. *Hum Mutat*, 31(3):335–46, 2010.

[97] J.M.G. Izarzugaza, O.C. Redfern, C.A. Orengo, and A. Valencia. Cancer-associated mutations are preferentially distributed in protein kinase functional sites. *Proteins*, 77(4):892–903, 2009.

[98] A. Dixit et al. Sequence and structure signatures of cancer mutation hotspots in protein kinases. *PLoS ONE*, 4(10):e7485, 2009.

[99] I.A. Adzhubei et al. A method and server for predicting damaging missense mutations. *Nature Methods*, 7:248–249, 2010.

[100] M. Kimura. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature*, 267:275–276, 1977.

[101] Z. Yang and JP Bielawski. Statistical methods for detecting molecular adaptation. *Trends Ecol Evol*, 15:496–503, 2000.

[102] A. Torkamani, N. Kannan, S. Taylor, and N. Schork. Congenital disease snps target lineage specific structural elements in protein kinases. *Proceedings of the National Academy of Sciences*, 105(26):9011–16, 2008.

[103] L. Baum and J.A. Egon. An inequality with applications to statistical estimation for probabilistic functions of a markov process and to a model for ecology. *Bull. Amer. Metereol. Soc.*, 73:360–363, 1967.

[104] L.E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state markov chains. *Ann. Math. Statist.*, 37:1554–1563, 1966.

[105] L. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *Ann. Math. Statist.*, 41:164–171, 1970.

[106] L. Baum and G.R. Sell. Growth functions for transformations on manifolds. *Pac. J. Math.*, 27:211–227, 1968.

[107] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. Roy. Stat. Soc.*, 39:1–38, 1977.

[108] S.M. Omohundro. Best-first model merging for dynamical learning and recognition. *Advances in Neural Information Processing Systems*, 4:958–965, 1992.

[109] S. Gull. Bayesian inductive inference and maximum entropy. *Maximum-Entopy and Bayesian Methods in Science and Engeneering*, 1:53–74, 1988.

[110] M. Galassi. *GNU Scientific Library Reference Manual*. Number ISBN 0954612078. 3 edition. http://www.gnu.org/software/gsl/.

[111] N. Metropolis and S. Ulam. The monte carlo method. *Journal of the American Statistical Association*, 44:335–341, 1949.

[112] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1091, 1953.

[113] W.H. Press, S.A. Teukolsky, W.T. Vettering, and B.P. Flannery. *Numerical recipes in C++: The art of scientific computing*. Cambridge University Press, 2002.

[114] S. Kirkpatrick, C.D. Gelatt, and M.P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.

[115] D. Vanderbilt and S.G. Louie. A monte-carlo simulated annealing approach to optimization over continuous variables. *Journal of Computational Physics*, 56:259–271, 1984.

[116] S. Eddy. Multiple alignment using hidden markov models. *ISMB-95 Proceedings*, pages 114–120, 1995.

[117] S.P. Otto and T. Lenormand. Rsolving the paradox of sex and recombination. *Nature Reviews Genetics*, 3:252–261, 2002.

[118] M. M. Desai, D. Weissman, and M. Feldman. Evolution can favor antagonistic epistasis. *Genetics*, 177:1001–1010, 2007.

[119] W. Stephan. The rate of compensatory evolution. *Genetics*, 144:419–426, 1996.

[120] P.G. Higgs. Compensatory neutral mutations and the evolution of rna. *Genetica*, 102/103:91–101, 1998.

[121] D.B. Weissman, M.W. Feldman, and D.S. Fisher. The rate of fitness-valley crossing in sexual populations. *Genetcis*, 186:1389–1410, 2010.

[122] A. Altland, A. Fischer, J. Krug, and I.G. Szendro. Rare events in population genetics: Stochastic tunneling in a two-locus model with recombination. *Phys.Rev.Lett.*, 106:088101, 2011.

[123] D.T Gillespie. Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry*, 81:2340–2361, 1977.

[124] C. Escudero and A. Kamenev. Switching rates of multiple-step reactions. arXiv:0811.0090v1 [cond-mat.stat-mech].

[125] L. Dagum and R. Menon. Openmp: an industry standard api for shared-memory programming. *IEEE Computational Science and Engineering*, 5:46–55, 1998.

[126] A. Melbinger, J. Cremer, and E. Frey. Evolutionary game theory in growing populations. *Physical Review Letters*, 105:178101, 2010.

[127] M. Lynch. Scaling expectations for the time to establishment of complex adaptions. *PNAS*, 107:16577–82, 2010.

[128] H.A. Kramers. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7:284–304, 1940.

[129] J.E. Moyal. Stochastic processes and statistical physics. *Journal of the Royal Statistical Society B*, 11:150–210, 1949.

[130] D. Bedeaux, K. Lakatos-Lindenberg, and Shuler K.E. On the relation between master equations and random walks and their solutions. *Journal of Mathematical Physics*, 12:2116, 1971.

[131] N.G. van Kampen. Elimination of fast variables. *Physics Reports*, 124:69–160, 1985.

[132] C. W. Gardiner. Adiabatic elimination in stochastic systems. i. formulation of methods and application to few-variable systems. *Physical Review A*, 29:2814–2822, 1984.

[133] S. Nakajima. On quantum theory of transport phenomena- steady diffusion. *Progress of Theoretical Physics*, 20:948–959, 1958.

[134] R. Zwanzig. Ensemble method in the theory of irreversibility. *J. Chem. Phys.*, 33:1338–1341, 1960.

Erklärung

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit - einschließlich Tabellen, Karten und Abbildungen - , die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie - abgesehen von unten angegebenen Teilpublikationen - noch nicht veröffentlicht worden ist sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen der Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Prof. Dr. Alexander Altland betreut worden.

Köln, 17.10.2011

Teilpublikationen:

1. A. Fischer, C. Greenman and V. Mustonen. *Germline fitness based scoring of cancer mutations*, Genetics, 188: 383-393, 2011
2. A. Altland, A. Fischer, J. Krug and I. Szendro. *Rare events in population genetics: Stochastic tunneling in a two-locus model with recombination*, Physical Review Letters, 106: 088101, 2011

Danksagung

Diese Dissertation ist das Ergebnis einer dreijährigen Arbeit am Institut für Theoretische Physik der Universität zu Köln und am Wellcome Trust Sanger Institute in Cambridge, UK. Viele Menschen haben mich auf diesem Weg begleitet und unterstützt. An dieser Stelle möchte ich einigen persönlich meinen Dank aussprechen.
Vor allem danke ich meinem Betreuer Prof. Dr. Alexander Altland und meinem Gastgeber Dr. Ville Mustonen am Sanger Institut für ihre Geduld und Freundschaft.
Meine Familie hat mir das Studium ermöglicht. Dafür bin ich meinen Eltern Olga und Bernd, meiner Schwester Diana und meinen Großeltern Alrun, Wolfgang und Maria zu großem Dank verpflichtet.
Es war meine Freundin Katharina, die mich jeden Tag begleitet hat und mich allzu oft mit meiner Arbeit teilen musste. Ohne ihre Toleranz und Unterstützung wäre diese Arbeit nicht möglich gewesen. Sie ist und bleibt mein Anker und dafür bin ich ihr ewig dankbar.
Ich danke allen Kollegen und Freunden, die ich in der Zeit in Köln und Cambridge um mich hatte. Besonderer Dank gilt Anna und Rabia, Gregor und Fanny für das aufmerksame und kritische Lesen der Arbeit.
Letztlich danke ich der Bonn-Cologne Graduate School und der DFG für finanzielle Förderung.