Diss. ETH No. 20486

The Exploration of HIV Fitness Landscapes

A dissertation submitted to
ETH Zurich

for the degree of
Doctor of Sciences

presented by
João Zambujo Ramos Martins
Dipl.-Ing. INSA Lyon, France
born October 3, 1983
citizen of Portugal

accepted on the recommendation of
Prof. Dr. Sebastian Bonhoffer, examiner
Prof. Dr. Christoph Adami, co-examiner

2012

To my grandmother, Maria de Lurdes,
my dearest shortcut to myself.

Acknowledgments

This dissertation would not have been the same without the support of quite a few people.

Foremost, I thank my adviser, Sebastian Bonhoffer. Sebastian is the embodiment of Ken Kesey's idea of leadership. "You don't lead by pointing and telling people some place to go. You lead by going to that place and making a case." I will always remember his utmost mercy on the day I told him that I had mistakenly erased an hour of his work, as well as his complete lack of mercy each time he hits back a shuttlecock or a ping-pong ball.

Along with Sebastian, I would also like to express my gratitude to Chris Adami for inspiring and convincing me to pursue my studies and for agreeing to co-examine my defense. Chris is one of the most inspiring and enlightening persons I know. I am privileged to have met Chris and Sebastian.

I thank Trevor Hinkley, Roger Kouyos, and Gabriel Leventhal for their vital help in the several projects of this thesis, and sure enough, Rita Jenny for solving all the paperwork issues and for helping me organize the defense. I also thank Roland Regös and Jan Engelstädter for their answers to all the questions I could think of. Besides, I would also like to express my gratitude to Alex Gomperts and to Mathieu Legros for proofreading parts of this thesis.

In addition, I would also like to thank all those who shared their energy with me during my stay at ETH. In particular, I thank my family who always supported me and my close friend who have been like a family to me. I also thank all my work-mates (especially my office-mates), my housemates, my sport-mates, and my soul mates for being the reason of the great time I had during the past years.

Lastly, I will not forget the financial support from the Swiss National Science foundation and the computational support from the Brutus computer cluster.

Abstract

One of the aims of systems biology is to decode genetic sequences in terms of biological activity and phenotypic expression. In particular, to describe evolutionary processes, it is important to characterize the fitness of organisms as a function of the genetic space they can explore. In other words, it is important to characterize the fitness landscape on which evolution takes place. Many theoretical studies in evolutionary biology have assumed simplistic fitness functions to be able to study the evolutionary process. This thesis however, explores and describes a fitness landscape based on real *in-vitro* fitness of HIV experimental data.

Background knowledge about the data is presented in chapter 1. In addition, this chapter also introduces two individual mutation-based models of the fitness landscape used in chapters 2 and 3. One, the main effects (ME) model, includes the estimates of the fitness effects of individual amino acid variants; whereas the other, the main and epistatic effects (MEEP) model, also accommodates the estimates of the pairwise epistatic interactions. The details of the fitting and the performance of the models are presented in the appendix A. Most importantly, the two individual mutation-based models of the fitness landscape are a great tool to investigate the roles of epistasis and pleiotropy.

Chapter 2 explores three complementary visual representations of the fitness landscape. One provides a polynomial surface fitting of the experimental fitness values of the viral sequences, represented by points in a plane and placed such that the information about the number of amino acid mutations between the sequences is maximally conserved. The second representation renders the fitness landscape as a network where edges link neighboring sequences and the size of the nodes accounts for fitness. The third representation uses the MEEP model to generate a three dimensional fitness surface based on a grid of 1-mutation neighboring sequences incorporating the most frequent individual mutations. All three representations indicate a high level of local ruggedness and support Kauffman's *massif central* hypothesis which states that high fitness genotypes tend to be close to each other. Most importantly, they show that low-dimensional fitness maps can still capture important features of complex fitness landscapes.

Chapter 3 uses the ME and the MEEP fitness landscape models to simulate the evolution of HIV populations and study the maintenance of genetic recombination, one of the most intriguing problems of evolutionary biology. On the basis of simplistic models of fitness landscapes, it has been shown that the interaction between genetic drift and natural selection favors recombination independent of epistatic interactions. The ME and the MEEP

fitness landscape models therefore offer an unprecedented opportunity to bridge the gap between simplistic models and real fitness landscapes. Although recombination is shown to be still generally favored under the ME and the MEEP fitness landscape models, evolved HIV populations cannot be kept in realistic regions of the sequence space, and therefore it remains unclear whether genetic drift outweighs epistasis as a factor for the maintenance of recombination in a more complex and rugged fitness landscape.

Chapter 4 provides an exploratory analysis of the change of the HIV fitness values across different common drug environments. Specifically, chapter 4 presents a principal component analysis of the fitness data of the different conditions which reveals structure and patterns associated with drug resistance and cross-resistance. In addition, by comparison with patterns generated by simulated data, it was possible to quantify which part of the total variance of the original data was due to non-specific, drug-class-specific and drug-specific effects of resistance mutations. Accordingly, it was shown that relative fitness is mainly drug-independent and that drug-specific effects are significantly different between drug classes. Further comparison of the results with known combination therapies indicates that principal component analysis can identify effective drug combinations to minimize the risk of emergence of resistance.

At last, chapter 5 sums-up the most important difficulties and challenges encountered in the previous chapters and puts the main results in perspective. First and last, this thesis shows that a better understanding of HIV's evolutionary process leads to a better understanding of HIV drug resistance and the other way around.

Résumé

L'un des principaux objectifs de la biologie des systèmes est celui de décoder des séquences génétiques en termes d'activité biologique et d'expression phénotypique. En particulier, afin de mieux comprendre le processus évolutif, il est important de caractériser l'aptitude des organismes en fonction de l'espace génétique qu'ils peuvent explorer. Autrement dit, il est important de caractériser le paysage adaptatif où l'évolution peut avoir lieu. En biologie évolutive, de nombreuses études théoriques sont basées sur des fonctions de fitness simples afin de pouvoir étudier les processus de l'évolution naturelle. Cette thèse, en revanche, explore et décrit un paysage adaptatif qui est basé sur des données expérimentales de fitness *in-vitro* du VIH.

Quelques précisions sur les données sont présentées dans le chapitre 1. Ce chapitre introduit également deux modèles du paysage adaptatif utilisés dans la suite de la thèse. L'un permet de calculer la fitness d'un virus à partir de l'effet de chaque mutation individuelle (ce modèle est appelé ME) ; tandis que l'autre prend en compte non seulement l'effet de chaque mutation individuelle, mais aussi les effets des interactions entre les paires de mutations (ce modèle est appelé MEEP). Les détails sur l'ajustement et la performance des modèles sont présentés dans l'annexe A. Ces deux modèles constituent un outil sans précédent pour étudier le rôle des interactions de mutations (épistasie et pléiotropie) dans le processus évolutif.

Le chapitre 2 explore trois représentations visuelles complémentaires du paysage adaptatif. La première consiste en une surface de régression polynomiale sur les valeurs de fitness des séquences virales, qui sont représentées par des points dans un plan et placées de telle sorte que l'information sur le nombre de mutations entre les séquences est maximale. La seconde représentation montre le paysage adaptatif sous forme d'un réseau dont les arêtes lient des séquences voisines et la taille des nœuds représente la fitness de chaque virus. Enfin, la troisième représentation utilise le modèle MEEP pour générer une surface adaptative tridimensionnelle basée sur un treillis formé par des séquences espacées d'une mutation les unes des autres et intégrant les mutations les plus fréquemment observées. Ces trois représentations indiquent toutes un haut niveau de rugosité locale et soutiennent l'hypothèse du *massif central* de Kauffman qui stipule que les génotypes dont la fitness est élevée ont tendance à être proches les uns des autres. Elles montrent également qu'un espace de dimension réduite (deux ou trois dimensions) est suffisant pour retrouver les caractéristiques les plus importantes des paysages adaptatifs complexes.

Le chapitre 3 utilise les modèles ME et MEEP pour reproduire l'évolution des populations virales et, de cette façon, étudier le maintien de la recombi-

naison génétique qui constitue l'un des problèmes les plus intrigants de la biologie évolutive. Sur la base de modèles simples de paysages adaptatifs, il a été observé que l'interaction entre la dérive génétique et la sélection naturelle favorise la recombinaison indépendamment des interactions entre les effets des mutations. Les modèles ME et MEEP permettent donc de faire le lien entre des modèles plus simples et les vrais paysages adaptatifs. Bien que la recombinaison soit, de façon générale, aussi favorisée pour les modèles plus complexes tels que le ME et le MEEP, les populations virales n'ont pas pu évoluer et rester dans les régions réalistes de l'espace de séquences. Il n'apparaît donc pas encore clairement si la dérive génétique est toujours plus importante que les interactions épistatiques pour le maintien de la recombinaison dans le cas d'un paysage adaptatif qui soit plus complexe et plus rugueux.

Le chapitre 4 présente une analyse exploratoire de la variation des valeurs de fitness du VIH mesurées en présence de différents médicaments. Plus précisément, ce chapitre présente une analyse en composantes principales des données de fitness qui révèle la structure et les profils de résistance aux médicaments. En outre, par comparaison avec des données simulées, nous avons pu quantifier la partie de la variance totale des données d'origine due à des effets non-spécifiques et à des effets spécifiques à chaque classe de médicaments. En conséquence, il a été montré que la fitness est essentiellement indépendante de l'environnement (du médicament) et que les effets spécifiques aux médicaments sont significativement différents entre les classes de médicaments. Par la comparaison de ces résultats avec les thérapies connues, il s'avère que l'analyse en composantes principales permet d'identifier des combinaisons de médicaments efficaces pour réduire au minimum le risque d'émergence de résistance.

Enfin, le chapitre 5 fait un résumé des difficultés les plus importantes et des défis rencontrés le long des chapitres précédents et fait une rétrospective des résultats les plus importants. Avant tout, cette thèse montre qu'une meilleure compréhension des processus d'évolution du VIH conduit à une meilleure compréhension de la résistance du VIH aux médicaments et vice-versa.

CONTENTS

CHAPTER

ONE

INTRODUCTION

Needless to try—your knowledge is insufficient to allow me to explain you what I want.

Jorge Zambujo

1.1 The concept of fitness landscape

Fitness landscapes have been the subject of a somewhat philosophical debate (Kaplan, 2008; Ruse, 1991), not so much about the definition of the concept itself, but rather about Wright's three dimensional pictorial metaphor which he created to illustrate his so-called shifting balance theory in non-mathematical terms (Wright, 1932). The definition of a fitness landscape is rather straightforward: a fitness landscape is but the mapping between a set of genotypes and their corresponding fitness values (Kauffman, 1993). However, the comprehension and description of such a mapping for a large set of genotypes presents many difficulties and is the source of discussions about the nature of the evolutionary process itself (Provine, 1989). Access to large-scale sequence and fitness data has led to new ways of tackling a problem which has primarily been only philosophical (Schuster, 2012).

1.2 HIV data collection

Being the cause of AIDS, one of the world's largest pandemics, HIV has been having devastating social and economic consequences in the last three decades (Fauci et al., 2003). The research effort which was mounted to take control of HIV led to the discovery and approval of, on average, one antiretroviral drug per year (Clercq, 2009). The introduction of new drugs has been accompanied by the emergence of multiple resistance mutations, especially in the *pol* gene which encodes for two main target proteins of drug treatment (Bennett et al., 2009). With such a large number of available drugs and the associated risk of the emergence of drug resistance mutations, genotypic and phenotypic testing became indispensable to help clinicians to better plan HIV patients' specific therapies. As a result of the adoption of genotypic and phenotypic HIV testing, large datasets gathering both genotypic and phenotypic data have been collected. One of these datasets is owned by Monogram Biosciences, an American company previously known as Virologic, which again, provides individualized genotypic and phenotypic testing for HIV patients. In a 10-year research collaboration, Monogram Biosciences has shared over 70'000 amino acid sequences of the HIV-1 *pol* gene, along with their respective fitness in the absence and the presence of a minimum of 15 antiretroviral drugs.

1.3 The fitness assay

The *in-vitro* replication capacity (RC) quantifies the total production of virions in a single round of infection. Thus, the RC can be regarded as a measure of viral fitness for each single sequence for a given drug condition. To stop infection after a single round, a viral vector is constructed based on an NL4-3 HIV-1 clone, where a luciferase expression cassette is inserted at the place of the envelope gene to disable the possibility of further replication (fig. 1.1 *b*). In essence, patient derived amplicons containing the protease (PR) and the reverse transcriptase (RT) are inserted in the NL4-3 clone (fig. 1.1 *b*), so that HEK 293 cells are co-transfected both with the NL4-3 clone and with a murine leukemia virus (A-MLV) which contains the envelope gene to guarantee that the HIV clone is able to replicate only once (fig. 1.1 *c*). The RC (also referred to as fitness) is obtained by measuring and normalizing the luciferase activity relative to a NL4-3 reference virus of fitness 1. Fitness measurement errors have been estimated to be around 20% (Petropoulos et al., 2000).

1.4 Sequence data

In total, Monogram Biosciences shared 70'081 amino acid sequences of HIV-1 subtype B PRs and RTs derived from patients receiving multi-drug combination therapy. The sequences have a total length of 404 amino acids: 99 correspond to the full length of the PR and 305 to the first part of the RT. Both the PR and the RT lie next to each other in the *pol* gene and are essential enzymes for viral replication. For this reason, they have been the two main targets of anti-viral drug therapy and most of the mutations associated with drug treatment occur in the first 400 residues of the *pol* gene (Bennett et al., 2009). In essence, the PR is a homodimer of 2 times 99 residues long and plays an essential role in the maturation of viral particles budding from the cell's membrane. It cleaves the Gag and the Gag-Pol poly-proteins into smaller core proteins (MA, CA, NC, p6, PR, RT and IN of Fig. 1.1 *b*) encoded by these two genes. Without the PR processing of these two protein precursors, the virus cannot be infectious. Another important aspect is that viral assembly and maturation are highly coordinated. This means that small changes in PR activity may induce drastic changes in viral fitness. For a three-dimensional representation of the structure of a PR dimer please refer to figure A.3 of the appendix A. The RT is a heterodimer responsible for the reverse transcription of the viral RNA into the DNA duplex. The bigger subunit (p66) has 560 residues and contains the two catalytic domains of the

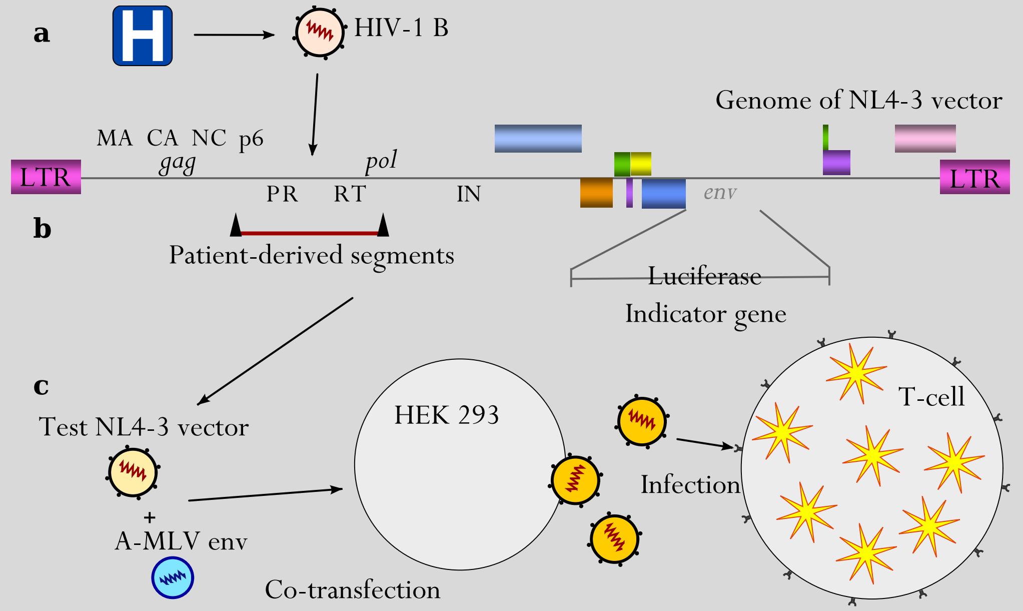


Figure 1.1: Schematic illustration of the processes of data collection and of the fitness assay. **a.** Collection of HIV-1 sequences from blood samples of HIV infected patients. **b.** Insertion of a patient's derived segment into NL4-3 vectors containing a Luciferase as an indicator gene instead of the full *env* gene. **c.** Co-transfection to HEK 293 using a helper virus (A-MLV *env*) to substitute the missing *env* gene; the full round of infection is completed with the infection of T-cells which is followed by light emission associated with the indicator gene.

molecule, a polymerase triad (Asp 110, Asp185 and Asp186) and a ribonuclease H (RNase H) active site. The first domain takes single stranded viral RNA and transcribes it into a RNA/DNA hybrid double-helix. The second cleaves the RNA from the hybrid and finally, the polymerase again completes the remaining DNA to allow the integration of the DNA double-helix into the host cell genome. Even though both subunits arise from the same amino acid sequence, the smaller has 440 residues and is arranged differently: in a closed conformation that deactivates RT's catalytic sites (Frankel and Young, 1998). Typically, combination therapies have consisted of PR inhibitors (PIs) in association with RT inhibitors (RTIs). (The RTIs can still be classified in two classes, either as nucleoside analog reverse transcriptase inhibitors (NRTIs) or as non-nucleoside reverse transcriptase inhibitors (NNRTIs)). Both RTI classes inhibit the polymerase active site (Beerenwinkel, 2004). For further details regarding the different drugs, please refer to chapter 4.

1.5 Individual mutation-based models

Two individual mutation-based models have been developed to estimate the fitness effects of 1'857 single mutations and of 257'536 pairs of mutations which were found in Monogram Biosciences' 70'081 amino acid sequences (see the appendix A). The fitness w of a sequence i was calculated as

$$w_i = exp(b_0 + \sum_{j=1}^{N_M} M_{ij}\gamma_j + \sum_{k=1}^{N_E} E_{ik}\chi_k)$$

where b_0 stands for the intercept, γ_j the estimated main effect of the j^{th} mutation M_{ij} and χ_k estimates the epistatic interaction of the k^{th} combination of mutations E_{ik}. One variant of the individual mutation-based models assumed no epistasis ($\chi_k = 0$) and was referred to as the main effects (ME) model, and the second variant also included pairwise epistasis, and was therefore referred to as the main effects and epistatic effects (MEEP) model. The fitting of the individual fitness estimates was obtained by means of a Generalized Kernel Ridge Regression (GKRR) which is a computational method of the family of the support vector machines and well suited for cases where the number of parameters to be estimated exceeds by far the number of experimental observations (number of available sequences). (A full description of the individual mutation-based models and of the GKRR is found in the supplementary material of Hinkley et al. (Hinkley et al., 2011).) Chapter 2 uses the MEEP model and chapter 3 uses both the ME and the MEEP variants. The details of the performance of the models in the different drug environments is found in the appendix A.

CHAPTER

TWO

REPRESENTATIONS OF THE FITNESS LANDSCAPE IN FEW DIMENSIONS

> Alle sagten: das geht nicht.
> Dann kam einer, der wusste das
> nicht und hats gemacht.
>
> Unbekannt

Contents

Submitted as: Martins, JZR; Leventhal, G; Hinkley, T; Kouyos, R; Haddad, M; Whitcomb, JM; Petropoulos, CJ; Bonhoeffer, SL. ***Visual representations of an HIV derived fitness landscape.***

Abstract

Fitness landscapes are generally high-dimensional and therefore hard to depict and conceptualize. Here, we explore three approaches to visually capture the main properties of a large fitness landscape derived from HIV sequences which were assayed for *in-vitro* fitness. First, we apply a multidimensional scaling to data consisting of 4’000 HIV *pol* amino-acid sequences to obtain an approximation of the sequence space in 2D, which we use to create a 3D smooth trend surface of the fitness landscape by means of a polynomial surface of *in-vitro* fitness values. Second, we build a network-based representation of the fitness landscape where edges link neighboring sequences and the size of the nodes represent the fitness values. Third, we apply a model which estimates the fitness effects of single and double mutations to calculate a 3D fitness surface of a discrete sequence space consisting of a mesh of one mutation neighboring sequences. Our results show evidence for a high degree of local epistasis and biophysical constraints, as well as empirical support for Kauffman’s *massif central* hypothesis which states that high fitness genotypes tend to be close to each other. Overall, this chapter shows that low-dimensional representations can capture important features of complex fitness landscapes.

2.1 Introduction

In 1932, Sewall Wright presented the heuristic concept of fitness landscape to explain his Shifting Balance theory (Wright, 1932). More specifically, Wright plotted maps of adaptive values on what he called "fields of gene combinations", nowadays referred to as the genetic space. These maps were later used to define the concept of fitness/adaptive landscapes (Gavrilets, 2004; Kauffman, 1993; Stadler, 2002). If the number of gene combinations, and thereby fitness landscape, is sufficiently small it is possible to map the fitness of each gene combination such that the representation is intelligible (Wiles and Tonkes, 2006; Wright, 1932). However, if the fitness landscape is large, then the genetic space is high-dimensional. As a consequence, it is difficult to get a legible map of all the gene combinations to their fitness values in two dimensions.

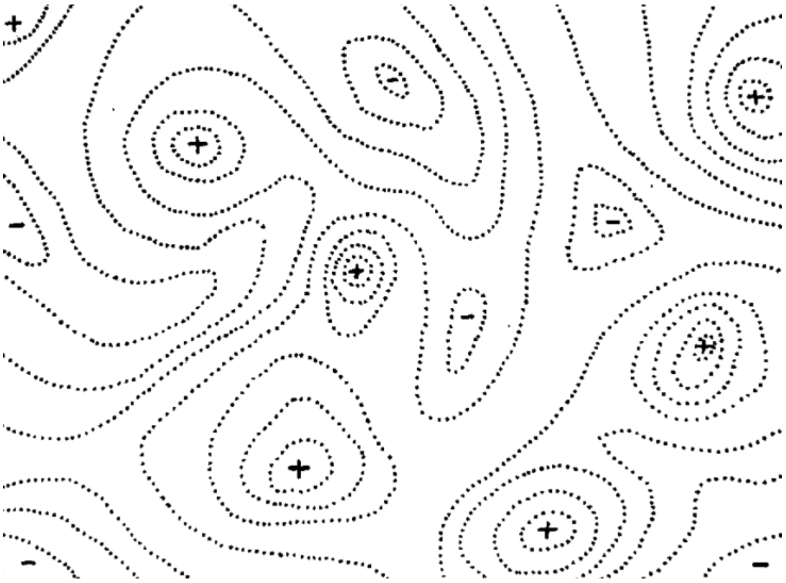


Figure 2.1: Wright's original diagram of a fitness landscape (Wright, 1932).

Wright argued that a very large number of pairwise genetic distances could be approximately represented by an equally very large number of Euclidean distances in two dimensions, on top of which he assumed and drew a continuous, smooth fitness surface (see fig. 2.1). As in topographic maps, Wright's representations showed peaks, ridges, and valleys, and served to intuitively illustrate the mathematical results of his Shifting Balance theory. Accordingly, populations tend to evolve to the peaks of high fitness of the landscape which are separated by valleys of low fitness, and the topography of the landscape determines how many peaks exist and how accessible they are. The intuitive nature of these representations made them common in textbooks and in research papers on evolutionary biology (Barton et al., 2007; Orr, 2009; Poelwijk et al., 2007; Smith et al., 2002). Yet, the over-simplifying assumptions inherent to these representations and the lack of methods to produce them led to an ongoing epistemological debate on whether Wright's representations should be kept or abandoned (Kaplan, 2008; Provine, 1989; Ruse, 1991; Skipper, 2004). On the one hand, for very high-dimensional spaces, the concept of peaks and valleys is likely to be meaningless (Gavrilets, 2004). On the other hand, except for recent studies such as (Gavrilets, 1997; McCandlish, 2011), there were no real attempts to create and test concrete applications of these representations. However, the development of multivariate analysis and spatial statistics, the increased availability of computing power, and the access to large sequence and fitness data sets constitute an unprecedented opportunity to finally investigate their application to experimental data.

In this chapter we first implement a set of methods to construct a Wrightean diagram using data which consists of amino acid sequences and the corresponding fitness values. The data have been described in detail in chapter 1 and in the appendix A. Specifically we consider three approaches to represent fitness landscapes. The underlying idea of our first approach is similar to the one proposed by McCandlish (McCandlish, 2011), in the sense that we seek a representation of the sequence space in two dimensions. McCandlish analyzes the eigenvalues and eigenvectors of evolutionary transition matrices, which best suits populations evolving in a same environment and undergoing small mutation rates (McCandlish, 2011). HIV populations, however, are characterized by high mutation rates; therefore, this method does not apply. Instead, a non-metric multidimensional scaling (MDS) is a more suitable method to obtain a scatter plot that maximally conserves the information about the distance between HIV sequences. We measure the distance between sequences by counting the number of substitutions. This distance is commonly referred to as Hamming distance (HD). We use MDS to produce a scatter plot of the sequence space, which we then use to fit the fitness surface by trend surface analysis; i.e. the surface is obtained by fitting a low degree polynomial surface to the fitness values of every HIV sequence (Li et al., 2000; Venables and Ripley, 2002).

As any smooth surface most likely misrepresents some aspects of high-dimensional landscapes (Gavrilets, 2004; Kouyos et al., 2012), we also investigate the fitness landscape using a network-based representations. As in the works of McCandlish (McCandlish, 2011) and of Ashlock and Schonfeld (Ashlock and Schonfeld, 2005), we plot fitness values at the nodes of a network, with the edges linking the closest sequences in terms of number of mutations. Finally, we use a model that estimates the effect of both main and epistatic effects of mutations (see appendix A) to plot and to compare the fitness surface of a regular mesh of one-mutation neighboring sequences with the previous representations.

All in all, we find good qualitative agreement between the three representations in terms of the overall fitness distribution and local ruggedness along the sequence space.

2.2 Materials and methods

This section consists of five parts. In part one, we explain how the subset of sequences analyzed here was chosen from the larger data set described in chapter 1 and in the appendix A. In parts two and three, we briefly present the mathematical methods that allow us to produce a Wrightean-like diagram from sequence and fitness data. In parts four and five, we present two alternative methods to also visually inspect the fitness landscape.

2.2.1 The sequence alignment

We consider a subset of 4’000 unique amino acid sequences from the data set described in chapter 1 and in the appendix A, which contains 70’081 virus samples derived from HIV-1 infected patients. The sequenced region contains all of the protease and the first 305 amino acid positions of the reverse transcriptase. We chose a subset of 4’000 sequences for the following reasons: the majority of the original set of sequences contains unresolved positions indicating that the sampled virus population was polymorphic at these positions. Such unresolved positions can induce significant errors in the calculation of the HD. To minimize this problem, sequences with more than 2.5% of unresolved positions were excluded from the sample. The choice of this threshold offers a “level-headed” compromise between the number of available sequences and the sequence quality. In addition, to obtain a more uniform distribution of the sampled sequences across the range of HD, we force the distribution of the HD between the sampled sequences and the consensus sequence of the alignment to be uniform. Thus, we choose 100 sequences, at random, for each class of HD, in the range of 3 to 42 substitutions to the consensus sequence (the maximum range containing at least 100 sequences per class of HD). Hence, we cover an important part of the range of possible classes of HD of the original alignment, with as many sequences as possible, while respecting the constraint of an equal number of sequences per class of HD. (For a detailed description of the sequence data and of the fitness assay, see chapter 1) Note that a sample size of 4'000 sequences implies 7'998'000 pairwise HD. This number is still tractable by numerical optimization methods such as MDS.

2.2.2 Multidimensional scaling of pairwise Hamming distances

We use a non-metric MDS to maximize the correlation between the sequences' pairwise HD and their pairwise Euclidean distances between the sequences in a xy plane. The purpose of MDS is to find a configuration of points in a reduced dimensional space, here a two dimensional plane, such that the distances between these points best match those of the original pairwise distance matrix. Put differently, we use MDS to create a perceptual two dimensional map of the sequences' genetic space, in order to display the fitness values on a configuration of points that is optimized according to the pairwise HD between sequences. In essence, MDS is a numerical optimization technique that starts with a prior configuration of axis and points. This prior configuration is typically given by principal coordinate analysis of the original distance matrix (as it was the case in this chapter). MDS then incrementally improves the initial configuration of points, by moving the positions of the points by small amounts, and by choosing the new configuration that will gradually increase the goodness of fit to the initial distance matrix (Kruskal, 1964). The goodness of fit of a new configuration is typically given by Kruskal's stress function S which is written as follows,

$$S = \sqrt{\frac{\sum_{i<j}^{n} (d_{ij} - \hat{d}_{ij})^2}{\sum_{i<j}^{n} d_{ij}^2}}.$$

Here, d_{ij} is the initial distance between two points, point i and point j, and $\hat{d}_{ij}$ the corresponding distance in the configuration of points that is being optimized. For a detailed mathematical description of MDS please refer to (Cox and Cox, 1994). The method has been implemented in the `isoMDS` function which is part of the `R:MASS` package (R Development Core Team, 2009; Venables and Ripley, 2002).

2.2.3 Spatial interpolation of the fitness landscape

We use the scaled approximation of the sequence space which we obtain from the MDS to map the fitness of the sequences on a three dimensional surface (i.e. x and y coordinates span the sequence space, and the z axis to the fitness values). To represent this surface, we fit a polynomial regression surface, also known as trend surface, to the entire set of points. The trend surface is fitted by least-squares, a method described in detail in (Ripley, 1981) and implemented in the function `surf.ls` of the `R:spatial` package (Ripley, 1981; Venables and Ripley, 2002). Put differently, the procedure consists in fitting polynomials to the fitness values and the sequences' two dimensional coordinates of the sequence scaled space which is given by MDS of the sequences' pairwise HD.

2.2.4 Network-based representation

A limitation of MDS is that it is computationally expensive to optimize over many data points. An intermediate step between the pairwise HD and the two-dimensional coordinates produced by MDS is the distance matrix. This $N \times N$ matrix of pairwise HD can be interpreted as the adjacency matrix of a fully connected weighted graph. The two dimensional representation then is a graph drawing problem and in fact many graph drawing algorithms are variants of MDS. Within this graph representation it is now possible to reduce the complexity by deleting edges in order to highlight important relationships between sequences and ignore those that contain little information. For example, if we are most interested in those sequences that are most similar to a particular sequence then it may be beneficial delete all edges except for the K closest neighbors. Another example would be if we are interested in the density of close neighbors. In that case we would delete all edges that are above a certain HD threshold, retaining connections between similar sequences.

Once the reduced graph is constructed, it is then possible to draw the graph using MDS (or other graph drawing algorithms (Fruchterman and Reingold, 1991; Kamada and Kawai, 1989)), but only optimizing over a subset of edges rather than the complete distance matrix.

2.2.5 One-mutation neighbors mesh

We also construct a one-mutation neighbors mesh of the fitness landscape to investigate its shape. The mesh is defined by discrete two-dimensional coordinates (i, j) of one-mutation neighboring sequences. Since we lack experimental measurements for all the sequences in the mesh, we use estimated fitness values based on the MEEP model as described in the appendix A. This model explains approximately 45% of the variance of the experimental fitness measurements (see figure A.1 of the appendix A). The mesh represents a section through the fitness landscape. To ensure maximal accuracy of the estimates and the coverage of the most plausible region of the sequence space, we place the consensus sequence of the protein alignment at $(0,0)$, the center of the mesh. At $(1,0)$, $(0,1)$, $(-1,0)$, and $(0,-1)$, we place four sequences differing of only one mutation from the consensus (labeled in blue in figure 2.2). Once more, to ensure a maximal accuracy of the estimates, the mutations are chosen according to their frequencies of occurrence in the alignment, Each sequence is mutated at a different locus to ensure that the pairwise HD between sequences can be given by the axial directions of the mesh and to prevent double and multiple mutants on the same locus. At $(2,0)$, $(0,2)$, $(-2,0)$, and $(0,-2)$, we place another four sequences, but differing by two mutations from the consensus, and we repeat this procedure until we reach thirty mutations from $(0,0)$. At $(1,1)$ we place a double mutant that contains the single mutations of both $(1,0)$ and $(0,1)$. Again, we repeat this operation until the entire mesh is filled up (see figure 2.2).

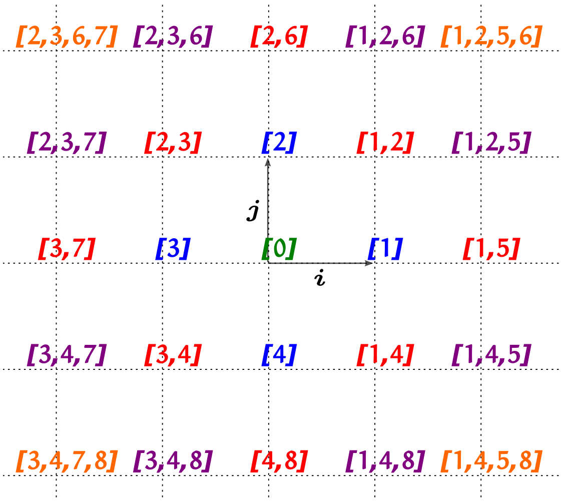


Figure 2.2: Schematic illustration of a one-mutation neighbors mesh of axis i, j. Green indicates the position of the consensus sequence. The colors blue, red, violet, and orange indicate the position of sequences with one, two, three, and four substitutions to the consensus sequence, respectively. The numbers in brackets designate the order of choice of new mutations. Number one corresponds to the most common mutation found in the sequence alignment; number two to the second most common mutation, and so forth.

2.3 Results

We use three complementary approaches to visualize an HIV derived fitness landscape. First, we consider a method based on multi-dimensional scaling, second, a network-based representation, third, a one-mutation neighbors mesh based on predicted fitness values.

2.3.1 Spatial interpolation of the fitness landscape

We construct a Wrightean fitness surface in two steps. First, we use MDS to find a two dimensional point configuration optimizing the pairwise HD between the sequences. Second, we fit a three dimensional surface by interpolating the fitness values of the sequences with a polynomial trend surface.

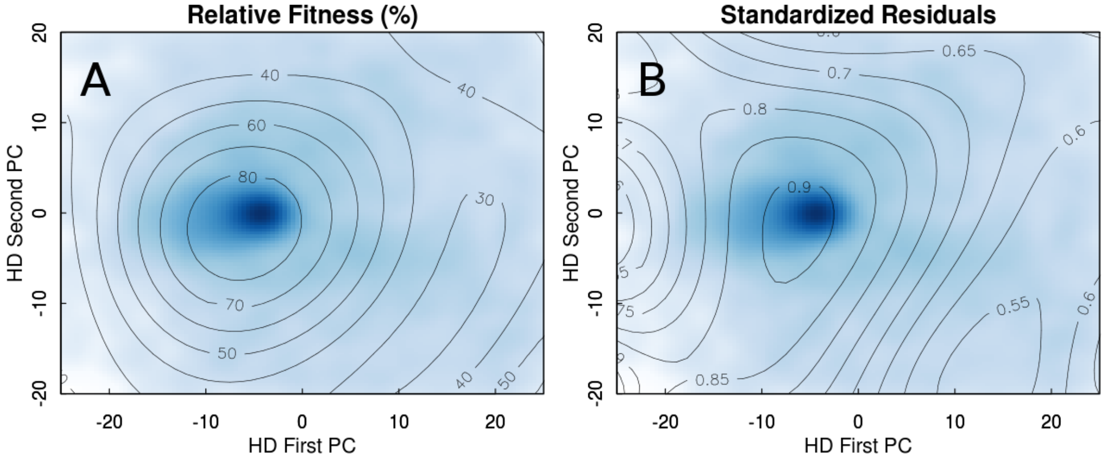


Figure 2.3: 6-degree polynomial interpolation by least-squares of the fitness of 4'000 unique sequences placed on a two dimensional MDS space of pairwise HD. The fitness surface is given by the isolines, in panel **A**. The sequence density in the two dimensional MDS space of pairwise HD given by levels of blue. Dark blue stands for regions of high sequence density, light blue for regions of low sequence density. In panel **B** are depicted the isolines of an interpolation of the absolute values of the standardized residuals of the surface in panel **A**.

In the first step, we find that MDS provides a surprisingly good representation of the sequence alignment in two dimensions. The Pearson correlation coefficient between the pairwise HD and the multi-scaled pairwise HD reaches 0.85, after 500 iterations of the MDS algorithm. The R^2 between these distances indicates that a two dimensional configuration of points can explain up to 72% of the total variance of the sequence space. The plane represents a sequence space of approximately 50×40 pseudo substitutions.

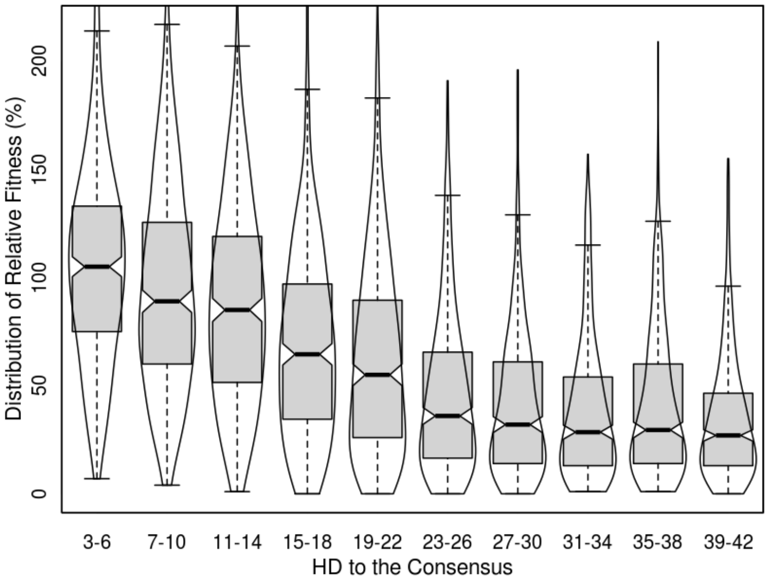


Figure 2.4: Boxplots showing the median and the quartiles of the fitness distribution of classes of 400 sequences of increasing HD to the consensus. The area between the vertical curves that is superimposed to the boxplots gives the kernel density estimation of the fitness distribution of for each class of HD. The fitness corresponds to the relative replicative capacity of sequences relative to that of an NL4-3 based control virus.

In the second step, we use a polynomial regression surface of low degree to fit the sequences' fitness values and capture the overall fitness trend over the sequence space. A smooth trend typically implies a low degree polynomial fit. We thus find that degrees 5 and 6 are low enough to capture the overall fitness trend and high enough to eliminate the artefactual slopes due to fewer data points at the borders of the surface (see figure 2.7). The fitted surface is depicted in figure 2.3A for the polynomial interpolation of degree 6.

Despite of the fact that the sequence alignment contains an equal number of sequences of each class of HD to the consensus, the density of points (given by the levels of blue in figure 2.3) is manifestly higher in the area of higher fitness than in the rest of the plot. (This can be partly explained geometrically, for the simple reason that the surface of concentric circles increases quadratically with HD, whereas the number of points increases linearly, given the uniform sampling along the range of HD.) The residuals associated with the fit are also manifestly larger in this area of higher fitness and point density (figure 2.3B). Hence, most of the variation in the data is poorly represented by the surface, especially in the area of higher fitness. This indicates a high local ruggedness in this region and is in line with the distribution of the variance of the fitness values across the different classes of HD, which is shown in figure 2.4. Also in line with figure 2.3, both the average fitness values and their variances decrease with the accumulation of mutations from the consensus sequence (figure 2.4).

Network-based representation

The most straightforward way to reduce the number of edges in the graph is to keep a link between two nodes (i.e. sequences) if their respective HD is less than some threshold value θ_{HD}. When θ_{HD} is too small, however, this method will result in largely disconnected graphs, as many of the sequences will be more than θ_{HD} away from their neighbors for sparsely sampled sequence spaces. This also strongly depends on the sampling distribution in sequence space. Since the sampling density is highest close to the consensus sequence for the HIV data, this method would then preferentially connect nodes that are close to the consensus sequence, whilst leaving nodes far away from the consensus sequence disconnected (figure 2.8).

Another way to reduce complexity is to keep only those edges, which lead to the K closest neighbors (i.e. smallest HD) of each sequence, ignoring duplicate edges. Each node in the resulting graph will then have at least K neighbors. This graph is more of a qualitative representation of how sequences would evolve when moving across the landscape if we assume that it is easier to reach sequences that are close, but disregards the number of mutations required to reach the neighbor sequence. In this sense, nodes with a high degree (i.e. number of neighbors) are accessible from many other nodes in the sample, while nodes with a low degree are only accessible from few sequences.

When $K = 1$, each node is only connected to one other sequence. The resulting graph is disconnected, with several clusters of sequences that are closer to each other than to other sequences. When $K = 2$ (Figure 2.5A), the graph becomes connected, but both high and low fitness nodes are spread out on the plane. This shows that if we just consider the closest neighbors of a particular sequence, we do not see any type of correlation between high and low fitness sequences. As K is increased the high fitness nodes (blue) start to move closer to each other, though they are interspersed with low fitness nodes (red). This is compatible with the main results obtained with both

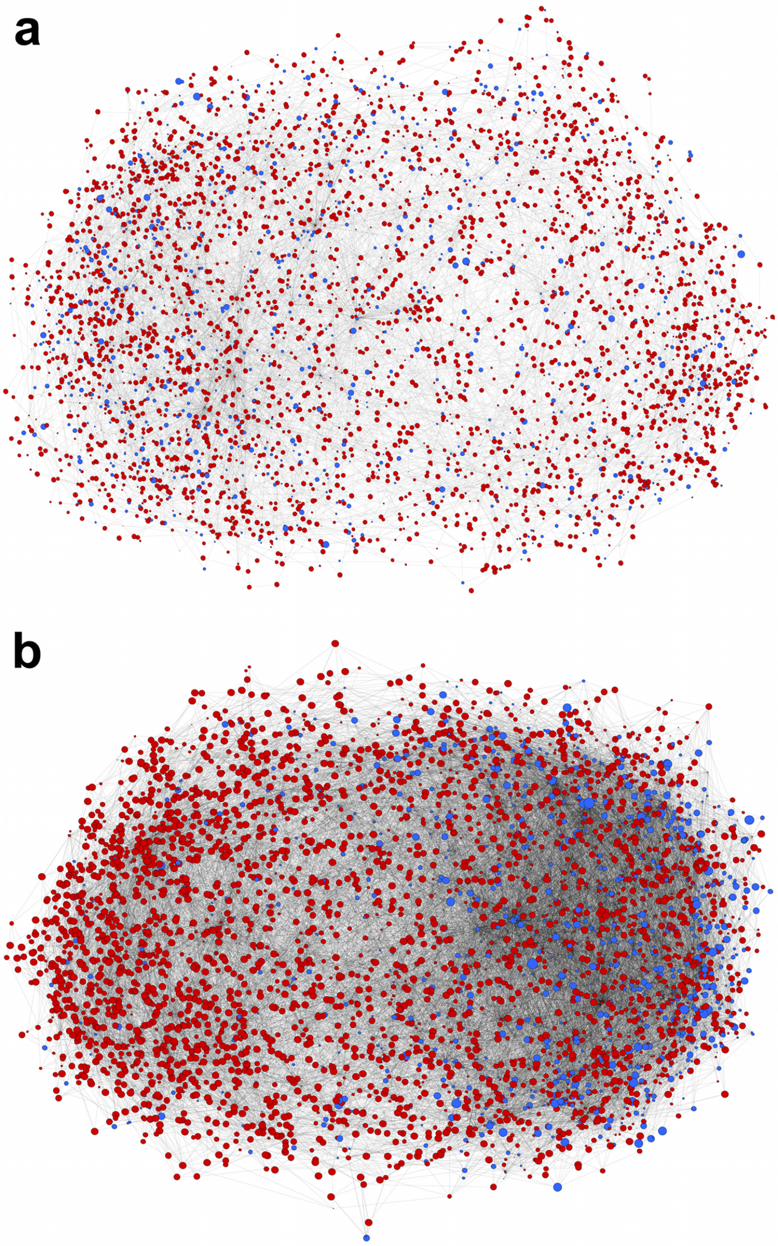


Figure 2.5: Reduced graphs using the nearest-neighbor reduction scheme for $K = 2$ (A) and 5 (B) closest neighbors and $N = 4100$ sequences. The graphs were drawn using the Graphviz sfdp algorithm (Gansner and North, 2000). Blue nodes are sequences that have a higher fitness than the reference sequence, red node sequences with a lower fitness. The size of the node is representative of the absolute value of the relative fitness (courtesy of Gabriel Leventhal).

the spatial interpolation and the one-mutation neighbors mesh, in the sense that there clearly is a general region of high fitness, but that this region is locally rugged.

One-mutation neighbors mesh

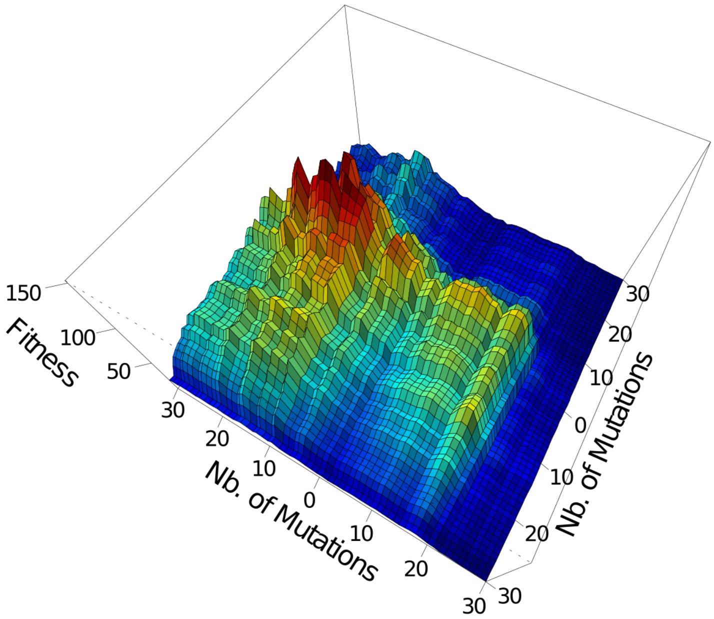


Figure 2.6: Surface view of the predicted fitness of a one-mutation neighbors mesh of sequences accumulating the most frequency mutations from the consensus sequence. The xy axis show the number of mutations to the consensus and the z axis shows relative fitness (%), where blue stands for low and red for high values of relative fitness to that of an NL4-3 based control virus. The fitness based on a model including main and epistatic effects of mutations (see the appendix A for a detailed description of the model).

The construction of a one-mutation neighbors mesh allows a direct three dimensional representation of a section through the sequence space. The construction of such a mesh is illustrated in figure 2.2. More specifically, we draw thirty unique mutations for each of their axial directions of the mesh. In total, the mesh consists of 61×61 unique sequences which always differ in a single mutation to the four adjacent sequences.

This surface covers about the same order of magnitude as the domain of HD as it can be found in the original sequence alignment. The sequences are placed by order of frequency of occurrence of their mutations from the center $(0,0)$, where the consensus sequence is placed, to the edges of the mesh. The sequences that do not directly lie on the axes result from the combination of the mutations of the sequences that lie on the axes (see figure 2.2). The sequences at the four corners of the mesh thus accumulate the greatest number of substitutions — sixty mutations to the consensus sequence.

Given that there are no experimental measurements for most of the sequences in the mesh, we use a mode which predicts fitness for the amino acid sequences. This fitness is given by the MEEP model as described in the appendix A. The overall predictive power of the model that is used to estimate the fitness of sequences has been shown to be approximately 45% based on an independent cross-validation data set of 5'000 sequences (see figure A.1 of the appendix A).

Figure 2.6 shows the surface formed by the fitness estimates of the sequences that constitute the mesh of one-mutation neighbors. This representation of the fitness landscape reveals several local optima. As observed in figures 2.3 and 2.5, the ruggedness of the surface tends to be higher in the region of higher fitness (in red). Interestingly, the consensus sequence (the center of the mesh) is located a few mutations away from the region of highest fitness. All the four corners of the mesh (i.e. sequences with high numbers of mutations) have nevertheless very low fitness (in blue). Overall, fitness decreases with the accumulation of mutations from the center of the mesh.

2.4 Discussion

All our results show some structure in the distribution of the average fitness across the sequence space in a way that suggests that local fitness peaks are more likely to be located close to each other, supporting the *massif central* hypothesis (Kauffman, 1993). This hypothesis was corroborated for the case of a theoretical fitness landscape (Østman et al., 2010), and our results now also offer empirical support from a biologically realistic derived HIV fitness landscape. A difference to theoretical fitness landscape studies is that one has no control regarding the locations of the sequences relatively to the fitness optima. It is unlikely to observe any sequence at a fitness optimum of the landscape because, although the fitness values result from in-vitro measurements, the viruses were directly obtained from patients. Thus, the fitness values were measured in an environment that is different from that in which they evolved.

Applying MDS to visualize fitness landscapes has several shortcomings.

As mentioned before, one limitation of MDS is that it is computationally demanding as it uses all the combinations of pairwise distances between points, which means it works best for a limited number of data points. Yet, we found that a two dimensional approximation of the sequence space retained as much as 72% of its original variance in terms of pairwise HD between sequences. This finding means that most of the information concerning the pairwise HD can still be retained in only two dimensions, despite of the drastic cut in the number of dimensions. Such high value of conserved variance may be due to several selection conditions on which sequences evolved prior to the drug free environment on which the fitness measurements were taken. If sequences evolved in such a way that they express only a few main phenotypic traits, then there is a sampling bias that reduces the complexity of the sequence space. Biophysical constraints (constraints in the secondary and tertiary protein structures) and pleiotropic constraints have been shown to limit the degree of mutational freedom of sequences (Lozovsky et al., 2009; Weinreich et al., 2006) and might thus constitute another non-exclusive explanation for such high value of conserved variance. These constraints can lead to "holes" in the sequence space (Gavrilets, 1997). If the number of "holes" is sufficiently high, then it is possible that most of the complexity of the sequence space can be captured in only a few dimensions.

It is important to note that the sequence space has a clear interpretation in terms of neighborhood — two sequences that are close to each other differ only by few amino acids. In a MDS space, however, this is only true in a statistical sense — sequences that are close to each other likely have only a few amino acids difference, but this need not to be the case. Hence, neighborhood has a very different interpretation.

A low-degree polynomial interpolation of the sequences' fitness values yields a coarse-grained picture of the fitness distribution along the MDS approximation of the sequence space. We observe that, on average, the sequences in a focal area of the sequence space (closer to the consensus) tend to be fitter than the sequences at the periphery. The observation that the goodness of fit is the lowest also in the focal area of the sequence space, suggests that the ruggedness (the degree to which fitness changes within a few mutations) of the landscape is higher in the region of high fitness.

Due to the discrete nature of the sequence space, graph theory provides a good framework to represent the fitness landscape (Ashlock and Schonfeld, 2005; McCandlish, 2011; Stadler, 2002). We have further reduced the complexity from the full distance matrix by only including links between the closest neighbor sequences. Only including the most immediate neighbor sequences results in a relatively even distribution of high and low fitness sequences across the two dimensional plane. When including links between medium distance neighbors the high fitness sequences begin to cluster together. Thus the most immediate neighbors don not have similar fitness values, but the intermediately distanced neighbors of high fitness sequences also have a high fitness. This is in support of the MDS analysis, in the sense that fitness peaks tend to be close to each other, but this region of high fitness is interspersed with many low fitness values, indicating high ruggedness.

For a more local analysis of the surface, we constructed a one-mutation neighbors mesh of estimated sequence fitness values based in the MEEP model described in the appendix A. Although fitness values are estimated on potentially unrealistic sequences (due to biophysical constraints), the fitness model filters part of the experimental noise that can otherwise be confounded with ruggedness. Again, the one-mutation neighbors mesh shows that ruggedness tends to be higher in the region of higher fitness, in accordance with the other methods.

The analyses presented in this chapter also point out the limitations of any reductive representation of a fitness landscape. However, if the sequence space does have a much lower complexity than what is usually assumed, then such representations can reveal the main features of fitness landscapes, such as ruggedness and overall fitness distribution. This chapter suggests that a good mathematical model for biologically realistic fitness landscape should ideally not only account for high ruggedness (such as highly epistatic NK landscapes (Kauffman and Weinberger, 1989)), but also integrate a region of higher fitness and highly constrained/pleiotropic sequence spaces (Weinreich et al., 2006).

Overall, our analyses support the *massif central* hypothesis, at least for HIV, and thereby underlines the utility of simple representation of fitness landscapes.

2.5 Supplementary figures

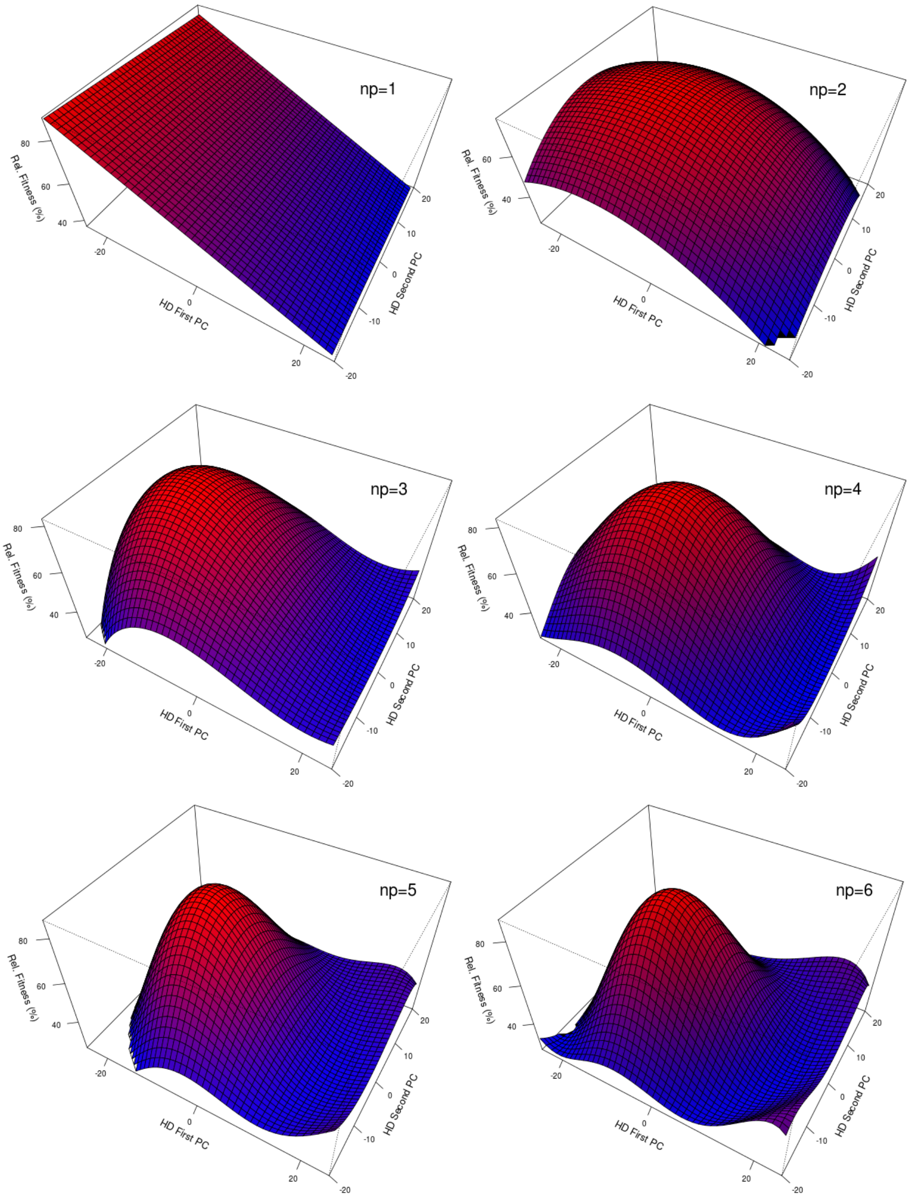


Figure 2.7: Interpolated polynomial fitness surfaces by least-squares on the genetic scaled spaces by MDS. The six different figures show the result of the interpolation for polynomial degrees (np) from 1 to 6. Blue regions correspond to domains of low relative fitness and red for regions of high relative fitness. The contour plot that corresponds to the interpolation with np $= 6$ is shown in the left panel of figure 2.3.

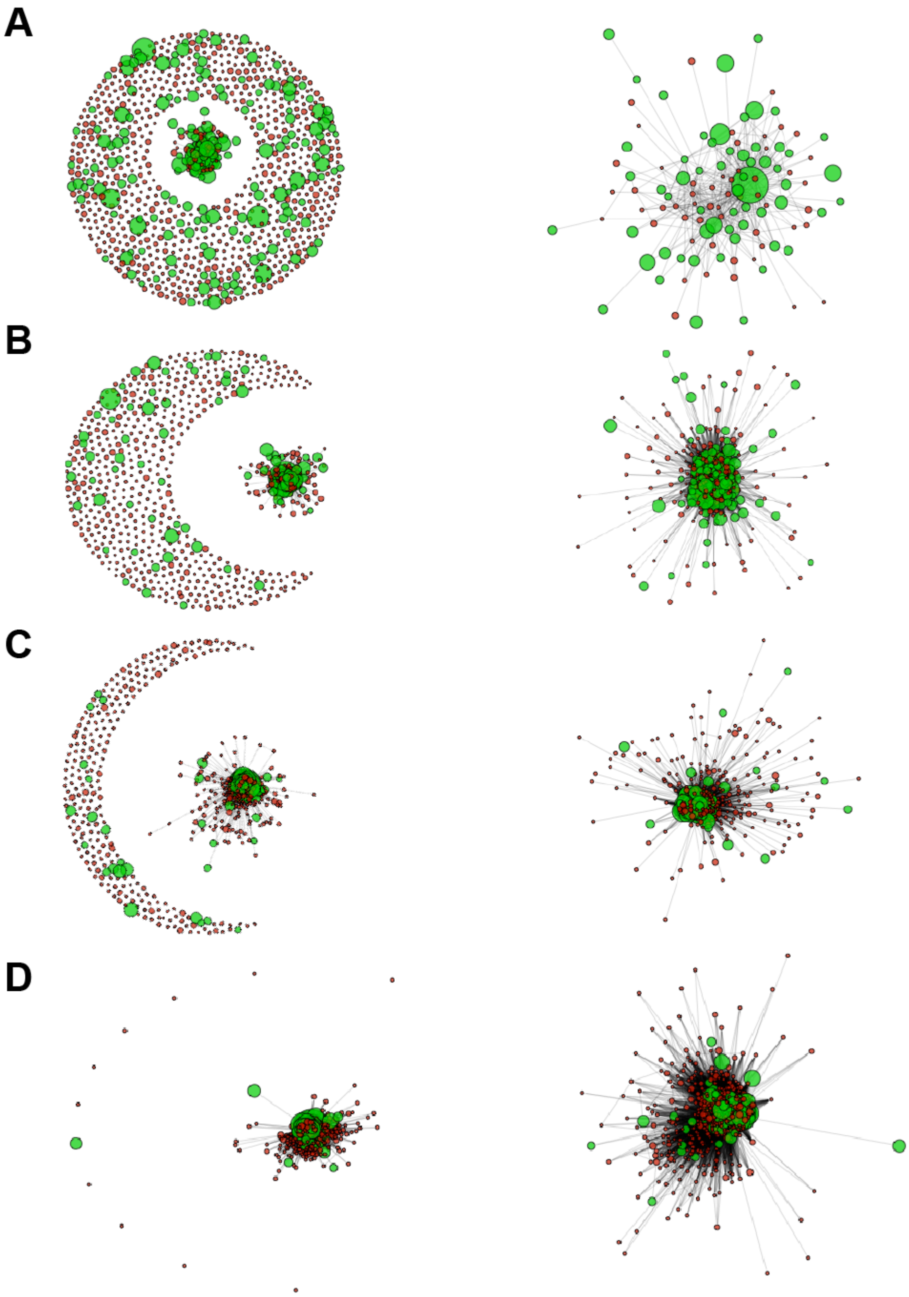


Figure 2.8: Network with nodes connected that are less than $\theta_{HD} = 10$ (A), 20 (B), 30 (C) and 40 (D) HD apart. The graph on the right are close-up of the largest component in the four cases (courtesy of Gabriel Leventhal).

CHAPTER THREE

ODDS FOR THE EVOLUTION OF RECOMBINATION

> We have to remember that what we observe is not nature itself, but nature exposed to our method of questioning.
>
> Werner Heisenberg

Contents

To be submitted as: Martins, JZR; Hinkley, T; Kouyos, R; Haddad, M; Whitcomb, JM; Petropoulos, CJ; Bonhoeffer, SL. ***Evolution of recombination on an HIV derived fitness landscape.***

Abstract

The Hill-Robertson effect, which states that the interaction between genetic drift and selection generates unfavorable linkage disequilibrium (hence favoring recombination), offers one of the most promising hypotheses to explain the evolution and the maintenance of sexual reproduction and recombination. On the basis of simple models of fitness landscapes, it has been shown that the Hill-Robertson works independently of epistatic interactions. In this chapter, we test whether this is also valid in the case of individual mutation-based models of fitness landscapes which are based on estimates of the fitness effects of 1'857 single mutations and of 257'536 pairs of mutations found in a 70'081 HIV-1 B pol-genotypes assayed for in vitro replication capacity. We use computer simulations to mimic the evolution of HIV populations and we address the question of whether genetic drift also outweighs epistasis as a factor for the evolutionary maintenance of recombination in the case of more complex and rugged fitness landscapes. Although recombination is shown to be generally favored in finite population for individual mutation-based models, evolved HIV populations cannot be kept in realistic regions of the sequence space, and therefore it remains unclear whether genetic drift outweighs epistasis as a factor for the maintenance of recombination in the case of a more complex and rugged approximation of the landscape.

3.1 Introduction

Understanding the evolution and the maintenance of sexual reproduction and recombination has been one of the most intriguing problems of evolutionary biology of the last thirty-five years (Kondrashov, 1993). The most prominent population genetic theories that have been proposed to answer this problem argue that the benefit of sexual reproduction and recombination arises from breaking apart harmful genetic linkage disequilibria (statistical associations between alleles at different loci), and assume that this benefit outweighs the cost of breaking apart co-adapted gene combinations (commonly referred to as recombination load) (Barton and Charlesworth, 1998). As for the cause of the emergence of linkage disequilibria, there are two dominant views. The stochastic view states that genetic linkage disequilibria is primarily due to the interaction between genetic drift and selection, whereas the deterministic view postulates that linkage results rather from epistasis, in other words, the way genes or alleles interact (Kouyos et al., 2006). Using computer simulations, Keightley and Otto have shown that varying levels of epistasis did not significantly affect the benefit of recombination in finite populations (Keightley and Otto, 2006). Besides, it was shown that the benefit of recombination increases with population size, given that there is selection on sufficient loci (Iles et al., 2003; Keightley and Otto, 2006). In brief, it has been argued that epistasis is negligible in comparison to the interplay between drift and selection as a mechanism generating linkage disequilibria, on which the benefit of recombination relies. These results are based on simple models of fitness landscapes. Later, de Visser et al. (de Visser et al., 2009) found a general disadvantage of sex and recombination on an empirical fitness landscape. As this landscape relied on five loci only, de Visser's results are not directly comparable to the previous ones assuming selection on many loci. Nevertheless, de Visser's results pointed out that the topography of the fitness landscape and, in particular, the presence of sign epistasis (when the sign of the fitness effects of an allele varies across genetic backgrounds (Weinreich et al., 2005)) have a significant effect on the benefit of recombination (de Visser et al., 2009).

A system analysis of the mutational effects in HIV-1 *pol* genes provided two individual mutation-based models of a large fitness landscape derived from experimental data (see appendix A). These models constitute a framework of unprecedented biological realism, on which it is possible to examine the evolution and the maintenance of recombination for the case of large fitness landscapes. We use these models to test the earlier results of Keightley and Otto (Keightley and Otto, 2006) on large-scale and biologically realistic fitness landscapes. Accordingly, we simulate competition assays between recombination and non recombinant HIV genomes, and track the fate of recombination modifier alleles for different parameters. Hence, we test whether recombination is still robustly selected for, not only in the case of simple fitness fits, but also in the case of the more complex individual mutation models, with and without epistasis, also in the form of sign epistasis. We find that recombination is usually favored, but in contrast to the results of Keightley and Otto (Keightley and Otto, 2006), we observe that the increase of population size does not always increase the strength of selection for recombination. Unfortunately, we cannot draw any conclusions on the determinants of the benefit of recombination due to important limitations of the application of the different models of fitness landscapes which we highlight in the discussion section.

3.2 Materials and methods

To mimic the evolution of HIV populations, we implemented a standard genetic algorithm for N sequences with a maximum of resemblance to the HIV-1 sequences of the sequence alignment. Each sequence consisted of 404 polymorphic amino acid residues[1]. In total, 1455 single mutations were allowed, which corresponded to the set of single mutations found in the HIV sequence alignment used to fit the fitness models[2]. The number of mutations per generation and per sequence followed a Poisson distribution with mean μ. The mutations were randomly and uniformly distributed from residues 1 to 404. The fitness was calculated for four different fitness models. Two models inferred fitness based on the number of substitutions to the consensus sequence, commonly referred to as Hamming distance (HD), and the other two models inferred fitness based on estimates of main and epistatic effects of individual mutations.

In case of the two HD-based models, the fitness w of a sequence i was calculated as $w_i = exp(b_0 + \alpha n_i + \beta n_i^2)$, where b_0 is the intercept of the model, n_i the HD to the consensus sequence, α the independent fitness effect of a mutation and β its epistatic contribution. In the simplest variant of the model, we assumed no epistasis ($\beta = 0$). In this simple case, we assumed that the log-fitness decreases linearly with the accumulation of substitutions to the consensus. In essence, the sequences were sorted according to their HD to the consensus and the average log-fitness of each class of HD fitted to a linear model $log(w_i) = b_0 + \alpha n_i$. In the second variant, to take

[1] 99 Protease residues plus 305 Reverse Transciptase residues.

[2] For a detailed description of the sequence data and of the fitness assay, please refer to chapter 1 and to the appendix A.

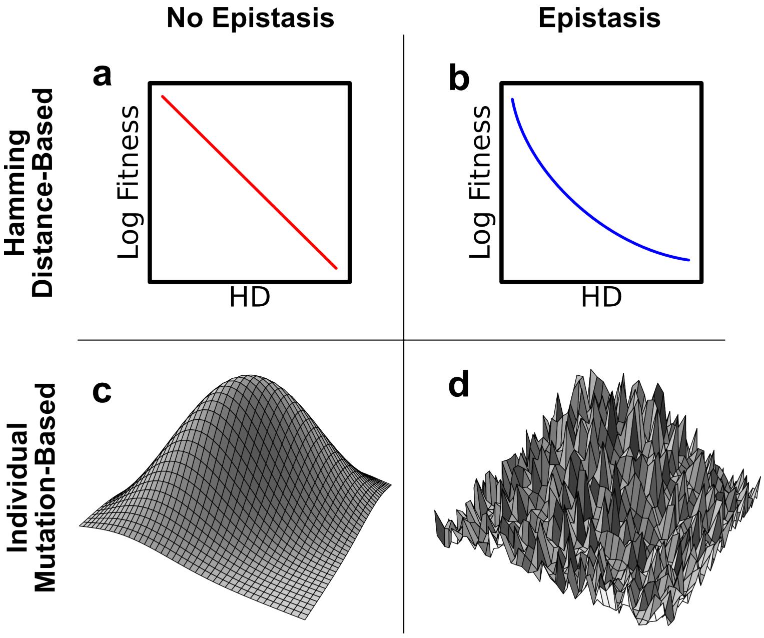


Figure 3.1: A diagrammatic representation of the four different fitness models used in this chapter. **a** illustrates the log-fitness linear regression model based on the number of substitutions to the consensus sequence and **b** the analog quadratic regression model, **c** illustrates the individual mutation main effects (ME) model and **d** the individual mutation main and epistatic effects (MEEP) model.

epistasis into account, the average log-fitness of each class of HD was fitted to the quadratic function with parameters b_0, α, and β. In the case of the two individual mutation-based models, we used the ME model (which excludes epistasis) and the MEEP model (which includes epistasis). The descriptions of the individual mutation-based models is given in chapter 1. For a better overview of the models, a schematic outline of the four different fitness models is provided in figure 3.1.

A binary modifier of the recombination rate was added at the end of each sequence. The start frequencies of the modifiers were set 50% at state `0` and 50% at state `1`. The modifier locus was left mutation free. Pairs of sequences were sampled with replacement from the parental population with a probability proportional to their fitness (by “roulette-wheel” selection) and the number of recombination events followed two Poisson distributions with means r_{01} (with $\mathrm{r}_{01} = \mathrm{r}_{10}$) and r_{11} depending on the combination of the pair of modifiers. For all cases, we assumed $\mathrm{r}_{00} = 0$. The positions of the recombination breaks were also randomly and uniformly distributed along the entire sequence. The sampling took over until the number of offspring sequences reached N so as to maintain a constant population size. Each run of the simulation started with an homogeneous population consisting of N repeats of a sequence corresponding to a local optimum, which was reached by adaptive walks climbing the fitness landscape by means of steepest ascent from the consensus of the sequence data. We also implemented an initial equilibration (*burn-in*) period of a minimum of N generations to ensure that the population was allowed to reach a state of selection-mutation balance around one or more optima of the fitness landscape. After the equilibration period, we let the modifiers to freely change in frequency until their complete extinction or fixation in the population.

3.3 Results

To determine the odds for the fixation of a recombination modifier for different parameter combinations, we implemented computer simulations which reproduced the evolution of HIV populations in the four different fitness models shown in figure 3.1. The models can be grouped in two different ways. They can be either HD-based or individual mutation-based, or they can either exclude or include epistasis. Although both HD-based models have nearly the same fitness predictive power (see figure 3.2 B), the quadratic model fits the data significantly better than the linear model (see figure 3.2 A), which is evidence for a considerable amount of epistasis in the empirical fitness landscape.

The individual mutation-based models offer a substantial increase in terms of fitness predictive power and also highlight the significance of epistasis in fitness determination, with the MEEP model providing a significant higher predictive power than the ME model. For a detailed analysis of the predictive power of the ME and the MEEP models please see the appendix A. To limit the number of parameter combinations of the genetic algorithm, we explored the parameter space in a twofold manner. We took into consideration both small and medium/large population sizes ($N = 1'000$ and $N = 10'000$) (Brown, 1997; Kouyos et al., 2006). The recombination rate was left constant, at a level of the same order of magnitude of previous estimates (Rajaram, Minin, Suchard, and Dorman, Rajaram et al.; Zhuang et al., 2002) ($\mathrm{r}_{11} = 0.1$ crossovers per sequence per generation). We also allowed an intermediate recombination rate to consider not only a scenario where the modifier is completely linked to its genetic background, similar to what is to be expected from competing sexual and asexual genotypes of a same species ($\mathrm{r}_{01} = 0$ and $\mathrm{r}_{10} = 0$ crossovers per sequence per generation), but also the case where the modifier is only partially linked to the selected

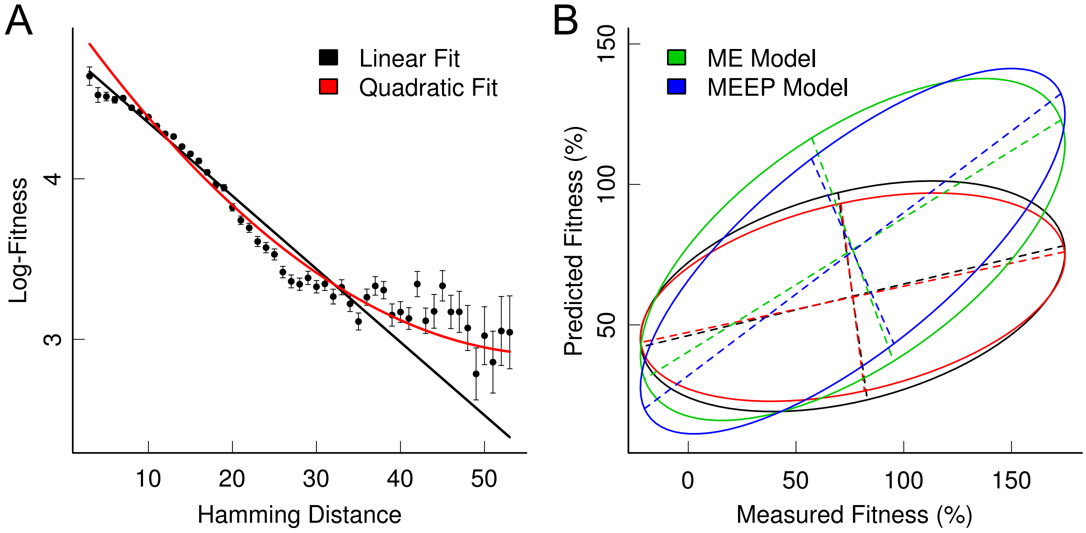


Figure 3.2: A. Mean and standard errors (black dots and gray vertical bars) of log fitness as a function of the number of mutations (Hamming distance) to the consensus sequence for all the sequences in the data set. The black slope shows the linear fitness fit, whereas the red curve shows the quadratic fitness fit. **B.** Illustration of the R-squares of the four fitness models by dispersion ellipses of predicted against measured fitness values of 5'000 sequences sampled at random from the data set. Black stands for the linear model, red for the quadratic model, green for the ME model, and blue for the MEEP model.

loci by allowing an intermediate recombination rate ($r_{01} = 0.05$ and $r_{10} = 0.05$ crossovers per sequence per generation). Finally, we tested both a low and a high mutation rate ($\mu = 0.1$ and $\mu = 0.5$ mutations per sequence per generation). The order of magnitude of biological estimates for the mutation rate is assumed to be in the range of the chosen mutation rates (although $\mu = 0.5$ is presumably exaggerated) (Brown, 1997; Wain-Hobson, 1993).

We ran 500 instances of the genetic algorithm for each combination of parameters and for each fitness model. Each instance is initialized with an homogeneous population of N sequences which, for a given fitness model, corresponds to the nearest local fitness optimum to the consensus sequence. In essence, we subjected the consensus sequence to adaptive walks that climbed the fitness landscape by means of the steepest ascent. To bring the homogeneous population to mutation-selection balance, we allowed an equilibration period of a minimum of N generations of mutation, selection, drift and recombination. After the equilibration period, the modifier for recombination was introduced in half of the population. And finally, the population was allowed to evolve until complete fixation or extinction of the modifier for recombination. As the proportion of fixations and extinctions follows a binomial distribution of sample size 500, it is possible to statistically test the evolution of recombination for a given combination of parameters and a given fitness model as it is shown in figure 3.3. Figure 3.3 reveals that recombination is favored in all four fitness models. This result, however, is not significant for 5 out of the 32 possible parameter combinations, namely when the population size is high and the mutation rate is low. On the positive note, for a high mutation rate, recombination is always strongly favored.

There is no significant difference between the outcome of the two different intermediate recombination rate scenarios: no clear pattern can be assigned to the introduction of the intermediate recombination rate

The fixation frequencies of the recombination modifier show greater similarity between the linear and the quadratic models (pairwise comparison of the frequencies of the two upper panels of figure 3.3) than between the ME and the MEEP models (pairwise comparison of the frequencies of the two lower panels of figure 3.3).

Interestingly, it is in the MEEP model for fitness landscape that recombination is the least favored (see low right panel of figure 3.3) — a benefit for recombination is still present but to a weaker degree.

3.4 Discussion

The fixation frequencies of the recombination modifier show an overall advantage of recombination, also in the fitness landscapes that correspond to the more complex individual mutation-based models. In this respect, our results are in line with previous findings by Keightley and Otto (Keightley and Otto, 2006). Yet, we see no generalized increase of the advantage of recombination with the increase of the population size. Instead, an increase in the benefit of recombination when both the population size and the mutation rate are large suggests that the advantage of recombination depends on an interaction between the population size and the mutation rate. Furthermore, when the population size is large but the mutation rate is small, the benefit of recombination is weaker or even absent.

The fitness landscapes defined by the ME and the MEEP models are characterized by vast neutral regions of the sequence space (Kouyos et al., 2012); therefore, the interactions and the effects of the parameters might only be better revealed for very high levels of rates of recombination and of mutation. As we tested the parameters in a somewhat reasonably realistic range, we might fail to catch some of their effects and interactions. Further-

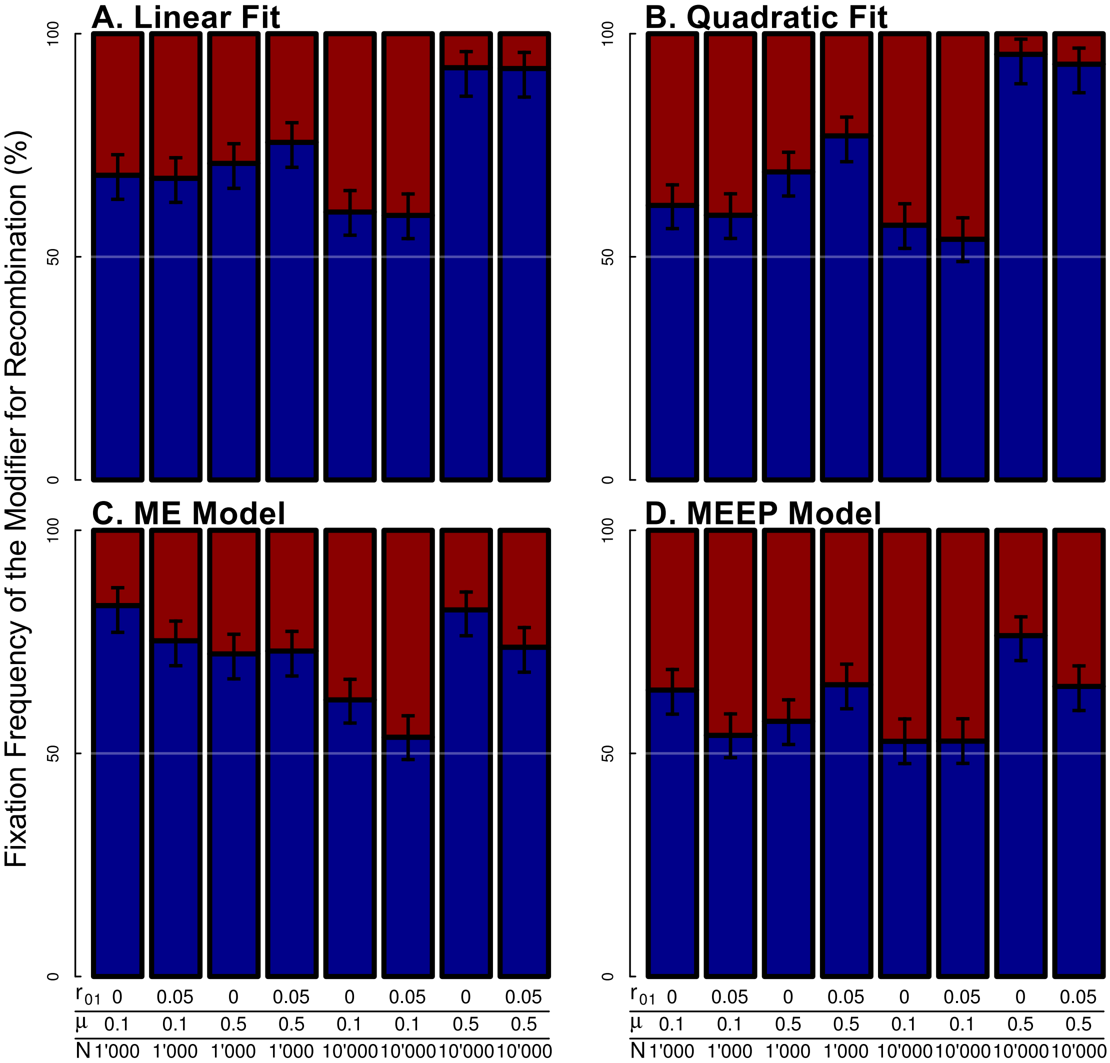


Figure 3.3: Barplots showing the relative frequencies of the recombination modifier for 500 independent runs. Blue represents the proportion of the runs for which the recombination modifier fixates in the population and red the proportion of the runs for which the recombination modifier goes extinct. Each panel corresponds to one of the four fitness models of this chapter (linear and quadratic fits, and ME and MEEP models). For a diagrammatic representation of the models see figure 3.1. The parameters are displayed below the x axis. r_{01} stands for the intermediate recombination rate, μ the mutation rate, and N the population size. Error bars represent the 95% confidence intervals of the relative frequencies of the modifier.

more, although the ME and the MEEP models provide more than a two-fold increase in terms of fitness predictive power, their highest fitness optima are located in unobserved and, in all likelihood, unrealistic regions of the sequence space — more than a hundred substitutions away from the consensus sequence, whereas the observed range of number of substitutions to the consensus does not exceed 60. This is due to the fact that the distribution of estimated fitness effects obtained in (Hinkley et al., 2011) contains a high number of beneficial mutations which are fixated by natural selection (see figure 3.4).

As a consequence, in these two models, after the initial equilibration phase, evolved populations consist of presumably unrealistic sequences. (Without the equilibration phase recombination is selected for also because it provides an extra source of variability that speeds up the adaptation of the recombination positive fraction of the population by means of the Fisher-Muller hypothesis (de Visser and Elena, 2007).) In the case of the HD-based models, there is only one fitness optimum, which corresponds to the consensus sequence. Thus, in these two models, and for reasonable rates of mutation and recombination, evolved populations remain in a realistic region of the sequence space.

The introduction of an intermediate recombination rate partially breaks the linkage between the recombination modifier and its genetic background. For this reason, we expected a consistent reduction of the benefit of recombination in this scenario. However, the introduction of an intermediate recombination rate showed no coherent effect on the benefit of recombination.

Interestingly, the fitness landscape defined by the MEEP model is generally less favorable for the evolution of recombination than those defined by the other models. As the MEEP model specifically allows for sign epistasis, we can speculate that the inclusion of this form of epistasis dampens the generalized benefit of recombination provided by the Hill-Robertson effect. Sign epistasis is assumed to be detrimental for the evolution of recombination because the more sign epistasis is present in the fitness landscape, the fewer the mutational pathways that can be traversed by natural selection. If this is the cause for the decrease of the general advantage of recombination in case of the MEEP model, then this result would indicate that the benefit of recombination depends on the topography of the fitness landscape, which would agree with (de Visser et al., 2009). Yet, only the implementation of a deterministic version of the simulations on the individual mutation-based models would allow us to quantify with exactitude which part of the benefit of recombination is attributable to the stochastic versus the deterministic effects. Unfortunately, this is not feasible because it requires the monitoring of the frequencies of approximately 2^{1000} genotypes.

Alternatively, we can imagine that the epistatic interactions of sequences at the fitness optima of the landscape defined by the MEEP model are par-

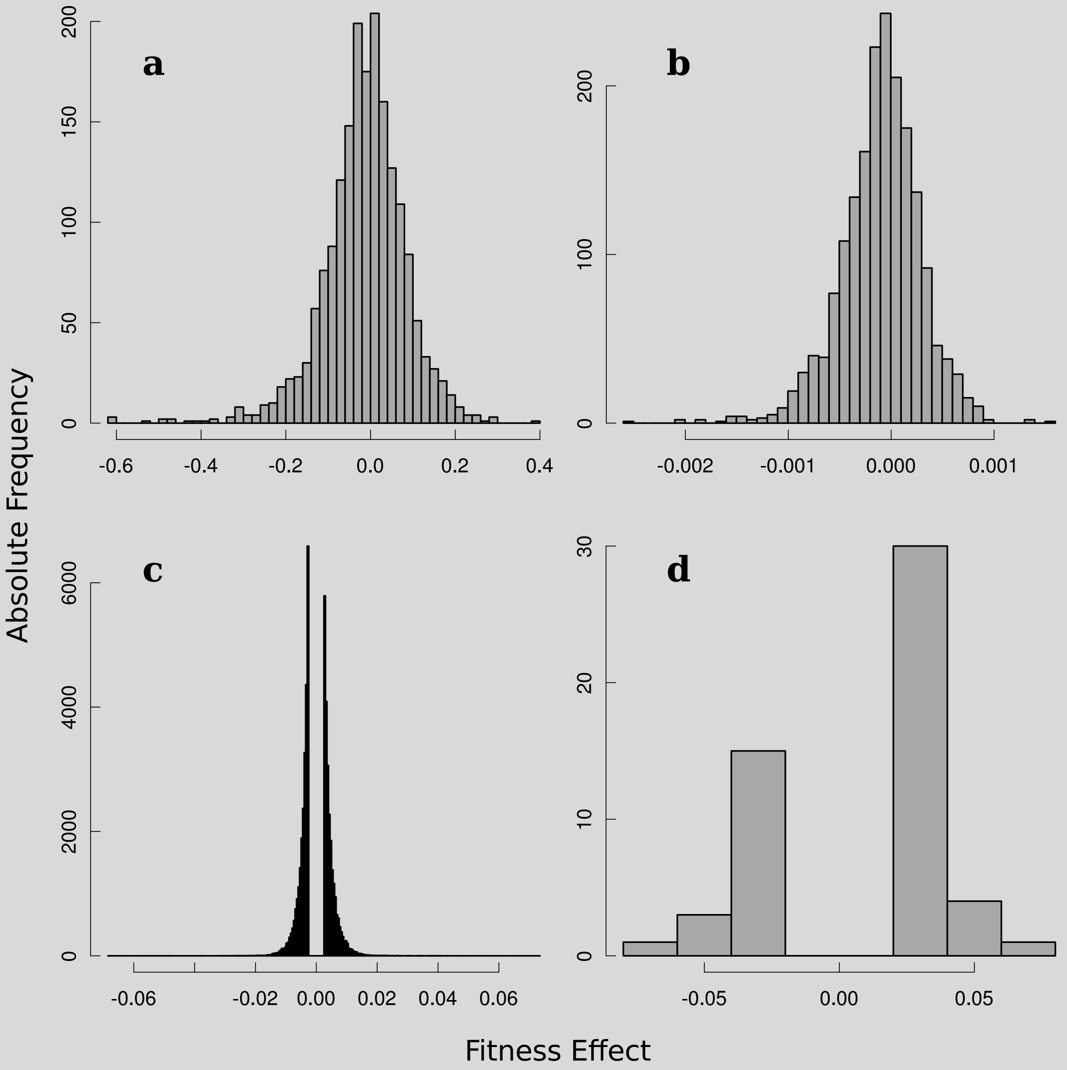
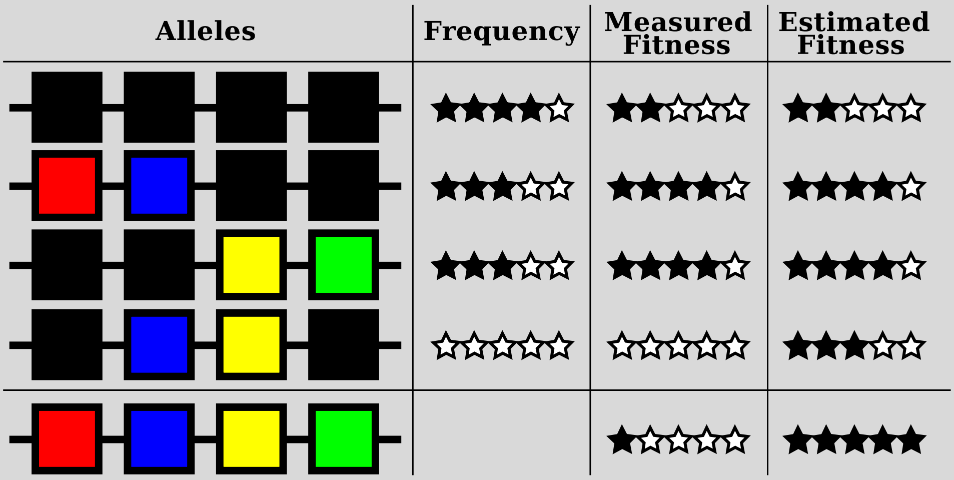


Figure 3.4: Distribution of the fitness effects of the ME and the MEEP models. **a.** Distribution of the fitness effects of the ME model. **b.** Distribution of the main effects of mutations of the MEEP model. **c.** Distribution of the 0.5% strongest epistatic effects of the MEEP model. **d.** Distribution of the $5 \cdot 10^{-5}\%$ strongest epistatic effects of the MEEP model.

tially meaningless. In the sequence region corresponding to these fitness optima, sequences are likely composed by series of alleles whose combinations are presumably poorly represented (if not absent) in the sequence data. One reason for the absence of certain allele combinations can be their possible highly detrimental fitness effects. Thus, even though it was shown that pairwise epistatic interactions of the MEEP model play an important role in the determination of fitness of unseen biological sequences (see the appendix A), they are insufficient to justly characterize the fitness landscape of unrealistic regions of the sequence space. Ideally, it would be necessary to distinguish between genuinely neutral allele interactions and the false negatives — the highly detrimental allele interactions which lead to unrealistic sequences (and therefore are assumed to be neutral for their absence in the sequence data). An illustration of this limitation is shown in figure 3.5.



Alleles	Frequency	Measured Fitness	Estimated Fitness
⬛⬛⬛⬛	★★★★☆	★★☆☆☆	★★☆☆☆
🟥🟦⬛⬛	★★★☆☆	★★★★☆	★★★★☆
⬛⬛🟨🟩	★★★☆☆	★★★★☆	★★★★☆
⬛🟦🟨⬛	☆☆☆☆☆	☆☆☆☆☆	★★★☆☆
🟥🟦🟨🟩		★☆☆☆☆	★★★★★

Figure 3.5: Illustration of one limitation of the application of individual mutation models. In the first row, we assume that the most common combination of four alleles (reference sequence) has average fitness, which is accurately predicted by the model. In the second and third rows, we assume two less common pairwise interactions presenting a fitness advantage, but which are still well predicted by the model. In the fourth row, however, we assume a highly detrimental pairwise interaction which results from two single mutant alleles present in the second and in the third rows. As this combination is never or seldom observed, the model assumes that its effect is neutral and incorrectly predicts a fitness value which migh be lower than the two previous ones but higher than the reference. In the last row, we assume the case of a sequence consisting of all four mutants present in the previous three rows. In this case, the model will incorrectly estimate a very high fitness and this sequence, although unrealistic, will likely be selected.

All in all, and despite a substantial increase in fitness predictive power, the individual mutation-based models have limited fitness predictive power for sequences composed of poorly represented alleles or combinations of alleles. Nevertheless, our results support the idea that there is no general disadvantage of evolving recombination; on the contrary, regardless of the fitness model, recombination is favored especially when mutation rates are high, which what is typically found in studies using recombination modifiers (Hartl et al., 1997).

CHAPTER

FOUR

DRUG PROFILES AND LEVELS OF RESISTANCE AND CROSS-RESISTANCE

Gring ache u seckle.

Anita Weyermann

Contents

Published as: Martins, JZR; Chappey, C; Haddad, M; Whitcomb, JM; Stawiski, E; Petropoulos, CJ; Bonhoeffer, SL. ***Principal component analysis of general patterns of HIV-1 replicative fitness in different drug environments.*** Epidemics, 2010. Martins et al. (2010)

Abstract

To detect general patterns and temporal trends of HIV-1 resistance we apply principal component analysis (PCA) to *in vitro* fitness data. 28’000 virus samples, obtained between 2002 and 2008, were assayed for fitness in 16 to 21 selective environments. Fitness measurements are based on replication capacity (RC), which quantifies *in vitro* viral replication in a single cycle of infection. RC is determined both in the absence of drugs and in the presence of 6-7 nucleoside analog reverse transcriptase inhibitors (NRTIs), 3-4 non-nucleoside reverse transcriptase inhibitors (NNRTIs), and 6-9 protease inhibitors (PIs). PCA shows remarkable structure in RC across the different environments, which reveals differences in the patterns of resistance and cross-resistance between drugs or between drug classes. To probe the causes of the observed patterns, we develop a model to generate simulated data and subject these simulated data to an equivalent analysis. By comparing the outcomes of PCA on the original and the simulated data, we quantify which part of the total variance of the original data is due to non-specific effects, class-specific effects, and drug-specific effects of resistance mutations. We find that relative fitness is mainly drug-independent and that drug-specific effects are substantially bigger than class-specific effects for NRTIs, but not for NNRTIs or PIs. The observed patterns remain remarkably stable over the period of observation. Comparison with known potent combination therapies suggests that PCA helps to identify combinations that act synergistically in preventing the emergence of resistance.

4.1 Introduction

Antiretroviral therapy has fundamentally changed the face of the HIV epidemic in the developed world. Over the last two decades more than 20 new antiretroviral drugs have been developed and the treatment with combinations of these drugs significantly reduces mortality and morbidity (Egger et al., 1997; Hammer et al., 1996; Mocroft et al., 1998; Palella et al., 1998). The clinical benefit of treatment, however, is compromised by the remarkable capacity of the virus to evolve resistance. A sustainable use of the existing antiretrovirals thus necessitates a sound understanding of the general patterns of resistance that emerge in the HIV epidemic. To establish these patterns from epidemiological data, however, presents considerable challenges. The large number of drugs, the even larger number of drug resistance mutations, and the varying levels with which individual mutations confer resistance to multiple drugs (i.e. cross-resistance) result in complex data structure, and detecting general patterns requires methods that reduce the high dimensionality of the data (Bennett et al., 2009; Egger et al., 1997; Harrigan and Larder, 2002).

The aim of the present chapter is to provide a comprehensive picture of HIV-1 drug resistance and its temporal evolution for three major drug classes, and to develop a quantitative understanding for the underlying mutational effects. We focus on the effect of protease and reverse transcriptase inhibitors on replication capacity (RC), an *in vitro* measure of viral fitness based on a single round of replication (Petropoulos et al., 2000). Fitness is measured in 16-21 selective environments, consisting of a drug-free environment, 6-7 environments containing nucleoside reverse transcriptase inhibitors (NRTIs), 3-4 environments containing non-nucleoside reverse transcriptase inhibitors (NNRTIs), and 6-9 protease inhibitors (PIs). Using 4'000 virus samples per year between 2002 and 2008 we employ principal component analysis (PCA) to detect general patterns and time trends in resistance and fitness. To assist the interpretation of these patterns, we develop a model to simulate data and compare the outcome of PCA on the original and the simulated data to develop a time-resolved view of the role of class-specific versus drug-specific effects.

4.2 Materials and methods

4.2.1 Experimental data

28'000 virus samples derived from HIV-1 infected patients were assayed for RC. These virus samples were submitted to Monogram Biosciences, Inc, for drug resistance testing between 2002 and 2008. The assay is described in chapter 1 and was originally presented elsewhere (Petropoulos et al., 2000). In brief, patient virus-derived amplicons representing all of the protease (PR) and residues 1-305 of the reverse transcriptase (RT) are inserted into the backbone of a resistance test vector. The amplicons reflect the diversity of PR and RT in the patient. The resistance test vector is a modified NL4-3 HIV clone such that it can only undergo a single round of infection. The RC is then determined as the total production of infectious progeny virus after a single complete replication cycle of the patient-derived virus relative to that of an NL4-3 based control virus. (Note that there are corresponding multi-cycle assays (Dykes et al., 2006; Miao et al., 2008), with different advantages and disadvantages.) The RC of the NL4-3 based control virus is thus equal to 1. For the virus samples analyzed here, the RC is measured not only in the absence of drugs, but also in the presence of 15-20 different single drugs (6-9 PIs, 6-7 NRTIs, and 3-4 NNRTIs) at a series of drug dilutions. The RC on drugs was obtained by interpolating measurements at different drug concentrations. For each drug, the RC of a virus was given by the interpolated value that corresponds to the drug concentration at which the NL4-3 based control virus has 10% of its RC in the absence of drugs. Specifically, the initial set of drugs is the following: 6 PIs—amprenavir (AMP), indinavir (IDV), lopinavir (LPV), nelfinavir (NFV), ritonavir (RTV), and saquinavir (SQV); 6 NRTIs—abacavir (ABC), didanosine (ddI), lamivudine (3TC), stavudine (d4T), zidovudine (ZDV), and tenofovir (TFV); and 3 NNRTIs—delavirdine (DLV), efavirenz (EFV), and nevirapine (NVP). 1 PI is introduced in 2003—atazanavir (ATV). In 2005, 1 PI—tipranavir (TPV)—and 1 NRTI—emtricitabine (FTC)—are introduced. 1 PI is introduced in 2006—darunavir (DRV)—and, at last, 1 NNRTI is introduced in 2008—etravirine (ETR).

4.2.2 Principal Component Analysis

PCA is a statistical method used to identify structure in a cloud of points in a multidimensional coordinate system. PCA identifies a new ordered set of orthogonal coordinates such that they capture decreasing amounts of maximum variance. An intuitive understanding of the PCA can be obtained as follows. Imagine a projection of the multidimensional cloud of points onto a line. Then this cloud is rotated until that an orientation is found for which the variance of the projected points along the line is maximal. This orientation defines the first principal component (PC). The remaining PCs are orthogonal to the first PC, but are ordered such that they explain decreasing maximum amounts of variance in the cloud of points. The more the variance is explained by the first PCs, the more the cloud of points is structured (Gnanadesikan, 1977; Jolliffe, 1986).

The mathematical procedure yields eigenvalues that quantify the fractions of the variance of the original cloud of points along the PCs. The result of a PCA is usefully summarized in scatter-plots that show a projection of the cloud of points onto planes defined by pairs of principal components. The normed vectors spanning up the original coordinate system are shown as arrows projected onto the plane of the scatter-plot. These scatter-plots are commonly called biplots (Gabriel, 1971; Gower and Hand, 1996).

Here, the cloud of points corresponds to the RC data consisting of a table with 28'000 rows (i.e. the virus samples) and 16-21 columns (i.e. the selective environments: absence of drugs; 6-7 NRTIs; 3-4 NNRTIs; 6-9 PIs). The RC values were log-transformed as this eliminated positive skewness of the values' distribution and improved the homogeneity of their variance. Our analysis was performed with the `ade4` `R` package (Chessel et al., 2004; R Development Core Team, 2009).

4.2.3 Simulated data

The PCA is a powerful tool to reveal hierarchical structure in variance and covariance in the data. However, it does not allow to infer directly the underlying processes that are responsible for the observed patterns. To obtain a better understanding of the patterns observed in PCA of the HIV samples, we develop a model based on assumed effects of mutations on RC in different selective environments. To this end we produce data with similar structure based on biologically motivated assumptions in order to test whether this model is compatible with the data.

We generate simulated data by the following approach. First, we generate sequences that are maximally similar to the original data with regard to sequence length and drug resistance polymorphism. Specifically, as the experimental data is based on the 99 amino acids of the PR and the first 305 amino acids of the RT, we generate corresponding random sequences with a total length of 404 positions. Each position in the sequence can be in one of two states corresponding to either the drug sensitive wild-type or a drug resistance mutation. We define a vector $\mathbf{m} = (m_1, m_2, \ldots, m_{404})$, where m_k are binary random variables. The number of mutations in each sequence are sampled from numbers of resistance mutations found in a similar dataset based on 9'466 sequences (Bonhoeffer et al., 2004) . In addition, we subdivide $\mathbf{m}$ into $\mathbf{m}^{\mathrm{PR}} = (m_1, m_2, \ldots, m_{99})$ and $\mathbf{m}^{\mathrm{RT}} = (m_{100}, m_{101}, \ldots, m_{404})$. Second, we generate the corresponding RC values for the sequences, which we store in a $N \times E$ matrix $\mathbf{\Phi} = [\phi_{i,j}]$, where N is the number of viral samples and E the number of drug environments. We assume that the effect of each mutation has up to three components depending on the environment. The first component (non-specific component) reflects that part of the effect which is present in all environments, which we write $\mathbf{a}^{\mathrm{PR}} = (a_1, a_2, \ldots, a_{99})$ and $\mathbf{a}^{\mathrm{RT}} = (a_{100}, a_{101}, \ldots, a_{404})$, where the a_k are sampled from a normal distribution with mean zero and variance 1 (i.e. $N(0,1)$) . The second component (class-specific component) reflects that part which is specific to all drugs of a given class. They are also normally distributed with mean 0 but different variances. Specifically we have

$$\begin{aligned}
\mathbf{b}^{\mathrm{PI}} &= (b_1^{\mathrm{PI}}, b_2^{\mathrm{PI}}, \ldots, b_{99}^{\mathrm{PI}}), \quad b^{\mathrm{PI}} \sim N(0, \sigma_\beta^{\mathrm{PI}})\\
\mathbf{b}^{\mathrm{NRTI}} &= (b_{100}^{\mathrm{NRTI}}, \ldots, b_{404}^{\mathrm{NRTI}}), \quad b^{\mathrm{NRTI}} \sim N(0, \sigma_\beta^{\mathrm{NRTI}})\\
\mathbf{b}^{\mathrm{NNRTI}} &= (b_{100}^{\mathrm{NNRTI}}, \ldots, b_{404}^{\mathrm{NNRTI}}), \quad b^{\mathrm{NNRTI}} \sim N(0, \sigma_\beta^{\mathrm{NNRTI}})
\end{aligned}$$

The third component (drug-specific component) reflects that part which is specific to only one drug. It is given by

$$\begin{aligned}
\mathbf{c}^{\mathrm{PI}} &= (c_1^{\mathrm{PI}}, c_2^{\mathrm{PI}}, \ldots, c_{99}^{\mathrm{PI}}), \quad c^{\mathrm{PI}} \sim N(0, \sigma_\gamma^{\mathrm{PI}})\\
\mathbf{c}^{\mathrm{NRTI}} &= (c_{100}^{\mathrm{NRTI}}, \ldots, c_{404}^{\mathrm{NRTI}}), \quad c^{\mathrm{NRTI}} \sim N(0, \sigma_\gamma^{\mathrm{NRTI}})\\
\mathbf{c}^{\mathrm{NNRTI}} &= (c_{100}^{\mathrm{NNRTI}}, \ldots, c_{404}^{\mathrm{NNRTI}}), \quad c^{\mathrm{NNRTI}} \sim N(0, \sigma_\gamma^{\mathrm{NNRTI}})
\end{aligned}$$

As the effect of the third component must be specific to a single drug rather than to all drugs of the same class, we create three specificity matrices ($\mathbf{S}^{PI}, \mathbf{S}^{NRTI}$, and $\mathbf{S}^{NNRTI}$) with E columns. The number of rows is either 99 (PI) or 305 (NRTI and NNRTI). Each row of these matrices contains a 1 at a randomly chosen column of the one drug for which the mutation is specific and all other entries are set to 0. We define $\mathbf{d}_j^l$ as the j^{th} column-vector of $(\mathbf{c}_l)^{\mathrm{T}} \cdot \mathbf{S}^l$, with l = PI, NRTI, or NNRTI. Finally, for a given set of $\sigma_\beta^{\mathrm{NRTI}}$, $\sigma_\beta^{\mathrm{NNRTI}}$, $\sigma_\beta^{\mathrm{PI}}$ and $\sigma_\gamma^{\mathrm{NRTI}}$, $\sigma_\gamma^{\mathrm{NNRTI}}$, $\sigma_\gamma^{\mathrm{PI}}$, we generate sequences $\mathbf{m}_i$ with $i = 1 \ldots N$. For each sequence, we proceed to calculate the RC for all different environment as follows:

$$log(\phi_{i,j}) = \begin{cases} \mathbf{a}^{\mathrm{PR}} \cdot \mathbf{m}_i^{\mathrm{PR}} + \mathbf{a}^{\mathrm{RT}} \cdot \mathbf{m}_i^{\mathrm{RT}} & j \in \mathrm{X0} \\ (\mathbf{a}^{\mathrm{PR}} + \mathbf{b}^{\mathrm{PI}} + \mathbf{d}_j^{\mathrm{PI}}) \cdot \mathbf{m}_{\mathrm{PR},i} + \mathbf{a}^{\mathrm{RT}} \cdot \mathbf{m}_i^{\mathrm{RT}} & j \in \mathrm{PI} \\ \mathbf{a}^{\mathrm{PR}} \cdot \mathbf{m}_i^{\mathrm{PR}} + (\mathbf{a}_{\mathrm{RT}} + \mathbf{b}^{\mathrm{NRTI}} + \mathbf{d}_j^{\mathrm{NRTI}}) \cdot \mathbf{m}_i^{\mathrm{RT}} & j \in \mathrm{NRTI} \\ \mathbf{a}^{\mathrm{PR}} \cdot \mathbf{m}_i^{\mathrm{PR}} + (\mathbf{a}_{\mathrm{RT}} + \mathbf{b}^{\mathrm{NNRTI}} + \mathbf{d}_j^{\mathrm{NNRTI}}) \cdot \mathbf{m}_i^{\mathrm{RT}} & j \in \mathrm{NNRTI} \end{cases}$$

where j is an index that refers to the particular drug considered.

4.2.4 Similarity measures of experimental and simulated data

To quantify the similarity between the results of the PCA of the experimental data with those of the simulated data, we use two measures of similarity based on the shape and orientation of the two clouds of points. The first measure is calculated by adding up the absolute differences between the sorted eigenvalues of the PCAs based on the experimental and the simulated data. This measure therefore quantifies how similar the simulated and the experimental data are with regard to the variance along the PCs. More intuitively, it measures the similarity between the overall shape of the two clouds of points. The second measure is based on the orientation of the arrows that represent the normed vectors spanning up the original coordinate system based on RC value in each environment. Specifically, we first calculate both the centroids and the axial standard deviations of groups of arrows that belong to the same drug class, and then add up the absolute differences between the experimental and the simulated data with regard to the centroids and axial standard deviations for each drug class. (When comparing groups of arrows, we took into account that the axis' orientation can be flipped because its choice is arbitrary.) Intuitively, the second measure thus quantifies the similarity of the orientation of the two clouds of points in the original coordinate system. In the main text we refer to the first measure as the shape similarity and to the second as the orientation similarity.

4.3 Results

4.3.1 PCA of experimental data

To identify patterns of cross-resistance we apply PCA to replicative fitness data in 21 selective environments. Fig. 4.1 shows the PCA of 4'000

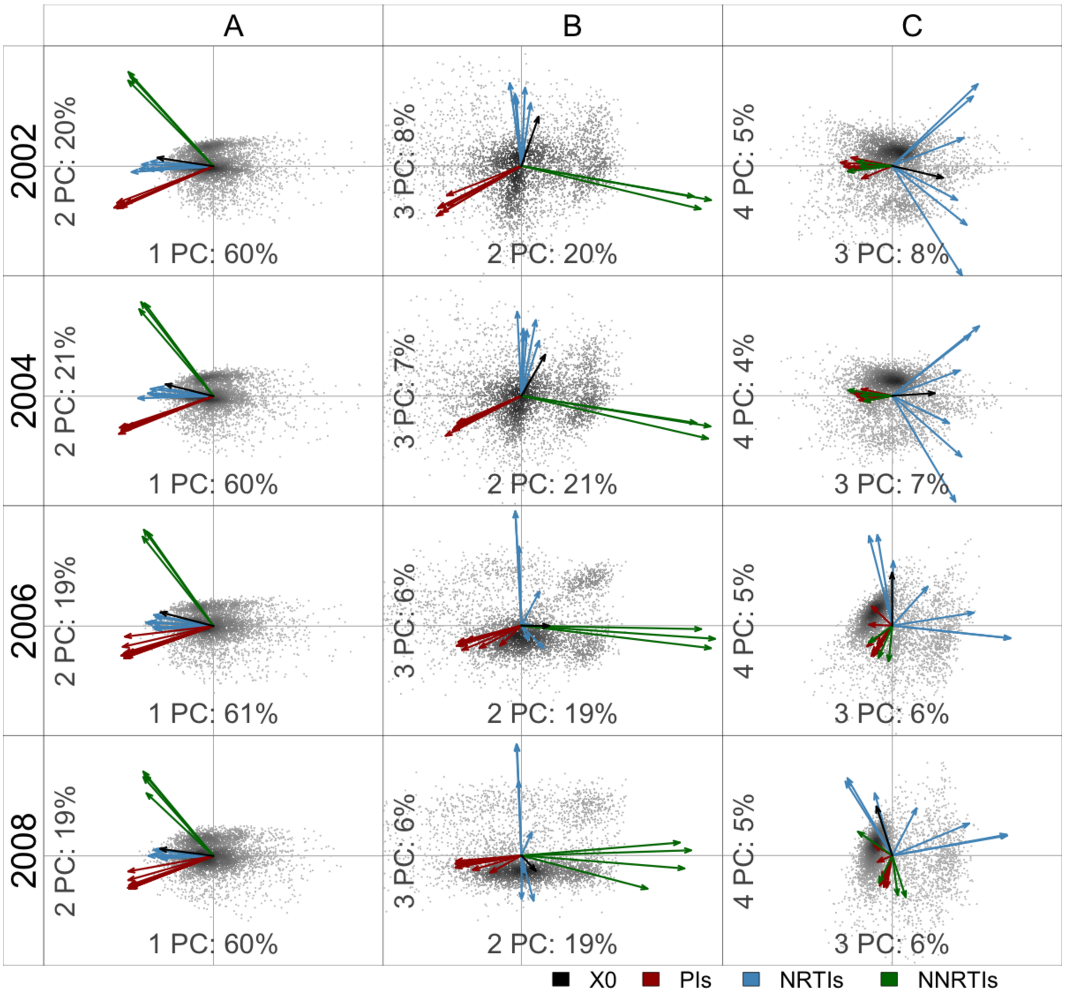


Figure 4.1: Biplots illustrating the PCA of the experimental data: for each of the 4 rows, panels *A* to *C* show the RC values of 4'000 viruses in 16-21 environments with respect to their projection onto the first 4 PCs. Column A, B, and C shows first against second, second against third, and third against fourth PC, respectively. The rows correspond to the years 2002, 2004, 2006, and 2008 and are based 4'000 viruses per year. The arrows show the projections of the initial coordinate system consisting of the RC value for 6-7 NRTIs (blue), 3-4 NNRTIs (green), 6-9 PIs (red), and 1 drug free environment (black). The variance captured by each PC is given as a label of the corresponding axis. Gray levels account for the density of the point scattering (and are generated with `geneplotter` (Gentleman and Biocore, 2006)).

HIV samples each from years 2002, 2004, 2006, and 2008. The biplots document that 4 PCs are sufficient to capture $> 90\%$ of the total variance in all years. Higher PCs (accounting for the remaining $< 10\%$) are regarded as non-informative and will be ignored in the following. A closer examination of fig. 4.1–*A* shows that the first PC captures approximately 60% of the total variance. Given that all arrows have a component that points to the left of the plot in column *A* of fig. 4.1, this implies that the first PC contrasts overall low with overall high RC viruses. This is supported by the fact that the average log-RC values across all environments and the projection of the RC values onto the first PC are highly correlated (see fig. 4.2). In particular, it is remarkable that the arrow reflecting the drug free environment points to the left together with all other arrows in fig. 4.1–*A*. This implies that viruses that have high RC in absence of drugs also tend to have high RC in the presence of drugs. This can either be the case because many of the mutations characterizing these viruses have similar effect in presence and absence of drugs. An alternative but not mutually exclusive explanation is that the effect of the mutations that are specific either to a single drug or to a class of drugs is manifest only in the corresponding environments. These environments, however, always only constitute the minority of environments as in our data set there are always more environments to which such mutations are not specific. As a consequence, the predominating effect in the analysis will always be that of a mutation in a non-specific selective environment.

Inspection of fig. 4.1–*A* also shows that arrows belonging to the same drug class tend to fall on top of each other. This indicates that a large part of the variance in the data is due to class-specific effects. The second PC shows that around 20% of the variance contrasts viruses on their susceptibility/resistance to PIs and NNRTIs (fig. 4.1–*A*). The general pattern seen in fig. 4.1–*A* remains very stable over the time period 2002-2008. Changes are predominantly due to introduction of new drugs over the time period of observation (see supplementary figure 4.7).

The third PC accounts for 6-8% of the variance (fig. 4.1–*B*). In 2002 and 2004, the third PC contrasts NNRTI and PI against NRTI resistant viruses. In 2006 and 2008, the third component begins to explain also differences within the NRTI class. Thus, adding up the variance accounted for by the first three components, we find that 80-85% of the variance in the data can be explained without evoking any drug-specific effect. While in 2002 and 2004 all arrows of the NRTI point in one direction, from 2006 onwards the NRTIS point in opposing directions along the third PC. This pattern is mostly, but not exclusively, due to the introduction of new drugs, as it partially remains when restricting the analyses only to those drugs that are available from 2002 (see supplementary figure 4.8).

In 2002 and 2004 the fourth PC, accounting for 4-5% of the total variance (fig. 4.1–*C*), begins to reveal drug-specific effects among the NRTIs, while the

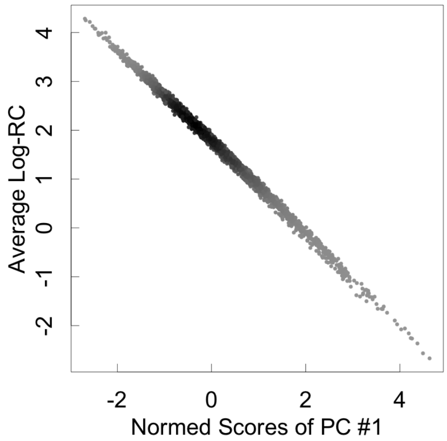


Figure 4.2: Average log-RC values over the 21 drug environments of 4000 samples from year 2008 against the projection onto the first PC ($R^2 = 0.996$).

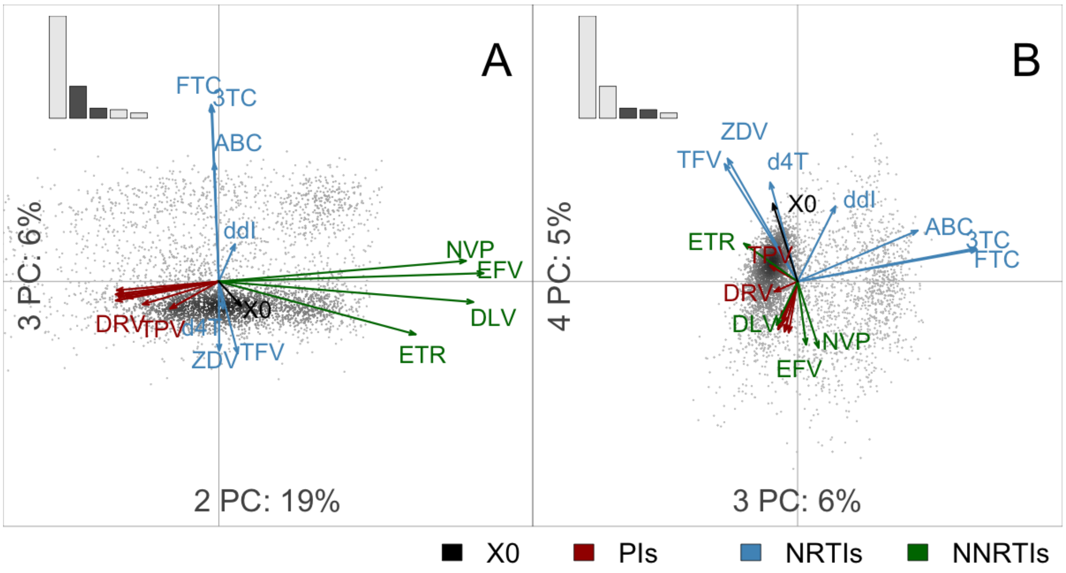


Figure 4.3: Biplots of the PCA of 2008 data with drug information. The plot corresponds to fig. 4.1–*B* and *C*, year 2008. Panels *A* and *B* show the RC values of 4'000 viruses in 21 environments with respect to their projection onto the second, third, and fourth PCs. The variance captured by each PC is shown in the inlays in the upper-left corner of each panel. Acronyms for the drugs or drug classes are found in *Methods*.

arrows for the NNRTIs and PIs remain much closer together. This implies that drug-specific effects are much more pronounced in NRTIs than in PIs or NNRTIs, which is in line with (Harrigan and Larder, 2002). In 2006 and 2008, the large contribution that NRTI specific effects make to the total variance begins to manifest itself already in the third PC (fig. 4.3–*A*), which also contrasts individual NRTIs against each other. In these same years, the fourth PC begins to reveal drug-specific effects, first in the PIs, and later in the NNRTIs (fig. 4.3–*B*). This is partially due to the introduction of new drugs and in part due to the fact that the 3rd PC begins to account for the drug-specific effects among the NRTI class (see supplementary figure 4.8).

4.3.2 Comparison of simulated and experimental data

The analysis of the experimental data reveals considerable hierarchical structure suggesting varying degrees of cross-resistance in the different drug classes. To assist the interpretation of the outcome of PCA of the experimental data, we simulate data according to a model in which the effect of each mutation on the RC is subdivided into an environment independent component (α), a class-specific component (β), and a drug-specific component (γ). For simplicity, these components are drawn from normal distributions. As we are interested in the relative importance of these components, we fix the standard deviation of the distribution from which α is drawn to 1 and we set the mean of all distributions to 0. For further details regarding the model see *Methods*.

By changing the variances of the normal distributions, we use this model to generate 100'000 datasets which differ in the relative contribution that these components make to the RC value in different environments. We explored how well the 100'000 simulated datasets exhibit generic features of the experimental dataset by using PCA based measures that quantify the similarity of the overall shape and of the orientation of the cloud of points formed by the experimental and the simulated data (see fig. 4.4–*A*). Thus our method is related to Approximate Bayesian Computing (Beaumont et al., 2002), with shape and orientation similarity being the summary statistics assessing the similarity between the actual data and the model.

For the top 100 simulated datasets that were maximally similar with regard to both similarity measures, we plotted the standard deviation of the class-specific effects, σ_β, against the standard deviation of the drug-specific effects, σ_γ (see fig. 4.4–*B*). To illustrate how similar the biplots of the simulated and the experimental datasets can be, we generated one dataset based on the average values of the standard deviations $\sigma_{\beta,NRTI}$, $\sigma_{\beta,NNRTI}$, $\sigma_{\beta,PI}$, $\sigma_{\gamma,NRTI}$, $\sigma_{\gamma,NNRTI}$, and $\sigma_{\gamma,PI}$ of the 100 most similar datasets (fig. 4.4–*B*). The biplot of the PCA of this simulated dataset is shown in fig. 4.4–*C*. The variances explained by the principal components are close for both datasets

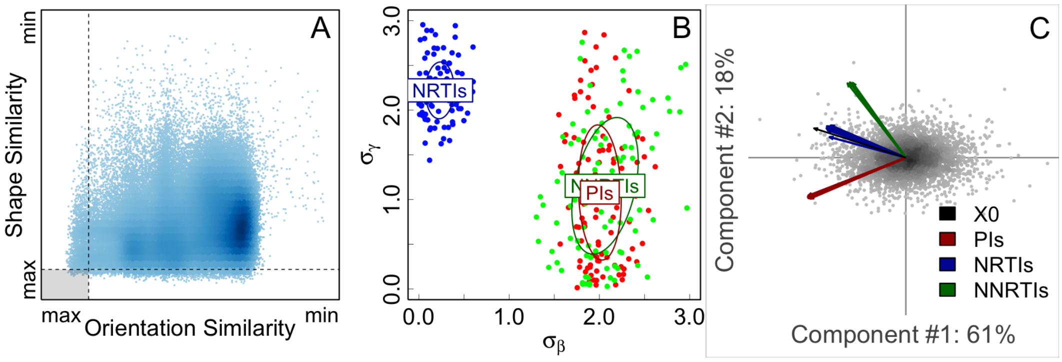


Figure 4.4: PCA of simulated data. Panel A shows the similarity between 100'000 simulated datasets and the experimental data based on shape and orientation of the cloud of points (see main text). Panel B shows the standard deviation of normal distributions underlying the class-specific effects, σ_β, against the standard deviation of the drug-specific effects, σ_γ for the 100 most similar data sets with regard to shape and orientation. Panel C shows the PCA of simulated data which was generated based on the average value of the standard deviations $\sigma_{\beta,NRTI}, \sigma_{\beta,NNRTI}, \sigma_{\beta,PI}, \sigma_{\gamma,NRTI}, \sigma_{\gamma,NNRTI}$ and $\sigma_{\gamma,PI}$ of the 100 datasets that were most similar to the PCA of the viruses from year 2002 (see fig. 4.1 *A* for comparison). The parameters σ_β and σ_γ were drawn randomly from a uniform distribution between 0 and 3.

(experimental dataset: 60%, 20%; simulated dataset: 61%, 18%. Moreover, the orientation of the arrows is similar in fig. 4.1–*A*, year 2002 (experimental data set) and fig. 4.4–*C* (simulated data set). Note, that although the PCA reveals considerable structural similarity between the simulated and the experimental data, comparison of the scatter-plots shows that the experimental data, in contrast to the simulated data, are not normally distributed.

Generating 100'000 datasets for each year and selecting the 100 best fits, we analyzed temporal trends in the contribution of the class-specific and drug-specific components (fig. 4.5). Across all years, we observe that the class-specific component outweighs the drug-specific component for NNRTIs and PIs, whereas the drug-specific component outweighs the class-specific component for the NRTIs. The class-specific component remains stable over time for the NRTIs and the NNRTIs, but decrease significantly over time in the PIs. The drug-specific component increases in all drug classes, but only significantly so in the NRTIs and the PIs. The temporal changes, however, are mostly attributable to the introduction of new drugs, with the exception of the decrease of the class-specific component of the PIs (see supplementary figure 4.9). The fact that the new drugs lead to an increase of the drug-

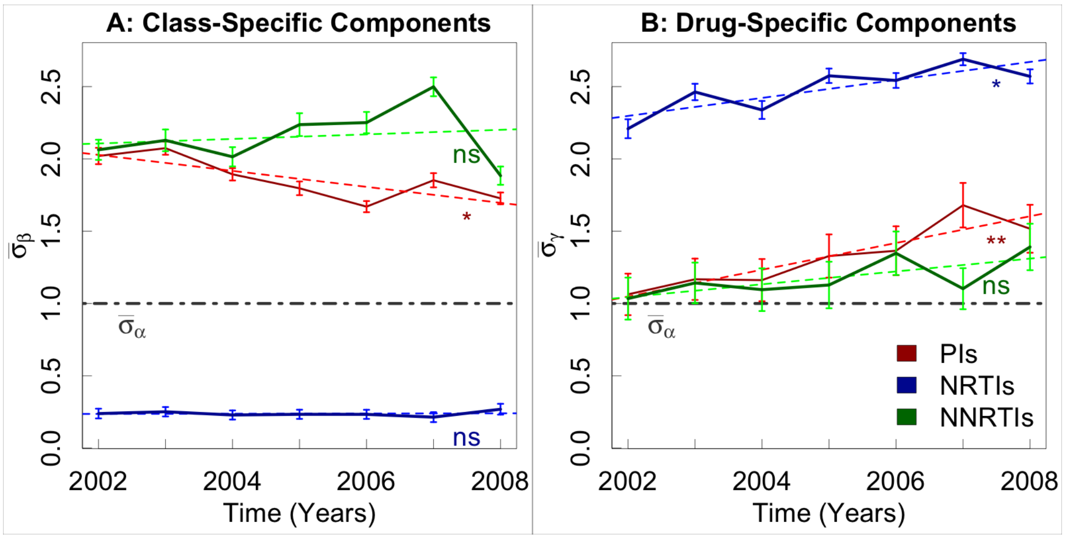


Figure 4.5: Time evolution of the class- and the drug-specific components of the mutational effects from 2002 until 2008. Panel A the time evolution of σ_β, the mean standard deviation (±SE) of the normal distributions from which the class-specific components are drawn. Panel B shows the time evolution of σ_γ, the mean standard deviation of the normal distribution from which the drug-specific components are drawn. Over the time window the class-specific components outweigh the drug-specific components for both the NNRTIs and the PIs, but not the NRTIs. The increase of the drug-specific components over time is mostly due to the introduction of new drugs (see supplementary figure 4.9).

specific component indicates that these drugs exhibit resistance patterns that are different from the previously existing drugs.

4.4 Discussion

PCA of the HIV samples ranging from 2002 to 2008 reveals considerable structure in the RC values across the 16-21 environments. The first PC essentially reflects the mean log-RC of the viruses across all environments and explains around 60% of the variance in the data. The comparison with the PCA of the simulated data supports the interpretation that the first PC is mainly due to the non-specific component of the mutational effects on RC. Although this non-specific component is responsible for most of the variance in the data, it is typically not the largest component of each mutation (compare σ_α with σ_β or σ_γ in fig. 4.5). The non-specific component nevertheless dominates the variance, because it is present for each mutation in all environments, and therefore contributes to RC everywhere. This also explains the somewhat counter intuitive observation that the arrow for the drug free environment points to the left together with the arrows of all drug environments (fig. 4.1, column *A*): the non-specific component is the only component that remains of a mutation in all environments to which this mutation is not specific.

The fact that the arrows for all members of a drug class essentially fall on top of each other when the data is projected onto the plane defined by the first and second PCs is remarkable. The likelihood that a random rotation of the data into a new coordinate system would achieve such an agreement is negligible, and hence points to the power of the PCA to reveal non-random patterns in the data. Moreover, it demonstrates the overriding effect of the drug class on the overall structure in the data.

Comparison of the experimental with the simulated data suggests that the variance captured in the second, third, and fourth PCs can mostly be explained by the class-specific component of the mutational effects. The drug-specific components are strongest for the NRTIs, and only for this class is the magnitude of the drug-specific component larger than that of the class-specific component. In line with earlier observations, our results thus suggest that cross-resistance is higher in the NNRTIs and PIs than in the NRTIs (Harrigan and Larder, 2002). Note, however, that the analysis presented here focuses on RC values in the absence and in the presence of drugs. Clinical studies such as (Harrigan and Larder, 2002) use changes in IC50, the drug concentration at which the virus is half maximally inhibited, as a measure of drug resistance. RC is a valid measure of *in vitro* fitness and it has been shown, in the absence of drugs, to correlate negatively with CD4 cell count and positively with viral load (Daar et al., 2005). Moreover, RC in the absence of drugs increases while IC50 decreases during long-term virologic failure (Barbour et al., 2002). However, the relation between RC on drugs and IC50 is not clear, as the present chapter is the first to investigate RC on drugs.

A temporal PCA analysis of HIV samples ranging from 2002 to 2008 shows that across all years the drug-specific component of the mutational effects outweigh the class-specific component for the NRTIs but not for the NNRTIs and the PIs. While this shows that the broad conclusions are robust over time, we emphasize that, like other datasets of comparable size, our dataset is not based on a randomly selected set of HIV patients. Hence, we cannot exclude that some part of the variance structure in the data is due to biased sampling or changing patterns in frequency of drug use. The temporal trends also show that the drug-specific component gains in importance over the years. Restricting the analysis to only those drugs that were present already in 2002, however, argues that this effect is mostly attributable to the introduction of new drugs that appear to have rather distinct resistance profiles from the drugs that previously existed in their class.

The model that we used to generate data greatly simplifies the complex nature of mutational effects. In particular, it assumes that the effects of mutations are independent of the genetic background and thus neglects that mutations may interact to determine fitness. Given the level of simplification it is remarkable how well the experimental and the simulated data agree with regard to the overall similarity of orientation and shape of the corresponding clouds of points provided that parameters are chosen appropriately. This argues that an additive model of mutational effects on log RC is able to describe the experimental data well in a statistical sense. However, it does not allow to infer that non-additive interactions between mutations are generally absent for two reasons. First, we have compared only a single model to the experimental data; therefore, we cannot exclude that there are other models that fit the data as well or even better. Second, the quality of a fit justifies the underlying assumption of a model in a statistical sense but not necessarily in a biological sense.

To assess the robustness of our findings we also produced simulated data based on modified models. To this end, we tested whether our findings are robust with regard to using other random distributions than the normal distribution for the three components underlying the mutational effects on log RC. In particular, it is plausible that the non-specific component of the mutational effect is typically deleterious, while the class- or drug-specific components are often beneficial. To account for this we assumed a negative exponential distribution for the first component and positive exponential distributions for the second and the third components. We also tested whether the fact that we have an uneven number of drugs in each class alters the results qualitatively by restricting the analysis to only four drugs per class. In all cases the results were qualitatively robust with regard to the main result of our analysis, namely that the class-specific component exceeds the drug-specific component for the NNRTIs and PIs, but not for the NRTIs (fig. 4.6).

In terms of the PCA, a potent combination therapy will be characterized by a non-overlapping fitness profile, i.e. by arrows that fan out maximally in the biplots. In fact, this is, to a considerable extent, reflected in current guidelines for combination therapies (Hammer et al., 2008), which recommend that an initial regimen should consist of a combination of EFV or an RTV boosted PI such as LPV together with two NRTIs (typically TFV and

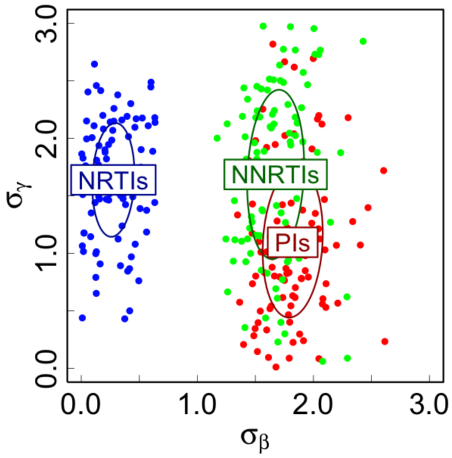


Figure 4.6: This figure corresponds to fig. 4.4–*B*, but here restricted to four randomly chosen drugs for each drug class.

FTC or ABC and 3TC). According to PCA, NRTIs show the strongest drug-specific effects, and thus it makes sense to use two of them as they allow the greatest opportunity to combine distinct profiles within a drug class. The A5202 study of the AIDS clinical trial group (Sax et al., 2008) showed that patients treated with the combination ABC/3TC had a shorter time to virologic failure than patients treated with TFV/FTC. It is interesting to note that PCA shows that TFV and FTC have more distinct profiles than ABC and 3TC (fig. 4.3). Obviously, many aspects of treatment such as side effects, simplicity of therapy, and pill number do affect the choice of the initial regimen but do not enter the data analyzed here. Nevertheless, the recommended combination regimens do appear to reflect the need to combine drugs that show a distinct profiles in the biplots (fig. 4.3).

The analysis presented here is based on RC measurement only and does not include any information about presence or absence of specific mutations in viral samples. The finding that 3TC, FTC and ABC cluster is expected as these drugs share the M184V resistance mutation (Turner et al., 2003; White et al., 2002) and thus validates the application of PCA on RC values as a means to delineate shared resistance profiles between drugs and drug classes. PCA, however, goes beyond more empirical approaches in that it allows to quantify the similarity between drugs in large and complex data sets.

Our results show that PCA is a powerful tool to reveal underlying structure in fitness and resistance patterns among a large number of HIV-1 samples. PCA allows us to infer the relative magnitude of the non-specific, the class-specific, and the drug-specific components of the mutational effects when used in conjunction with simulated data. Although simpler approaches such as pairwise regression of RC values in different environments may be more intuitive, PCA has the advantage that it reveals and quantifies hierarchical structure in the data that might easily be missed with pairwise regression. The hierarchy in the data structure reveals the dominant role of the drug class in determining variance in RC across different selective environments.

4.5 Supplementary figures

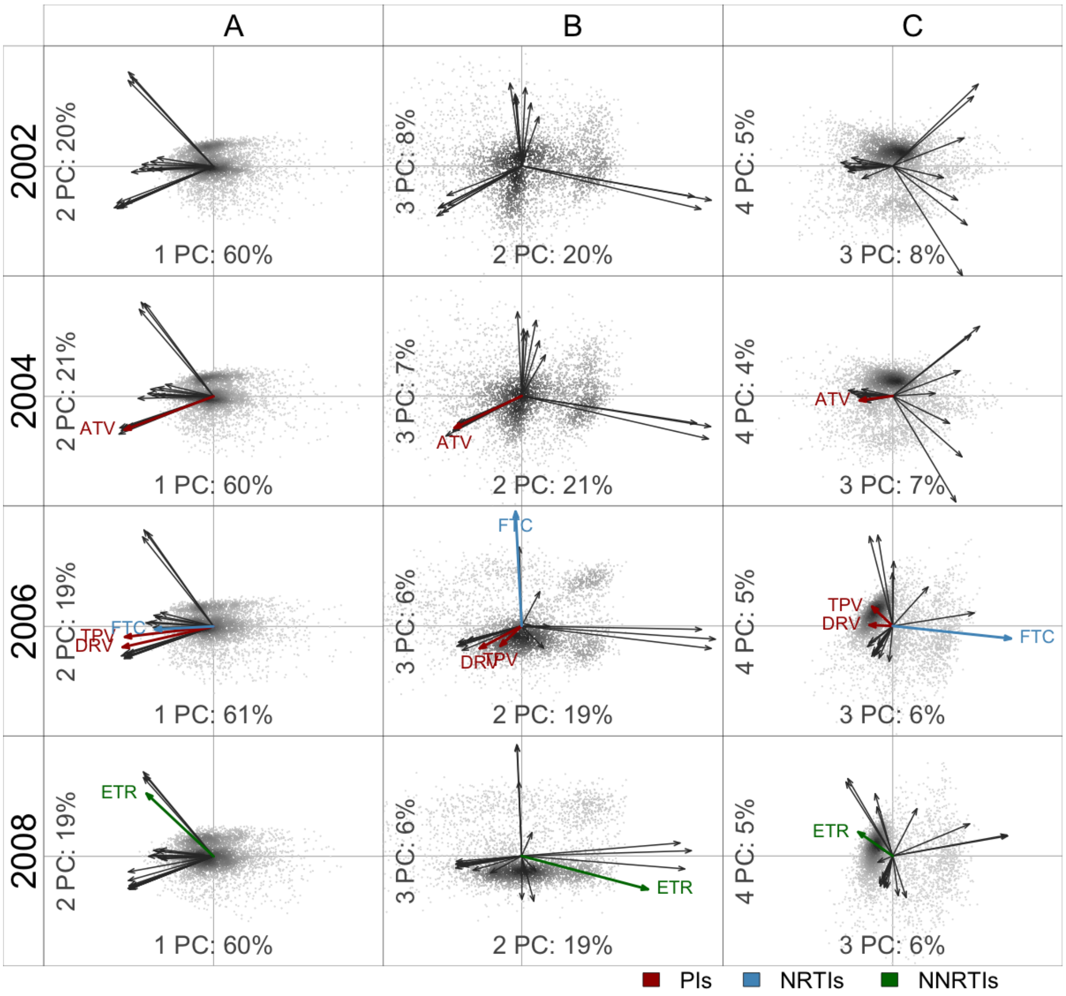


Figure 4.7: Biplots of the experimental data as shown in fig. 4.1, but here highlighting new drugs that were introduced between 2002 and 2008. Acronyms for the drugs or drug classes are found in *Methods*.

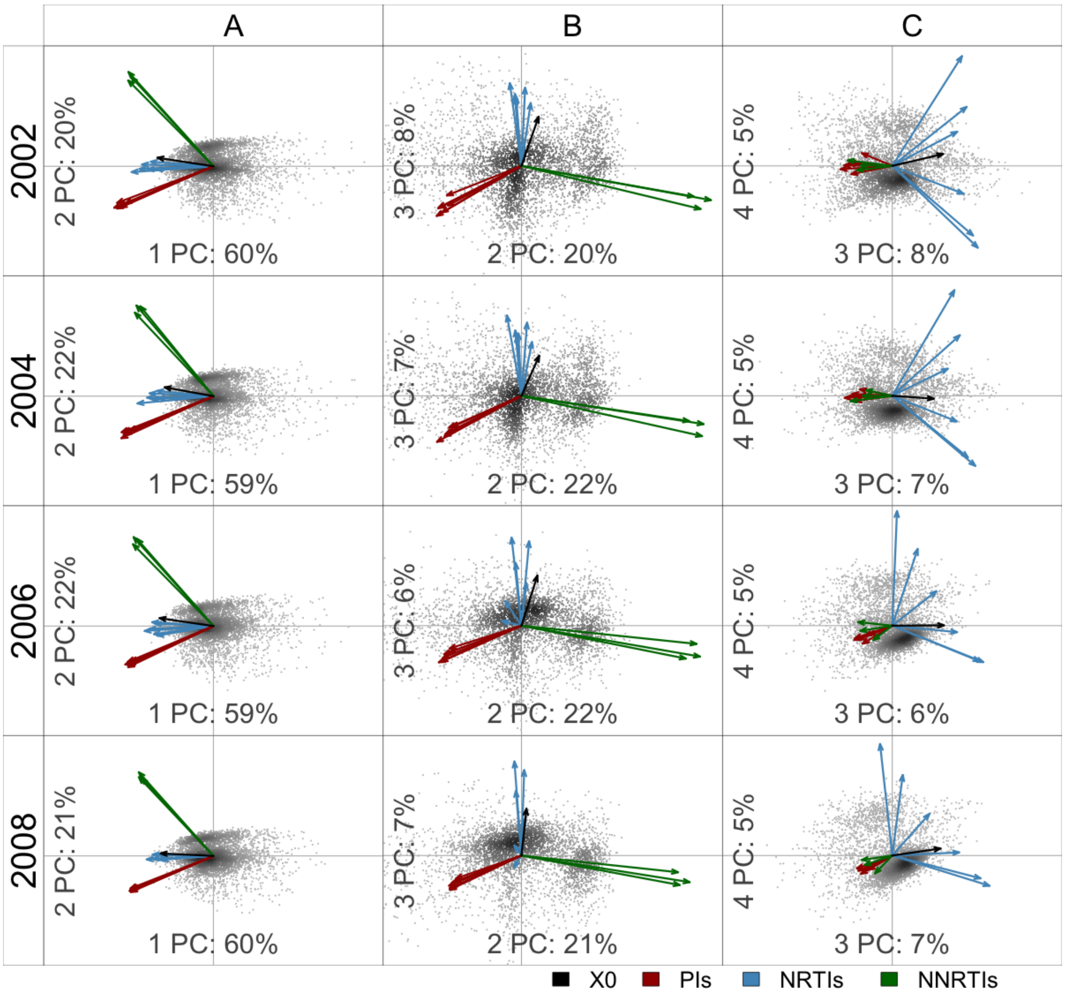


Figure 4.8: Biplots illustrating the PCA of the experimental data as shown fig. 4.1, but here restricted to only those 15 drugs for which data were available from 2002 onwards.

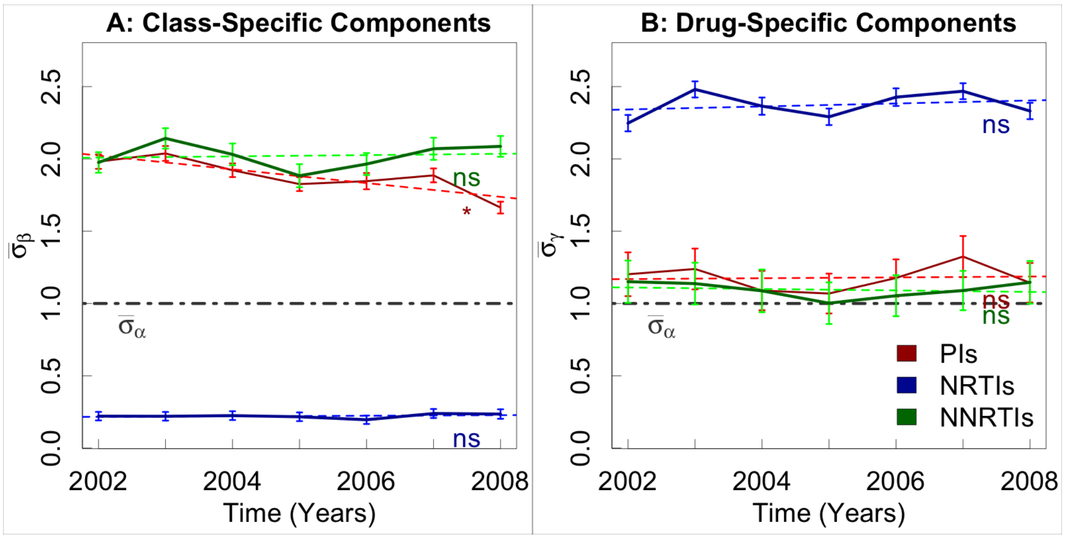


Figure 4.9: This figure corresponds to fig. 4.5, but here restricted only to those 16 environments (15 drug-containing and one drug-free environment) for which data was available from 2002 onwards.

CHAPTER

FIVE

OUTLOOK AND GENERAL CONCLUSION

For me there are no answers, only questions, and I am grateful that the questions go on and on. I don't look for an answer, because I don't think there is one. I'm very glad to be the bearer of a question.

P. L. Travers

5.1 Outlook

Several questions arose throughout the previous chapters and a few of them deserve further attention. They are presented in the following text by their order of appearance in the different chapters.

5.1.1 Visual representations of the landscape

An interesting result in chapter 2 involved the R^2 between the pairwise Hamming distances between sequences and the pairwise euclidean distances of the corresponding planar point configuration found by multidimensional scaling. It is remarkable that 72% of the total variance of the sequences' Hamming distances can be captured in only two dimensions in terms of euclidean distances. Therefore, it would be interesting to investigate which factors have an effect on the value of the R^2, with the hope of increasing it even further. In addition, it would also be interesting to investigate to what extent it would be possible to generate sequence alignments defining fitness landscapes of the same level of complexity as the one defined by experimental sequence data. Accordingly, it would be interesting to investigate how fitness landscapes defined by mathematical models (such as Kauffman's NK fitness landscape (Kauffman and Levin, 1987; Kauffman and Weinberger, 1989)) compare to the HIV landscape.

5.1.2 Odds for the evolution of recombination

A question which has arisen in chapter 3 is to know to what degree the setup of the genetic algorithm can reproduce the previous results of Keightley and Otto (Keightley and Otto, 2006) when using identical parameters — for the same the sequence length and for the same fitness functions. It would be also worthwhile to attempt to restrict the degree of mutation freedom of the sequences and allow only highly represented alleles to occur. In other words, to restrain sequences from diverging to unrealistic regions of the sequence space. The level of restrictiveness could also be subject to investigation. The more restricted the sequence space is, the lower the risk of evolving unrealistic allele combinations. Limiting the fitness landscape to the extreme of only five loci would also allow direct comparison with de Visser's results (de Visser et al., 2009). If a restriction of the sequence space would solve the issue of evolving unrealistic sequences, then the genetic algorithm could be used to look into other possibly relevant parameters of the evolutionary dynamics of HIV. An ambitious extension of this project would be to include a more individual oriented modeling of the evolutionary process and allow structure at the level of the viral population, either to mimic the evolution in different compartments of the same individual, or the evolution at the level of a metapopulation of different individuals.

5.1.3 Patterns of resistance and cross-resistance

In chapter 4, the similarity between the principal component analysis (PCA) patterns of experimental and simulated data was measured on the basis of an ad-hoc method. However, comparison of PCAs is a standard problem in G-matrix theory (used to study pleiotropy) (Mezey and Houle, 2003). More specifically, Common Principal Component (CPC) analysis is a method which is commonly used to quantify the similarities between G-matrices, which essentially correspond to PCA characteristic matrices. Replacing the ad-hoc method with a more standard CPC could possibly increase the accuracy of the parameters' estimation. It would be also interesting to investigate whether clusters of sequences (represented by points in the biplots) share common patterns of drug resistance mutations.

5.2 Conclusion

Overall, this thesis presents a few results whose importance relies on the fact that they are based on an extremely rich HIV dataset. First, it was shown that simple visual representations of fitness landscapes can reveal important features of the landscape. In particular, it was shown that a fitness landscape based on HIV-1 sequence data was locally very rugged, which corroborates Kauffmann's *massif central* hypothesis. Despite the fact that the simulation of the evolution of sequence populations according to the ME and the MEEP models cannot be kept in reasonably realistic regions of the sequence space, it has been observed that recombination is generally favored. Finally, it has been shown that the relative fitness of viral sequences is conserved very well across the different drug environments which suggests that HIV fitness landscapes are very much alike across the different drug environments.

APPENDIX

A

INDIVIDUAL MUTATION-BASED MODELS OF THE FITNESS LANDSCAPE

Contents

Published as: Hinkley, T; Martins, JZR; Chappey, C; Haddad, M; Stawiski, E; Whitcomb, JM; Petropoulos, CJ; Bonhoeffer, SL. ***A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase.*** Nature Genetics, 2011. Hinkley et al. (2011)

A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase

Hinkley, T. Martins, J. Z. R. Chappey, C. Haddad, M. Whitcomb, J. M.
Stawiski, E. Petropoulos, C. J. Bonhoeffer, S.

A.1 Introduction

With more than 20 drugs currently licensed to treat HIV infection(Clercq, 2009) and over 200 mutations associated with resistance(Bennett et al., 2009; Clavel and Hance, 2004; Johnson et al., 2008; Shafer and Schapiro, 2008), it is increasingly difficult to develop a comprehensive understanding of HIV drug resistance. Resistance mutations differ in their potency to resist drug pressure (Petropoulos et al., 2000; Rhee et al., 2004), they vary in their degree of cross-resistance to different drugs or drug classes (Harrigan and Larder, 2002), and they differ in the fitness costs induced in the absence of treatment (Croteau et al., 1997; Mammano et al., 2000; Martinez-Picado et al., 1999). Moreover their effects depend to varying degree on the context of accompanying mutations (Bonhoeffer et al., 2004; Rhee et al., 2004). The quantitative dissection of the fitness effects of resistance mutations in presence or absence of drugs and in particular the determination how the effect of mutations depend on the presence or absence of other mutations thus represents a major challenge.

The delineation of epistatic interactions between mutations is not only a matter of the size of the data set. The combinatorial complexity of the genetic context in which any mutations appears explodes to a degree such that the estimation of the fitness effects is not feasible with standard statistical approaches, because the number of parameters to be estimated easily outnumbers the number of data points available even for the largest data sets. Problems in which the combinatorial complexity overwhelms standard methods of parameter inference are a common challenge in systems biology, and various approaches have been developed that allow a reliable parameter estimation under conditions that lead to overfitting with standard statistical approaches. To overcome the problem of the large number of parameters and to account for non-normality in the error-structure we employ here generalised kernel ridge regression (GKRR), a regression method which, in essence, penalises against parameters that have low explanatory power. We use GKRR to quantify the fitness effects of amino acid variants using a data set that measures *in vitro* fitness of 70,081 HIV-1 samples in the absence of drugs and in the presence of 15 different individual drugs. The samples were obtained from HIV-1 subtype B infected patients undergoing routine drug-resistance testing. Our approach allows the reconstruction of an approximate fitness landscape of the HIV protease (PR) and reverse transcriptase (RT) and thus offers the first quantitative description of a large, realistic and biologically relevant fitness landscape.

In vitro fitnesses of viral isolates are measured by replicative capacity and sequenced in amino acids 1 to 99 of PR and 1 to 305 of RT. We quantify the fitness effects that are attributable to individual amino acid variants (main effects) and to pairwise epistatic effects between such variants (interactions) using GKRR. In particular we fit two alternative models: (i) The ME model, which predicts fitness only on the basis of the main effects and (ii) the MEEP model, which predicts fitness using both main effects and interactions. We applied GKRR because the size of the data-set used is too great for current implementations of other regularisation techniques such as the LASSO (Efron et al., 2002) or Dantzig selector (Candes and Tao, 2007) .

A.2 Results

Figure A.1 shows the predictive power of the ME and MEEP models based on a 6-fold cross-validation by randomly subdividing the data set into training and test sets of 65,000 and 5,000 independent virus samples, respectively. The goodness of the fit is quantified by the percentage deviance explained. Deviance is the standard measure of goodness of fit in generalised models (i.e. in models with non-normal error structure), and is analogous to the R^2 of linear models with normal error structure (Nelder and Wederburn, 1972). The predictive power across the environments ranges from 35.0% to 65.9% for MEEP and from 26.8% to 57.9% for ME. MEEP has an average predictive power of 54.8% across all 16 environments. MEEP represents on average an 18.3% improvement in predictive power relative to ME. Note, that in a regularised regression such as the GKRR, increase in predictive power measured by cross-validation is the appropriate model validation method. Hence, the substantial increase in predictive power of the MEEP over the ME model validates the inclusion of epistatic terms irrespective of their large number. Our kernelised approach allows to include higher order epistatic interactions without substantial increases in computational requirements. Including three-way epistasis marginally decreases predictive power (data not shown). This decrease is due to the substantial increase in effective coefficients and does not imply that higher order epistatic interactions do not contribute to fitness.

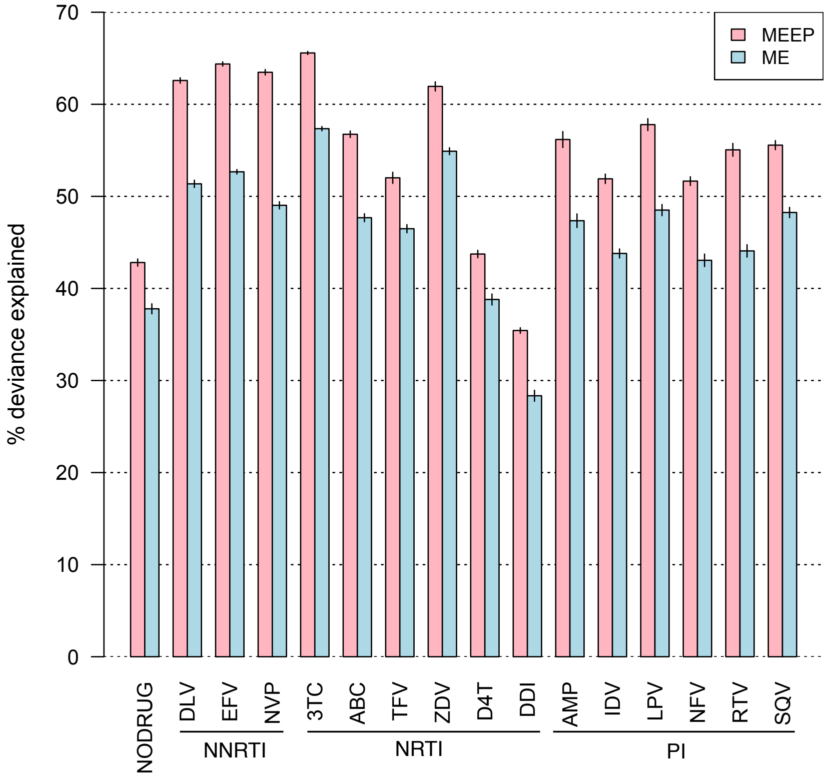


Figure A.1: Analysis of predictive power. The figure shows the predictive power of the ME and MEEP model in a drug free and 15 drug containing environments (for acronyms of drugs see chapter 4). The predictive power is measured by the percentage deviance explained in a cross-validation data set based on 5,000 independent virus samples. The bars represent mean and the whiskers the standard errors from a six-fold cross-validation. The MEEP model outperforms the ME model in all environments.

A analogous approach was taken to investigate the relative role of intra- versus intergenic epistasis (i.e. interactions within PR or RT versus interactions between PR and RT). Four models were fitted: ME only, ME + intragenic epistasis, ME + intergenic epistasis and the full MEEP model (see Figure A.2). Including intragenic epistasis consistently leads a much greater gain of predictive power than including intergenic epistasis. The ME + intragenic epistasis model is generally as good, and sometimes even marginally better, than the MEEP model, which indicates that adding intergenic epistatic effects to the ME + intragenic epistasis model does not further improve the predictive power. Decreases in predictive power are attributable to the fact that adding a large number of unnecessary parameters to a model can result in a reduction in predictive power in GKRR.

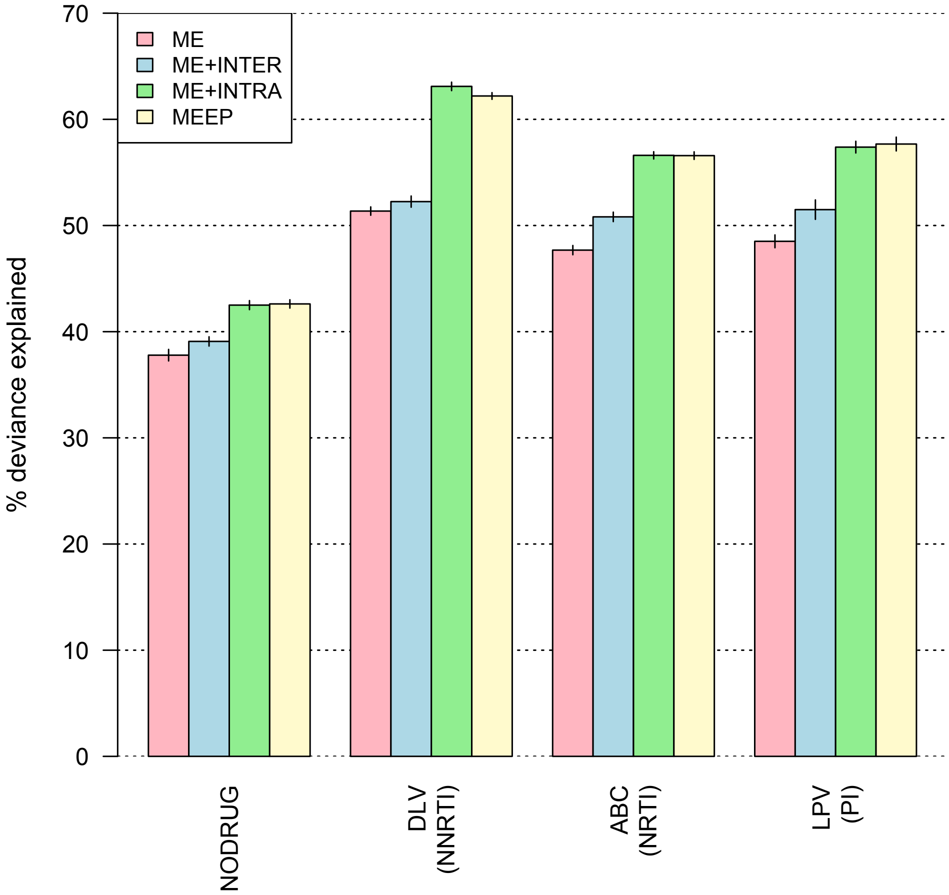


Figure A.2: Analysis of predictive power of different epistatic models for four representative environments. The figure shows that most of the predictive power attributable to epistasis is in fact attributable to intra- rather than intergenic epistatic interactions. In the NNRTI environment adding intergenic epistasis decreases predictive power. This decrease reflects that adding a large number of parameters with little or no explanatory power can reduce the predictive power of GKRR. The bars represent mean and the whiskers the standard errors from a six-fold cross-validation.

To verify that the estimates of the MEEP model are meaningful we obtained sequences of PR and RT of treated and untreated patients from Stanford HIV Drug Resistance Database (Shafer, 2006) and determined the change of frequency of amino acid variants in treated versus untreated patients. The change of frequency of amino acid variants is significantly correlated with the fitness gain of amino acid variants in presence versus absence of drugs relative to the consensus sequence ($p < 10^{-16}$ and $\rho = 0.30$, Spearman rank).

Because protein structure and epistasis are interrelated (Bershtein et al., 2006; Halabi et al., 2009) we investigated the relation between epistasis in the drug-free environment and protease structure as an independent verification that the estimates of the 802,611 epistatic effects are biologically meaningful. Fig. A.3 shows the strength of the epistatic effects between amino acid residues of the HIV-1 PR, revealing significant enrichment in epistatic interactions in the flap elbow, the cantilever and the fulcrum, structural units that have previously been described as being important to protein function (Hornak et al., 2006). Bootstrap analysis by random shuffling of the protein sequence reveals that epistasis is significantly enriched both within these structural domains and between the structural domains and the rest of the protein ($p < 10^{-5}$). Moreover, in accordance with expectation the strength of the epistatic interactions between amino acid residues correlates with physical proximity in 3D structure of PR ($p = 0.00857$ based on 100'000 bootstrap repeats, see supplementary figure A.5). The significant correlation between epistasis and secondary structure or proximity demonstrates that the estimated epistatic effects are biologically meaningful. Such correlations could not have been produced artifactually as our procedure includes no structural information for parameter estimation.

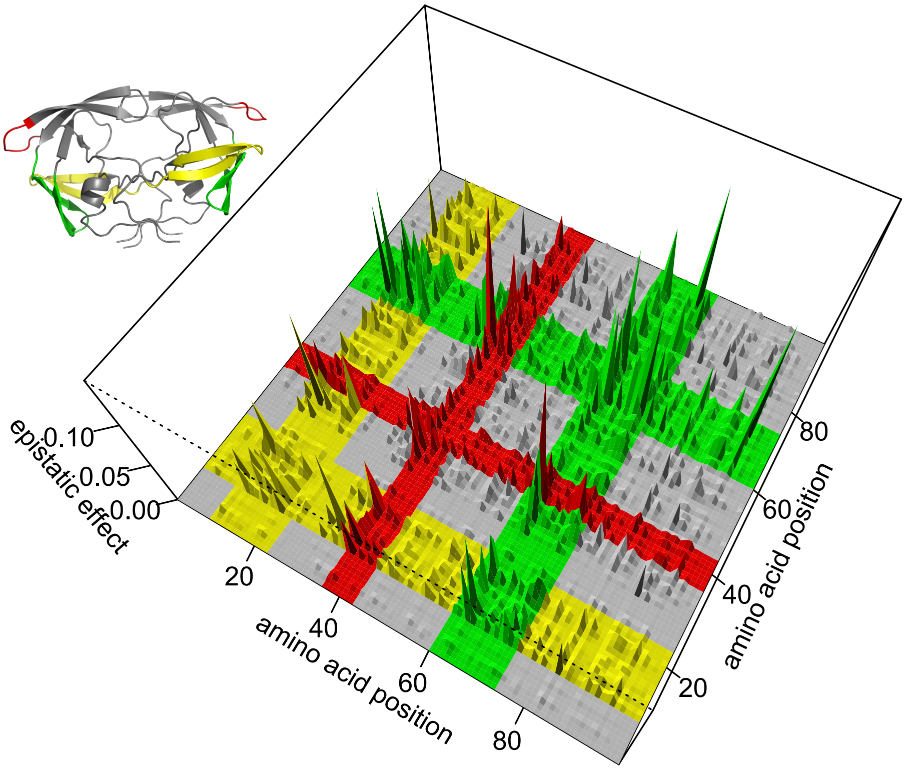


Figure A.3: Cumulative strength (CS) of the absolute epistatic effects in the HIV-1 PR as measured in the drug-free environment. The cumulative effect between two positions is calculated as the sum over the absolute values of all epistatic interactions between the amino acid variants at those positions as estimated by the MEEP model. We plot CS$^{1.5}$ to enhance visual clarity

. The regions corresponding to the flap elbow, fulcrum and cantilever, colored in red, yellow, and green, respectively, are significantly enriched in epistasis (see supplementary figure A.4). The inset shows the structure of the HIV-1 PR (Protein Data Bank ID 1A30, rendered with PyMOL). The region enriched in epistatic interaction, corresponding to the flap elbow, is somewhat larger than the literature description of this region (Hornak et al., 2006).

A.3 Discussion

Previous studies on epistasis in viruses did not allow a comprehensive quantification of individual fitness effects and epistatic interactions, because they focussed either on a limited set of interactions (Sanjuan et al., 2004), made use of sequence data only (Chen et al., 2004), or did not correct for the effect of the genetic background (Bonhoeffer et al., 2004). Our study demonstrates that despite the combinatorial complexity of the problem biologically meaningful estimates for main effects and epistatic interactions can be obtained from large data sets that link fitness measurements to genotype. We verified the estimated effects using independent data. First, we showed that models including epistatic interactions explain on average 54.8% of the deviance in fitness across the 16 different environments based on six-fold cross-validation. Second, we found a highly significant correlation between the change of the estimated main effects in presence versus absence of drugs and the change in frequency of the corresponding amino acid variants in treated versus untreated patients based on independent data from the Stanford HIV drug resistance database (Shafer, 2006). Finally, we found a correlation between epistasis and PR structural domains or physical proximity in the 3D structure of PR.

Ever since the synthesis of Darwinian evolution with genetic inheritance in the early 20^{th} century, the debate about the relative role of epistasis and main effects in determining fitness has remained at the heart of evolutionary genetics (Provine, 1971; Wolf et al., 2000) and, with the advent of systems biology, it is possible to measure these epistatic effects more comprehensively(Costanzo et al., 2010; de Visser and Elena, 2007; Jasnos and Korona, 2007; Kouyos et al., 2007; Yeh et al., 2009). Supporting Sewall Wright's view of the dominant role of epistasis (Provine, 1971; Wolf et al., 2000), we find that epistasis and, in particular, intragenic epistasis is crucial in determining fitness. For our data set the inclusion of epistatic interactions improves the predictive power by an average of 18.3% across all environments. Our approach provides us with a predictive model for realistic fitness landscapes, opening up new avenues to study evolutionary adaptation on complex fitness landscapes and to simulate the evolution of drug resistance.

A.4 Supplementary Figures

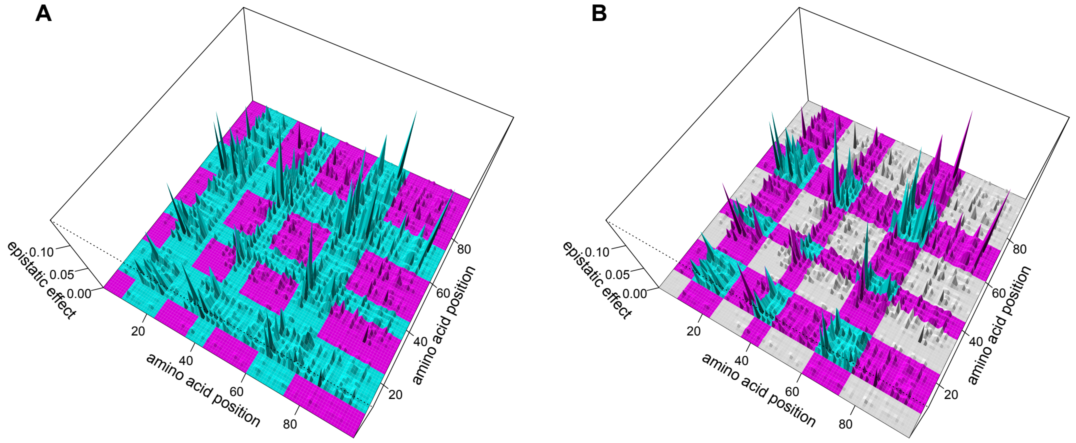


Figure A.4: Statistical test of enrichment of epistasis in fulcrum, cantilever and flap elbow in the HIV-1 protease. The plots are identical to figure 3 in the main text except for the coloring. For both plots we test whether interactions are enriched in the cyan compared to the magenta regions. Panel A thus compares the epistatic interactions between fulcrum, cantilever, and flap elbow and the rest of the protein to all other remaining interactions. The mean absolute epistasis in the cyan and magenta regions is 0.0202 and 0.00856, respectively. Bootstrap analysis by random shuffling of the positions in the protease reveals that the mean in the cyan region is significantly higher ($p < 10^{-5}$ based on 100'000 repeats). In panel B the cyan region corresponds to the interactions within and between the fulcrum, cantilever, and flap elbow and the magenta region corresponds to the interactions between these regions and the rest of the protease. The means of the cyan and magenta regions are 0.0355 and 0.0161. Bootstrap analysis also confirms that these means are significantly different ($p < 10^{-5}$ based on 100'000 repeats).

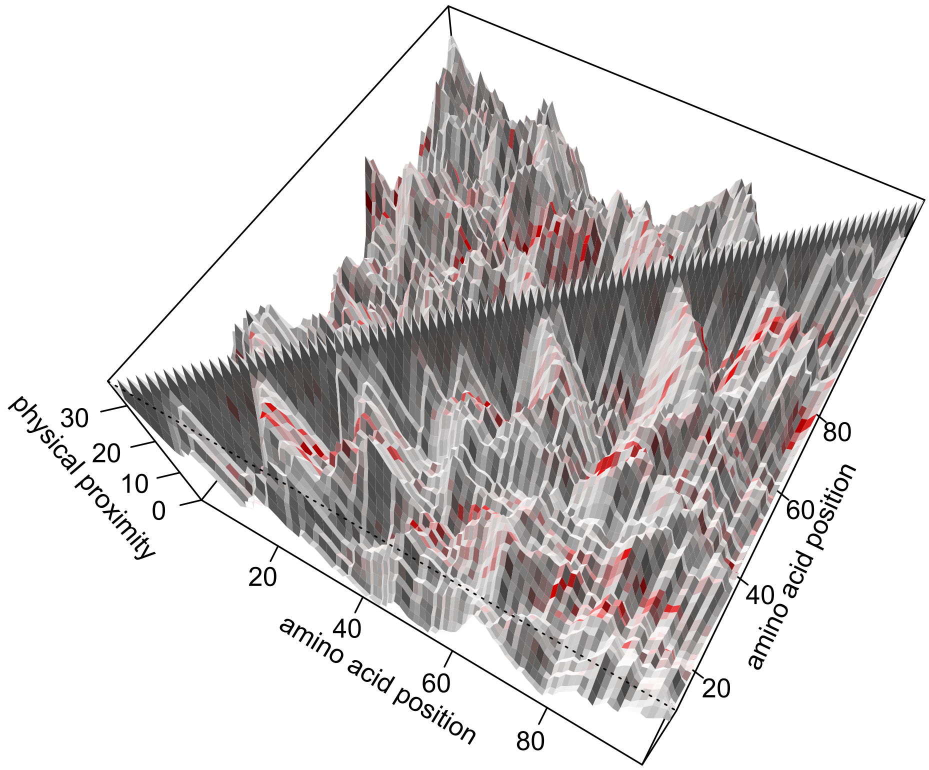


Figure A.5: Cumulative absolute epistatic effects versus physical proximity (Å) in the HIV-1 protease. The strength of the epistatic effect is measured as in figure 3 in the main text. Physical proximity is correlated to the cumulative absolute epistatic effect ($p = 0.00857$ based on 100’000 bootstrap repeats).

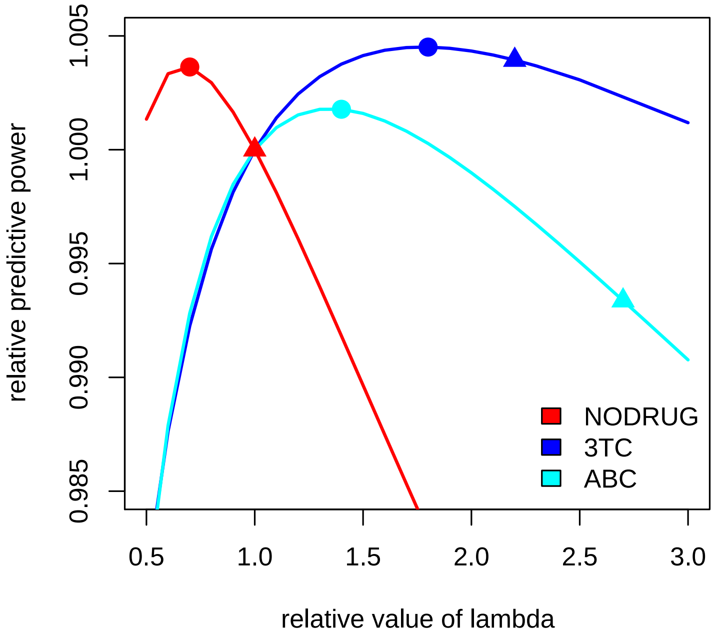


Figure A.6: Relative predictive power under varying lambda. Lambda was varied from its position as calculated with the square root approximation and the corresponding predictive power (relative to the predictive power for the calculated lambda) was measured against the cross validation set under environments NODRUG , 3TC, and ABC. The maximum possible predictive power is indicated by a circle (for optimal lambda choice). Lambda as would be calculated using a full GKRR for each bisection interval is shown by a triangle. NODRUG shows the same prediction for lambda, 3TC shown a better prediction for lambda and ABC shows a worse prediction. Important to note is that in all cases, the prediction (both for the square root approximation and for a GKRR approximation) for the final lambda differs from the optimal lambda, in predictive power, by less than 1%. It can therefore be concluded that our square root approximation for lambda is robust.

BIBLIOGRAPHY

Ashlock, D. and J. Schonfeld (2005). Nonlinear projection for the display of high dimensional distance data. In *The 2005 IEEE Congress on Evolutionary Computation, 2005.*, Volume 3, pp. 2776–2783. IEEE.

Barbour, J., T. Wrin, R. Grant, J. Martin, M. Segal, C. Petropoulos, and S. Deeks (2002). Evolution of phenotypic drug susceptibility and viral replication capacity during long-term virologic failure of protease inhibitor therapy in human immunodeficiency virus-infected adults. *Journal of Virology 76*(21), 11104–11112.

Barton, N., D. Briggs, J. Eisen, D. Goldstein, and N. Patel (2007). *Evolution* (1st ed.). Cold Spring Harbor Laboratory Press.

Barton, N. and B. Charlesworth (1998). Why sex and recombination? *Science 281*(5385), 1986.

Beaumont, M., W. Zhang, and D. Balding (2002). Approximate bayesian computation in population genetics. *Genetics 162*(4), 2025–2035.

Beerenwinkel, N. (2004). *Computational analysis of HIV drug resistance data*. Shaker.

Bennett, D., R. Camacho, D. Otelea, D. Kuritzkes, H. Fleury, M. Kiuchi, W. Heneine, R. Kantor, M. Jordan, J. Schapiro, et al. (2009). Drug resistance mutations for surveillance of transmitted HIV-1 drug-resistance: 2009 update. *PLoS ONE 4*(3).

Bershtein, M. Segal, R. Bekerman, N. Tokuriki, and D. Tawfik (2006, Jan). Robustness–epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature*.

Bonhoeffer, S., C. Chappey, N. T. Parkin, J. M. Whitcomb, and C. J. Petropoulos (2004, Nov). Evidence for positive epistasis in HIV-1. *Science 306*(5701), 1547–50.

Brown, A. (1997). Analysis of HIV-1 env gene sequences reveals evidence for a low effective number in the viral population. *Proceedings of the National Academy of Sciences 94*(5), 1862.

Candes, E. and T. Tao (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics 35*(6), 2313–2351.

Chen, L., A. Perlina, and C. J. Lee (2004, Apr). Positive selection detection in 40,000 human immunodeficiency virus (HIV) type 1 sequences automatically identifies drug resistance and positive fitness mutations in HIV protease and reverse transcriptase. *J. Virol. 78*, 3722–3732.

Chessel, D., A. B. Dufour, and J. Thioulouse (2004). The ade4 Package–I: One-table Methods. *R News 4*(1), 5–10.

Clavel, F. and A. J. Hance (2004, Mar). HIV drug resistance. *N. Engl. J. Med. 350*(10), 1023–35.

Clercq, E. D. (2009, Apr). Anti-HIV drugs: 25 compounds approved within 25 years after the discovery of HIV. *Int. J. Antimicrob. Agents 33*(4), 307–20.

Costanzo, M., A. Baryshnikova, J. Bellay, Y. Kim, E. D. Spear, C. S. Sevier, H. Ding, J. L. Y. Koh, K. Toufighi, S. Mostafavi, J. Prinz, R. P. S. Onge, B. VanderSluis, T. Makhnevych, F. J. Vizeacoumar, S. Alizadeh, S. Bahr, R. L. Brost, Y. Chen, M. Cokol, R. Deshpande, Z. Li, Z.-Y. Lin, W. Liang, M. Marback, J. Paw, B.-J. S. Luis, E. Shuteriqi, A. H. Y. Tong, N. van Dyk, I. M. Wallace, J. A. Whitney, M. T. Weirauch, G. Zhong, H. Zhu, W. A. Houry, M. Brudno, S. Ragibizadeh, B. Papp, C. Pal, F. P. Roth, G. Giaever, C. Nislow, O. G. Troyanskaya, H. Bussey, G. D. Bader, A.-C. Gingras, Q. D. Morris, P. M. Kim, C. A. Kaiser, C. L. Myers, B. J. Andrews, and C. Boone (2010). The genetic landscape of a cell. *Science 327*(5964), 425–431.

Cox, T. and M. Cox (1994). *Multidimensional scaling*. Monographs on statistics and applied probability. Chapman & Hall.

Croteau, G., L. Doyon, D. Thibeault, G. McKercher, L. Pilote, and D. Lamarre (1997, Feb). Impaired fitness of human immunodeficiency virus type 1 variants with high-level resistance to protease inhibitors. *J. Virol. 71*(2), 1089–96.

Daar, E., K. Kesler, T. Wrin, C. Petropoulo, M. Bates, A. Lail, N. Hellmann, E. Gomperts, S. Donfield, et al. (2005). HIV-1 pol replication capacity predicts disease progression. *Aids 19*(9), 871.

de Visser, J. and S. Elena (2007). The evolution of sex: empirical insights into the roles of epistasis and drift. *Nature Reviews Genetics 8*(2), 139–149.

de Visser, J., S. Park, and J. Krug (2009). Exploring the effect of sex on empirical fitness landscapes. *The American Naturalist*, 15–30.

Dykes, C., J. Wang, X. Jin, V. Planelles, D. An, A. Tallo, Y. Huang, H. Wu, and L. Demeter (2006). Evaluation of a multiple-cycle, recombinant virus, growth competition assay that uses flow cytometry to measure replication efficiency of human immunodeficiency virus type 1 in cell culture. *Journal of Clinical Microbiology 44*(6), 1930.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2002). Least angle regression. *Annals of Statistics 32*, 407–499.

Egger, M., B. Hirschel, P. Francioli, P. Sudre, M. Wirz, M. Flepp, M. Rickenbach, R. Malinverni, P. Vernazza, and M. Battegay (1997). Impact of new antiretroviral combination therapies in HIV infected patients in Switzerland: prospective multicentre study. *British Medical Journal 315*(7117), 1194–1199.

Fauci, A. et al. (2003). HIV and AIDS: 20 years of science. *Nature Medicine 9*(7), 839–843.

Frankel, A. and J. Young (1998). HIV-1: fifteen proteins and an RNA. *Annual review of biochemistry 67*(1), 1–25.

Fruchterman, T. and E. Reingold (1991). Graph drawing by force-directed placement. *Software: Practice and experience 21*(11), 1129–1164.

Gabriel, K. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika 58*(3), 453–467.

Gansner, E. and S. North (2000). An open graph visualization system and its applications to software engineering. *Software-Practice & Experience 30*(11), 1203–1233.

Gavrilets, S. (1997). Evolution and speciation on holey adaptive landscapes. *Trends in ecology & evolution 12*(8), 307–12.

Gavrilets, S. (2004). Fitness landscapes and the origin of species. *Austral Ecology 30*(5), 610–611.

Gentleman, R. and Biocore (2006). *Geneplotter: Graphics related functions for Bioconductor*. R package version 1.18.0.

Gnanadesikan, R. (1977). *Methods for statistical data analysis of multivariate observations*. Wiley New York.

Gower, J. and D. Hand (1996). *Biplots*. Chapman & Hall/CRC.

Halabi, N., O. Rivoire, S. Leibler, and R. Ranganathan (2009, Jan). Protein sectors: evolutionary units of three-dimensional structure. *Cell*.

Hammer, S., J. Eron Jr, P. Reiss, R. Schooley, M. Thompson, S. Walmsley, P. Cahn, M. Fischl, J. Gatell, M. Hirsch, et al. (2008). Antiretroviral treatment of adult HIV infection: 2008 recommendations of the International AIDS Society-USA panel. *JAMA 300*(5), 555.

Hammer, S., D. Katzenstein, M. Hughes, H. Gundacker, R. Schooley, R. Haubrich, W. Henry, M. Lederman, J. Phair, M. Niu, et al. (1996). A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine 335*(15), 1081–1090.

Harrigan, P. R. and B. A. Larder (2002, Mar). Extent of cross-resistance between agents used to treat human immunodeficiency virus type 1 infection in clinically derived isolates. *Antimicrobial Agents and Chemotherapy 46*(3), 909–12.

Hartl, D., A. Clark, et al. (1997). *Principles of population genetics*, Volume 7. Sinauer associates Sunderland, Massachusetts.

Hinkley, T., J. Martins, C. Chappey, M. Haddad, E. Stawiski, J. Whitcomb, C. Petropoulos, and S. Bonhoeffer (2011). A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase. *Nature genetics 43*(5), 487–489.

Hornak, V., A. Okur, R. Rizzo, and C. Simmerling (2006). HIV-1 protease flaps spontaneously open and reclose in molecular dynamics simulations. *Proceedings of the National Academy of Sciences (USA) 103*(4), 915–920.

Iles, M., K. Walters, and C. Cannings (2003). Recombination can evolve in large finite populations given selection on sufficient loci. *Genetics 165*(4), 2249.

Jasnos, L. and R. Korona (2007, Apr). Epistatic buffering of fitness loss in yeast double deletion strains. *Nat Genet 39*(4), 550–4.

Johnson, V. A., F. Brun-Vezinet, B. Clotet, H. F. Gunthard, D. R. Kuritzkes, D. Pillay, J. M. Schapiro, and D. D. Richman (2008, Dec). Update of the drug resistance mutations in HIV-1. *Topics in HIV medicine 16*(5), 138–45.

Jolliffe, I. (1986). *Principal component analysis*. Springer-Verlag New York.

Kamada, T. and S. Kawai (1989). An algorithm for drawing general undirected graphs. *Information processing letters 31*(1), 7–15.

Kaplan, J. (2008). The end of the adaptive landscape metaphor? *Biology & Philosophy 23*(5), 625–638.

Kauffman, S. (1993). *The origins of order: Self organization and selection in evolution*. Oxford University Press, USA.

Kauffman, S. and S. Levin (1987). Towards a general theory of adaptive walks on rugged landscapes. *Journal of theoretical Biology 128*(1), 11–45.

Kauffman, S. and E. Weinberger (1989). The NK model of rugged fitness landscapes and its application to maturation of the immune response. *Journal of theoretical biology 141*(2), 211–245.

Keightley, P. and S. Otto (2006). Interference among deleterious mutations favours sex and recombination in finite populations. *Nature 443*(7107), 89–92.

Kondrashov, A. (1993). Classification of hypotheses on the advantage of amphimixis. *Journal of Heredity 84*(5), 372–387.

Kouyos, R., C. Althaus, and S. Bonhoeffer (2006). Stochastic or deterministic: what is the effective population size of HIV-1? *TRENDS in Microbiology 14*(12), 507–511.

Kouyos, R., G. Leventhal, T. Hinkley, M. Haddad, J. Whitcomb, C. Petropoulos, and S. Bonhoeffer (2012). Exploring the complexity of the HIV-1 fitness landscape. *PLoS Genetics 8*(3), e1002551.

Kouyos, R., S. Otto, and S. Bonhoeffer (2006). Effect of varying epistasis on the evolution of recombination. *Genetics 173*(2), 589.

Kouyos, R. D., O. K. Silander, and S. Bonhoeffer (2007, Jun). Epistasis between deleterious mutations and the evolution of recombination. *Trends Ecol Evol 22*(6), 308–15.

Kruskal, J. (1964). Nonmetric multidimensional scaling: a numerical method. *Psychometrika 29*(2), 115–129.

Li, X., G. Cheng, and L. Lu (2000). Comparison of spatial interpolation methods. *Advance in Earth Sciences 3*.

Lozovsky, E., T. Chookajorn, K. Brown, M. Imwong, P. Shaw, S. Kamchonwongpaisan, D. Neafsey, D. Weinreich, and D. Hartl (2009). Stepwise acquisition of pyrimethamine resistance in the malaria parasite. *Proceedings of the National Academy of Sciences 106*(29), 12025–12030.

Mammano, F., V. Trouplin, V. Zennou, and F. Clavel (2000, Sep). Retracing the evolutionary pathways of human immunodeficiency virus type 1 resistance to protease inhibitors: virus fitness in the absence and in the presence of drug. *J. Virol. 74*(18), 8524–31.

Martinez-Picado, J., A. V. Savara, L. Sutton, and R. T. D'Aquila (1999, May). Replicative fitness of protease inhibitor-resistant mutants of human immunodeficiency virus type 1. *J. Virol. 73*(5), 3744–52.

Martins, J. Z., C. Chappey, M. Haddad, J. M. Whitcomb, E. Stawiski, C. J. Petropoulos, and S. Bonhoeffer (2010). Principal component analysis of general patterns of hiv-1 replicative fitness in different drug environments. *Epidemics 2*(2), 85 – 91.

McCandlish, D. (2011). Visualizing fitness landscapes. *Evolution 65*(6), 1544–1558.

Mezey, J. and D. Houle (2003). Comparing G matrices: are common principal components informative? *Genetics 165*(1), 411–425.

Miao, H., C. Dykes, L. Demeter, J. Cavenaugh, S. Park, A. Perelson, and H. Wu (2008). Modeling and Estimation of Kinetic Parameters and Replicative Fitness of HIV-1 From Flow-Cytometry-Based Growth Competition Experiments. *Bulletin of Mathematical Biology 70*(6), 1749–1771.

Mocroft, A., S. Vella, T. Benfield, A. Chiesi, V. Miller, P. Gargalianos, M. d'Arminio, I. Yust, J. Bruun, A. Phillips, et al. (1998). Changing patterns of mortality across Europe in patients infected with HIV-1. EuroSIDA Study Group. *Lancet 352*(9142), 1725.

Nelder, J. and R. Wederburn (1972). Generalized linear models. *J. Roy. Stat. Soc. A 135*(3), 370–384.

Orr, H. (2009). Fitness and its role in evolutionary genetics. *Nature Reviews Genetics 10*(8), 531–539.

Østman, B., A. Hintze, and C. Adami (2010). Critical properties of complex fitness landscapes. *Arxiv preprint arXiv:1006.2908*.

Palella, F., K. Delaney, A. Moorman, M. Loveless, J. Fuhrer, G. Satten, D. Aschman, and S. Holmberg (1998). Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection. *New England Journal of Medicine 338*(13), 853–860.

Petropoulos, C., N. Parkin, K. Limoli, Y. Lie, T. Wrin, W. Huang, H. Tian, D. Smith, G. Winslow, D. Capon, et al. (2000). A novel phenotypic drug susceptibility assay for human immunodeficiency virus type 1. *Antimicrobial Agents and Chemotherapy 44*(4), 920.

Poelwijk, F., D. Kiviet, D. Weinreich, and S. Tans (2007). Empirical fitness landscapes reveal accessible evolutionary paths. *Nature 445*(7126), 383–386.

Provine, W. (1971). *The Origins of Theoretical Population Genetics*. The University of Chicago Press.

Provine, W. (1989). *Sewall Wright and evolutionary biology*. University of Chicago Press.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rajaram, M., V. Minin, M. Suchard, and K. Dorman. Hot and Cold: Spatial Fluctuation in HIV-1 Recombination Rates. In *Bioinformatics and Bioengineering, 2007. BIBE 2007. Proceedings of the 7th IEEE International Conference on*, pp. 707–714. IEEE.

Rhee, S.-Y., T. Liu, J. Ravela, M. J. Gonzales, and R. W. Shafer (2004, Aug). Distribution of human immunodeficiency virus type 1 protease and reverse transcriptase mutation patterns in 4,183 persons undergoing genotypic resistance testing. *Antimicrobial Agents and Chemotherapy 48*(8), 3122–6.

Ripley, B. (1981). *Spatial statistics*, Volume 24. Wiley Online Library.

Ruse, M. (1991). Are Pictures Really Necessary ? The Case of Sewell Wright's "Adaptive Landscapes". *Philosoply of Science Association 2*(1990), 63–77.

Sanjuan, R., A. Moya, and S. F. Elena (2004). The contribution of epistasis to the architecture of fitness in an RNA virus. *Proceedings of the National Academy of Sciences (USA) 101*(43), 15376–15379.

Sax, P., C. Tierney, A. Collier, M. Fischl, C. Godfrey, N. Jahed, et al. (2008). ACTG 5202: shorter time to virologic failure (VF) with abacavir/lamivudine (ABC/3TC) than tenofovir/emtricitabine (TDF/FTC) as part of combination therapy in treatment-naive subjects with screening HIV RNA= 100,000 c/mL. In *Program and abstracts of the 17th International AIDS Conference*, pp. 3–8.

Schuster, P. (2012). A revival of the landscape paradigm: Large scale data harvesting provides access to fitness landscapes. *Complexity*, 1–5.

Shafer, R. W. (2006, Sep). Rationale and uses of a public HIV drug-resistance database. *J. Infect. Dis. 194 Suppl 1*, S51–8.

Shafer, R. W. and J. M. Schapiro (2008, Jan). HIV-1 drug resistance mutations: an updated framework for the second decade of HAART. *AIDS reviews 10*(2), 67–84.

Skipper, R. (2004). The heuristic role of Sewall Wright's 1932 adaptive landscape diagram. *Philos Sci 71*, 1176–1188.

Smith, T., P. Husbands, and M. O'Shea (2002). Fitness landscapes and evolvability. *Evolutionary Computation 10*(1), 1–34.

Stadler, P. (2002). Fitness landscapes. *Biological Evolution and Statistical Physics*, 183–204.

Turner, D., B. Brenner, and M. Wainberg (2003). Multiple Effects of the M184V Resistance Mutation in the Reverse Transcriptase of Human Immunodeficiency Virus Type. *Clinical and Vaccine Immunology*, 979–981.

Venables, W. and B. Ripley (2002). *Modern applied statistics with S-PLUS* (Fourth ed.). Springer, New York.

Wain-Hobson, S. (1993). Viral burden in aids. *Nature 366*(6450), 22–22.

Weinreich, D., N. Delaney, M. DePristo, and D. Hartl (2006). Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science 312*(5770), 111.

Weinreich, D., R. Watson, and L. Chao (2005). Perspective: sign epistasis and genetic costraint on evolutionary trajectories. *Evolution 59*(6), 1165–1174.

White, K., N. Margot, T. Wrin, C. Petropoulos, M. Miller, and L. Naeger (2002). Molecular mechanisms of resistance to human immunodeficiency virus type 1 with reverse transcriptase mutations K65R and K65R+ M184V and their effects on enzyme function and viral replication capacity. *Antimicrobial agents and chemotherapy 46*(11), 3437.

Wiles, J. and B. Tonkes (2006). Hyperspace geography: Visualizing fitness landscapes beyond 4d. *Artificial Life 12*(2), 211–216.

Wolf, J., E. Brodie III, and M. Wade (2000). *Epistasis and the evolutionary process*. The University of Chicago Press.

Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding, and selection in evolution. In *Proc of the 6th International Congress of Genetics*, Volume 1, pp. 356–366.

Yeh, P. J., M. J. Hegreness, A. P. Aiden, and R. Kishony (2009, Jun). Drug interactions and the evolution of antibiotic resistance. *Nat Rev Microbiol 7*(6), 460–6.

Zhuang, J., A. Jetzt, G. Sun, H. Yu, G. Klarmann, Y. Ron, B. Preston, and J. Dougherty (2002). Human immunodeficiency virus type 1 recombination: rate, fidelity, and putative hot spots. *Journal of virology 76*(22), 11273.