Intra-host HIV-1 evolution and the co-receptor switch

TECHNISCHE UNIVERSITÄT DARMSTADT

Vom Fachbereich Biologie der Technischen Universität Darmstadt zur Erlangung des akademischen Grades eines *Doctor rerum naturalium* genehmigte Dissertation von

MSc Bioinformatik Miriam Carbon-Mangels aus Zweibrücken

1. Referent: Prof. Dr. Kay Hamacher

2. Referent: Prof. Dr. Gerhard Thiel

Tag der Einreichung: 25.10.2013

Tag der mündlichen Prüfung: 18.12.2013

Darmstadt 2014

D17

Ehrenwörtliche Erklärung

Ich erkläre hiermit ehrenwörtlich, dass ich die vorliegende Arbeit entsprechend den Regeln guter wissenschaftlicher Praxis selbstständig und ohne unzulässige Hilfe Dritter angefertigt habe.

Sämtliche aus fremden Quellen direkt oder indirekt übernommene Gedanken sowie sämtliche von Anderen direkt oder indirekt übernommenen Daten, Techniken und Materialien sind als solche kenntlich gemacht. Die Arbeit wurde bisher bei keiner anderen Hochschule zu Prüfungszwecken eingereicht.

Darmstadt, den 25. Oktober 2013

_____

Abstract

The course of an infection with the human immunodeficiency virus type 1 (HIV-1) is characterised by three phases: primary infection, chronic infection and acquired immunodeficiency syndrome (AIDS). These stages are defined based on levels of the number of CD4-positive T-helper cells ($CD4^+$). This characteristic three-staged classification is also reflected in the course of the viral divergence and in the emergence of viral diversity.
It is known that the V3 loop, a region encoded in the HIV envelope gene, is important for T cell infection. The CD4 receptor of the cells is used as primary receptor for viral cell entry, and the CCR5 or CXCR4 are the most important co-receptors that are necessary for cell entry. In about half of all patients, HIV switches from CCR5 towards CXCR4 usage during the late stage of infection, which hints at the onset of AIDS. Since the co-receptor tropism is determined by the V3 loop sequence, an understanding of the mechanisms of its evolution and of the circumstances leading to the co-receptor switch is of high interest.
In the first part of the present work, we analysed longitudinal patient data, comprising information on $CD4^+$ cell count, viral load, medication, coinfections and V3 loop sequences. We examined the correlations among the clinical and evolutionary data as well as the co-receptor usage over time, guided by different questions: Is the course of disease one-directional? Can successful drug therapy influence co-receptor usage? What are the genetic differences between CCR5- and CXCR4-tropic viruses?
Due to the weak statistical support of our data, we only found few indications that successful HAART therapy influences the course of disease and the direction of the co-receptor switch. We hypothesise that successful therapy can pause or roll back the course of infection, enabling the $CD4^+$ cells to recover to high levels of immune pressure. A suppression of the viral load further can displace X4-tropic viral variants in the viral population in favour of R5-tropic variants.
In the second part of this work, we derived a fitness function to approximate the replication capacity of R5 and X4-tropic viruses. Based on a set of V3 loop sequences gathered from the Los Alamos HIV data base, the fitness function is composed of two components: the main fitness term describes the amino acid preferences found in the R5 and the X4 consensus sequence, and the additional epistatic term describes the effects of double mutations. While the impact of the main and epistatic fitness contribution can be influenced by a weighting parameter, an additional parameter controls the importance of available CCR5 and CXCR4 positive target cells. The fitness function enabled us to observe the differences of the underlying R5 and X4 fitness landscapes.
A comparison of the sequence data set showed that the R5-tropic viral sequences were highly conserved, in contrast to the X4 sequences. Network analyses confirmed the higher sequence variability of the X4 sequences, which we found to be distributed over a larger sequence space. Interestingly, our analyses revealed that the most weakly conserved sequence positions of the X4 data set were very sensible to mutations. Upon an alteration of the most weakly conserved nt positions, the X4 sequences showed an increased probability to acquire stop condons and to loose their replicative capacity.
The last part of the work describes an *in silico* approach of the V3 loop evolution based on the R5 and X4 fitness function. Simulations enable us to mimic the sequence evolution *in silico*, and to monitor the course of the viral diversity and divergence as well as the mean fitness of the simulated viral population over time.

First results indicated that our simulation is able to imitate the evolutionary course of the viral diversity and divergence of an HIV infection. In our simulations, the sequence evolution followed a chemically sensible course. Amino acids that differed from the favoured chemical properties were first replaced by amino acids belonging to the favourable chemical class and finally converged into the dominant amino acid in the specific sequence position. The present project was designed to prepare the ground for deeper insights into the evolutionary dynamics of the HIV V3 loop. Our work enabled us to gain broader knowledge of the properties of R5- and X4-tropic viral sequences.

Zusammenfassung

Eine Infektion mit dem humanen Immundefizienz-Virus (HIV) verläuft in drei charakteristischen Krankheitsphasen. Die Phasen können einerseits an Hand der Anzahl der CD4-Zellen unterschieden werden, andererseits kann eine Unterscheidung auf der Grundlage der viralen Diversität und Divergenz statt finden.
Für die Infektion der Wirtszellen durch das Virus ist die V3-Region, ein Abschnitt, der im Hüllprotein von HIV kodiert ist, von zentraler Bedeutung. Nach der Bindung von HIV an den CD4-Rezeptor der Zielzellen erfolgt eine Bindung der V3-Region an einen zellulären Korezeptor, welches in den meisten Fällen ein CCR5- bzw. CXCR4-Rezeptor ist. Im Verlauf der Infektion kann man bei etwa der Hälfte aller Patienten einen Wechsel des benutzten Korezeptors beobachten. Dieser findet im allgemeinen in einer späten Krankheitsphase statt und kündigt ein rasches Fortschreiten der Infektion an. Bisher ist es nicht gelungen, die Hintergründe und Mechanismen, welche zu diesem Korezeptor-Wechsel führen, komplett aufzuklären.
Die vorliegende Arbeit untersucht die Sequenzevolution von HIV-1 mit besonderem Augenmerk auf die Unterschiede zwischen R5-trophen und X4-trophen Viren. Der erste Teil beruht auf Daten von HIV-1-infizierten Patienten, die über mehrere Jahre beobachtet wurden. Basierend auf Publikationen aus der Zeit der beginnenden HIV-Forschung verglichen wir die Daten von akuellen Patienten mit den früheren Beobachtungen, um Unterschiede im Verlauf der Infektion zu untersuchen zwischen nahezu untherapierten Patienten und Patienten, die mit moderner Kombinationstherapie behandelt wurden. Wir fanden dabei erste Hinweise, dass die grundsätzlichen Beobachtungen der frühen Studien auch für Patienten mit modernen Therapieansätzen Bestand haben, wobei die Daten einen Unterschied im zeitlichen Verlauf der Infektion zwischen HAART-therapierten und therapie-naïven Patientengruppen andeuten. Unsere Untersuchungen lassen die Vermutung zu, dass aktuelle Therapien den Krankheitsverlauf verlangsamen und für begrenzte Zeit sogar stoppen oder zurück setzen können. Diese Hypothese konnte im Rahmen der vorliegenden Arbeit auf Grund der unzureichenden Datenlage allerdings nicht bestätigt werden.
Im zweiten Teil der Arbeit untersuchten wir die Unterschiede zwischen den R5-trophen und X4-trophen Viren an einem umfangreichen frei verfügbaren Sequenzdatensatz. Nach der Klassifizierung der Sequenzen in R5- und X4-trophe Varianten untersuchten wir zunächst die Unterschiede der R5 und X4 Konsensussequenz. Wir konnten frühere Ergebnisse bestätigen, nach denen die R5-trophen Viren stärker konserviert sind, und nach denen bei X4-trophen Viren eine Dominanz von positiv geladenen Aminosäuren in den Korezeptor bestimmenden Sequenzpositionen 11 und 25 vorliegt. Auf Basis des R5- und des X4-Datensatzes entwickelten wir zwei unabhängige Fitnessfunktionen, die die Replikationsfähigkeit der R5- beziehungsweise der X4-trophen Viren mathematisch beschreiben. Die Fitnessfunktionen bestehen jeweils aus zwei Beiträgen. Der erste Fitnessterm beschreibt die Fitness der Aminosäureabfolge der V3-Region der HIV-1 Sequenz, wohingegen der zweite Teil die Auswirkung von epistatischen Wechselwirkungen von Paaren von Sequenzmutationen auf die replikative Fitness berechnet.
Auf der Grundlage dieser Fitnessfunktionen waren wir in der Lage, die Fitnesslandschaften der R5- und X4-trophen Viren zu vergleichen. Wir stellten dabei fest, dass sich die stark konservierten Sequenzen der R5-trophen Viren in direkter Nachbarschaft im Sequenzraum befinden, während sich die Sequenzen der X4-trophen Viren über einen größeren Bereich

des Sequenzraumes erstrecken. Unsere Ergebnisse stimmten mit den Beobachtungen anderer Forschungsgruppen überein.
Die beiden Fitnessfunktionen bildeten das Herz einer sequenzbasierten Simulation der Evolution der V3-Region, die wir im dritten Teil dieser Arbeit beschreiben. Wir konnten zeigen, dass unsere Simulation die Evolution von zufälligen Sequenzen hin zur R5- bzw. zur X4-Konsensussequenz ermöglicht. Darüber hinaus folgen die Simulationen chemisch sinnvollen Pfaden. Wir konnten beobachten, dass sich anfänglich nicht optimierte, mutierte Sequenzpositionen zunächst in Richtung der korrekten chemischen Gruppe (z.B. positiv geladene Aminosäure) und in folgenden Replikationen weiter zur korrekten Konsensusaminosäure entwickelten. Unsere Simulationen ermöglichen daher Modelluntersuchungen der Evolution von artifiziellen Sequenzen der V3-Region, die nicht den Restriktionen einer groß angelegten Patientenstudie unterliegen.

Contents

List of Figures

List of Tables

1. Motivation

1.1. Motivation

In May 1983, the Science magazine published an article by Luc Montagnier and Françoise Barré-Sinoussi [9] describing a new virus isolated from the blood of a Caucasian patient with signs and symptoms of the acquired immune deficiency syndrome (AIDS). The virus, that was first termed human T-cell leukemia virus II (HTLV-II) and later named lymphadenopathy associated virus (LAV), today is known as human immunodeficiency virus (HIV). The discovery of the virus was granted with the Nobel Prize in Physiology or Medicine in 2008.
Since the discovery of HIV, the virus caused a global pandemic, with 33.4 million people worldwide living with an HIV infection [84]. Despite more than 30 years of research, HIV is still of major concern for public health. Neither a vaccine nor a curative therapy for HIV-infected patients exists and there are a number of open questions that hinder the development of therapy schemes.
The goal of the present work is to gain insight into the progression of the infection and to understand the factors that drive the dynamics of the viral evolution. We put special emphasise on the co-receptor tropism, since a change in cell tropism of the virus is observed in about 50 % of all patients [35] and is associated with a disease progression and a worse prognosis for the patients.
In our work, we combine biological approaches with *in silico* methods to seek new knowledge to support the ongoing combat against HIV.

1.1.1. Relevance of the work

According to the World Health Organisation, 2.2 million adults and 330,000 children acquired a new HIV infection in 2011, while 1.7 million people died from AIDS. At the end of the year, 34.0 million people worldwide were living with the virus [84].
HIV-infected patients are in constant need of HIV therapy to control the viral load. In highly active antiretroviral therapy (HAART), at least three drugs from different drug classes are combined to avoid or delay the occurrence of resistance mutations of the fast evolving virus.
Due to intensive research, an increasing number of anti-HIV drugs is available that enable therapy changes to cope with the occurrence of resistance mutations. The drugs belong to different classes, depending on the mode of action: nucleosidic and non- nucleosidic reverse transcriptase inhibitors (NRTI, NNRTI), protease inhibitors (PI), integrase inhibitors (II) and entry inhibitors. The drug Maraviroc [49, 67], a CCR5 co-receptor blocker and member of the drug class of entry-inhibitors, is one of the latest anti-HIV drugs.
Though the problem of resistance mutations became less severe due to the increasing

number of effective drugs, the frequent uptake of drugs often leads to a number of side effects, especially in long term therapy.
Therefore, new approaches to avoid infections have been developed, for example post-exposure prophylaxis (PEP) and pre-exposure prophylaxis (PrEP). First evidence for the success of PEP was described by Cardo *et al.* in 1997 [19], finding a 81% reduction of HIV seroconversion in health care workers after percutaneous exposure to the virus. Two years later, Guay *et al.* [66] reported a 47% reduction of PEP in mother to child transmission. In a review of Okwundu *et al.* [112], PrEP was described to reduce the incidence of HIV by 44% [63] to 62 % [148].
Despite these promising first reports, the clinical effectiveness of PEP and PrEP to avoid HIV infections is compromised. On the one hand, some trials were not able to show a protective effect in large scale use, and on the other hand the outcome of the therapy heavily depends on the compliance of the (uninfected) people taking the drugs.
Further preventive strategies, including gene therapy methods and approaches to elicit HIV-specific antibodies, so far were not effective.
In contrast, two recent publications reported success in a *functional cure* or *functional healing* of HIV-infected patients. At the 20th Conference of Retroviruses and Opportunistic Infections in March 2013, a group of researchers [116] presented a case of a newborn who was infected by mother to child transmission. The baby was treated with ART starting 30 hours after birth. Despite ART discontinuation at the age of 18 months, the plasma viral load of the child remained below the detection limit until the age of 26 month. Using ultra-sensitive methods, only a few single copies of HIV RNA could be detected. The study is still ongoing.
A few days after these exciting news, researchers from the Institut Pasteur [131] published the data of 14 HIV-infected patients with long-term virological remission after early initiated ART, so-called *post-treatment controllers*. During therapy of these patients, ART was initiated early post infection and continued for approximately three years. When the therapy was interrupted after that time, the patients were able to control the infection for at least 89 month without anti-HIV therapy. The study is still ongoing.
Despite these promising reports, a vaccine or a curative therapy for most of the 34.0 million HIV-infected people worldwide is still lacking and ongoing HIV research is of high demand.

1.1.2. Aim of the Project

The project was designed to reveal new insights into the evolution of HIV-1 within its host. An important aspect that guided our analyses is the central question about the occurrence of the co-receptor switch: Is the switch from CCR5- towards CXCR4-tropic viruses a cause or a consequence of the disease progression towards AIDS?
We examined the interrelations of clinical and evolutionary parameters from longitudinal data of HIV-infected patients. This enabled us to study more general evolutionary patterns in patients under antiretroviral therapy and to compare our findings to the one-directional evolution of mainly therapy naive patients described in the past [51, 138].
This biological view was supplemented by *in silico* studies of viral sequences of the V3 loop region of the HIV envelope protein. We derived a fitness function to describe the replicative fitness of the V3 loop and to study the fitness landscape of CCR5- and CXCR4-tropic HI

viruses. Based on the fitness function we furthermore developed a phylodynamic model to simulate the viral evolution *in silico*.

2. Introduction

2.1. Structure of the project

The present project addresses the properties of the HIV evolution and the co-receptor tropism by three associated approaches. Part one of this work is based on a clinical study observing the course of an HIV-1 infection over time. In part two, we developed a fitness function to describe the replicative fitness of the V3 loop, and to study the topology of the corresponding fitness landscape. The fitness function is the heart of the simulations that are presented in part three of this work.
In the following lines we give a short summary of the three parts of the project.

2.1.1. Project I: Correlation between clinical and evolutionary parameters of patients under HAART

In the early days of HIV research, most patients were treated with single antiretroviral drugs or did not receive any HIV specific therapy. The illustration of the clinical course of an HIV-1 infection of Pantaleo *et al.* and the description of the evolutionary course of the disease of Shankarappa date back to these early times, and the studies relied on data of untreated patients or patients with HIV mono-therapy. The early therapies did only slightly influence the course of infection and were only successful for short periods of time due to quickly upcoming drug-resistance mutations.
In recent HIV-1 therapy, antiretroviral drugs from different drug classes with different modes of action are combined into a highly active antiretroviral therapy, coined HAART. The combination therapies mainly comprise non-nucleosidic (NNRTI) and nucleosidic (NRTI) reverse transcriptase inhibitors, integrase (II) and protease (PI) inhibitors, and most recently co-receptor blockers. Due to the different modes of action and the different target sites of the drugs, combination therapy avoids drug-resistance mutations over long periods of time. Successful HAART therapy is defined as almost complete suppression of plasma viremia, with a level viral load below the limit of detection. In contrast to early therapy forms, recent HAART influences the course of the disease remarkably.
In the first part of the project, we ask whether the well-known illustrations of Pantaleo *et al.* [64] and Shankarappa [138] *et al.* introduced in the early 1990th are still suitable to describe the course of the infection of recent patients. We hypothesise that successful HAART therapy slows down or pauses the progression of the disease. In consequence, we expect the one-directional course of the disease described by Pantaleo *et al.* and Shankarappa to be turned into a bi-directional course. The success of the administered therapy could be reflected as a delay or eventually as a roll-back of the infection by stepping forward and backwards in the disease progression.
To address this question, we analysed correlations between clinical and evolutionary data of HIV-1-infected patients undergoing long-term HAART treatment. The analysis of the

longitudinal clinical and evolutionary data set is described in detail in the first part of this work.

2.1.2. Project II: Fitness function of HIV-1 V3 loop

In the second part of the project, we addressed the genetic differences between viral R5 and X4 populations and analysed deviations in the underlying fitness landscapes by utilising network theory. We selected a data set of eighty thousand HIV-1 V3 loop sequences from the Los Alamos HIV data base [100]. After data analyses and an *in silico* co-receptor prediction of the sequences, we discriminated a set of CCR5-tropic (R5) and a set of CXCR4-tropic (X4) V3 loop sequences.
Based on these two data sets, we derived a fitness function to describe the replicative fitness of the V3 loop sequence of HIV-1. We used the newly established fitness function to evaluate networks of V3 loop sequences and to analyse the fitness landscapes that are described by the fitness function.
We hypothesised that R5 sequences are favoured early in HIV infection due to a worse immune recognition, while the X4 sequences outnumber R5 strains in the late phase, as a consequence of the decreased immune pressure and due to a higher sequence variability of the X4 population.

2.1.3. Project III: Simulation of HIV-1 V3 loop evolution

In the third part of the project, we developed a software tool to simulate the course of an HIV-1 infection *in silico*. The fitness function we derived in the second part prepared the ground to estimate the replicative fitness of the simulated sequences. We examined the interplay of different biological parameters used in the simulation, e.g. the mutation rate and the population size. Furthermore, we balanced the strength of the main and epistatic fitness contributions and explored the influence of the R5 and X4 fitness term.
The simulated sequence data was used to monitor the *in silico* evolution with respect to the course of the diversity and the divergence of the viral quasispecies over time and the course of the position specific chemical properties over time.

2.1.4. Previous work within the project

The present work is part of the joint project *Monitoring of resistant HIV in newly and chronically infected HIV patients in Germany - Evolution of HIV-genotype and phenotypes during antiretroviral therapy*. The project was initiated in 1999 by Albrecht Werner at the *Paul-Ehrlich-Institut* and Hans-Reinhard Brodt at the *Universitätsklinikum Frankfurt am Main*.
When we started to work on this project in 2010, the study run for ten years and two previous research projects had been finished. In the first project, Binninger-Schinzel *et al.* [14] established a cell-based biological assay for *in vitro* co-receptor prediction (see the following Section 2.1.4).
The second research project addressed the peculiarities of HIV evolution of a patient who discontinued HAART to underwent stem cell therapy. The respective case studies published by Wolf *et al.* [162] and by Kamp *et al.* [85] built the basis for the correlation analyses in Chapter 3.

Biological assay for *in vitro* co-receptor prediction

The first *in silico* co-receptor prediction tools lacked sensitivity to detect X4 variants, a fact that became critical with the marketing authorization of Maravirioc [49, 67], the first CCR5 co-receptor blocker. The problem of a decreased X4 sensitivity is even harder to tackle in patients with a viral load level at or below the limit of detection, since the isolation of viral RNA sequences from the blood of those patients is often impossible.
Binninger-Schinzel *et al.* [14] addressed this problem by the development of a highly sensitive assay for *in vitro* co-receptor determination. They isolated peripheral blood mono-nuclear cells (PBMCs) from a homozygous CCR5 negative healthy donor. By a co-cultivation of the donor PBMCs with gamma-irradiated human leukaemia T-cells, they established an immortal CD4-positive cell line that was CCR5 receptor negative and hence non-permissive for R5-tropic strains. Due to this property, the cell line was named *isnoR5*. Infection of the isnoR5 cells with a dilution series of virus-containing supernatant proofed that the assay is highly sensitive for the detection of low amounts of X4 viruses in a viral population. As a marker to quantify the viral replication, a p24 antigen assay of the content of viral Gag protein was used.
Werner *et al.* [159] could show successful breeding of virus on the isnoR5 cell line in 87% of the experiments, independent of the therapy or the CDC state of the patients, which they stated to be a high rate in comparison to other labs.
In the course of this project, the isnoR5 cell line was used for the *in vitro* co-receptor determination and for the validation of the *in silico* co-receptor predictions generated by the bioinformatics tools geno2pheno [97] and FSSM [82, 81, 118], which are described in Section 3.3.3.

Case study of stem cell treated patient

Wolf *et al.* [162] described the course of the disease of an HIV-1-infected patient that underwent allogeneic stem cell transplantation due to severe aplastic anemia. After radiation and deletion of his own bone marrow cells, the patient got blood stem cells of a healthy donor. In contrast to the therapy of Thimothy Brown, better known as *Berliner Patient* [80, 3], who was treated with blood stem cells from a CCR5Δ32 donor, the blood stem cells for this patient originate from a donor without CCR5 deficiency.
The clinical perspective of this case study was complemented by Kamp *et al.* [85], who analysed the evolution of the viral V3 loop sequences of this patient using phylogenetic methods. A detailed analysis of the viral sequences of the patient, taken at six subsequent time points, showed that the stem cell therapy led to a decreased viral diversity, though HAART was disrupted during the period of the stem cell transplantation. Radiation and stem cell transplantation suppressed the patients immune system and simultaneously slowed down the course of the viral evolution.

3. Correlations between clinical and evolutionary parameters of patients under HAART

3.1. Introduction

In the first section of this chapter, we introduce some relevant details of the human immune system and describe the peculiarities of the human immunodeficiency virus type 1 (HIV-1), with special emphasise on its influence on the human immune system. The knowledge about the special properties of the virus and its interplay with the CD4-positive ($CD4^+$) immune cells is crucial to understand the relevance of the subsequent analyses. After the introduction of the biological background, we define some important biological and mathematical terms and explain the methods we used for the data analysis.
Central to this chapter is the description and the analysis of the longitudinal patient data set established in the course of this study. We examine the course of the disease of HIV-1-infected patients under HAART, with special emphasis on the clinical and evolutionary parameters of the infection.
We compare our findings with the well-known descriptions of the course of the disease of Pantaleo *et al.* [64] and Shankarappa *et al.* [138], which were established in the early 1990th.

3.1.1. HIV-1 infection and the human immune system

In the following lines, we give an overview over the role of $CD4^+$ in the immune system and the consequences of an HIV-1 infection. A more detailed description can be found in the article of Weber [157] or the book review of Levy [98], which are the main sources for the presented summary.

The human immune system

The human immune system is a complex interplay of innate and adaptive mechanisms, consisting of physical barriers and cell mediated responses, that protect us against diseases. An infection with the HI virus has large impact on the immune defence, since it alters the cell balance and disturbs the cell interactions of the human immune system.
The central aspect of an HIV-1 infection is a steady decline of immune cells bearing the CD4 receptor on the surface ($CD4^+$ cells) [98, 157]. While healthy young adults have $\sim 2 \cdot 10^{11}$ mature $CD4^+$ cells, the $CD4^+$ cell count of HIV-infected patients in a progressed state of the disease often drops below 200 cells per microlitre (which meets

approximately a bisection of the amount of cells) [68]. This threshold is commonly known as the AIDS-defining threshold [21, 160].

Innate immune system The first layer of immune protection is the innate immune system. The innate defence consists of a combination of physical barriers and cell-mediated immunity. The mucosa in the genital tract, populated by $CD4^+$ dendritic cells (DC), is one of these barriers.
DCs orchestrate the immune system via a signalling to T cells and B cells and stimulate resting T cells. Furthermore, DCs are important antigen presenting cells (APC) - they present foreign proteins to T cells via the major histocompatibility complex (MHC) on their surface. Efficient antigen-presentation is essential for an effective innate and adaptive immune response, since T cells can only recognise antigens that are presented on MHC molecules of APCs (e.g. DCs and macrophages).
The DCs in the mucosa of the genital tract are often the first cells that are infected by the HI virions. These early infected cells can transport the virus to the lymphoid tissue and enable the infection of other immune cells [98, 99].

Adaptive immune system The major component of the adaptive immune system are lymphocytes [98], a variant of white blood cells. The lymphocytes can be further discriminated into the thymus-derived T cells, the bursa-derived B cells, and the natural killer (NK) cells. While B cells are important for the humoral immunity, T cells are indispensable for the cell-mediated immunity. In contrast to B cells and T cells, NK cells are a part of the innate immune system.
The group of T cells can be further divided into precursor, effector, helper, and suppressor T cells, based on their specialised function within the immune system. T helper cells direct and regulate both humoral and cell-mediated immune responses via interaction with precursor and suppressor T cells, B cells, and monocytes. They orchestrate the immune defence by the secretion of cytokines, stimulate the antibody production of B cells, direct phagocytes, and activate other immune cells. Though T helper cells have no direct cytotoxic activity, they bear a central role in the human immune system.

3.1.2. The influence of the HIV-1 infection on the immune system

The $CD4^+$ T cells and the macrophages are the major target of the HI virions. NK cells can also be infected by HIV-1, since $\sim$ 50% express CD4, and a lower percentage in addition bear CCR5 and CXCR4 receptors [98, 99].
In the course of an HIV-1 infection, the $CD4^+$ cells are reduced in function, prior to an observed reduction in number [27]. The reasons for the reduced number of $CD4^+$ cells are manifold and comprise an increased cell death (e.g. due to apoptosis, necrosis, and bystander effects from the formation of syncytia (i.e. multi-nucleated cells) that induce the death of uninfected $CD4^+$ cells), and a decreased proliferation, life span, and regeneration (due to the cytokine release from the infected cells) [89, 98].
Furthermore, the rate of the remaining immune cells is altered [98, 157]. While the number of the long-term memory T cells is decreased, the number of short-lived naive effector T

cells is increased. This leads to a shift in the rate of $CD4^+$ to $CD8^+$ cells.
In advanced stages of the disease, different forms of anemia, leukopenia, and thrombocytopenia are observed in the HIV-1-infected patients (e.g. due to a bone marrow dysfunction), but it is still not understood in detail how HIV-1 disorders the immune system.
Finally, the HIV-1-infected patients develop immune deficiencies, which lead to the acquired immunodeficiency syndrome (AIDS) in the late stage of disease. Due to this property of the virus, an international virus taxonomy consortium coined the name *human immunodeficiency virus* (HIV) [28].

Characteristics of HIV-1

HIV is a positive single-stranded RNA (ribonucleic acid) virus of the genus of lentiviridae belonging to the retrovirus family. The complete HIV-1 genome has a length of 9,181 basepairs (bp) [58] and consists mainly of the genes gag, pol, and env, flanked by two long terminal repeats (LTR) of about 600 nucleotides (nt) with a 5′-cap and a 3′ poly-A tail. Figure 3.1 illustrates the sequence of genes of the HIV genome.
The Gag polyprotein is cleaved into three proteins: matrix, capsid, and nucleocapsid. The cleavage of the Pol polyprotein results in additional viral reverse transcriptase (RT), protease, and integrase molecules. The envelope gene (env) of HIV has a length of 2,571 bp. The gene product of env is a precursor glycoprotein (gp) 160, which is cleaved into the membrane proteins gp41 and gp120 as well as the HIV accessory proteins Vif, Vpu, Vpr, and Nef, and the regulatory proteins Rev and Tat. Central to the present project is a specific 105 bp nucleotide region of gp120, termed *V3 loop* region.

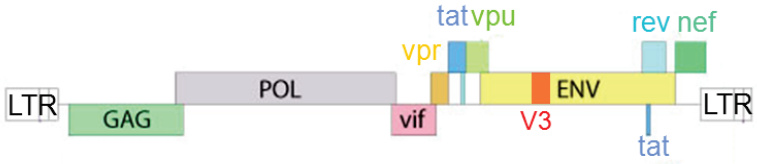
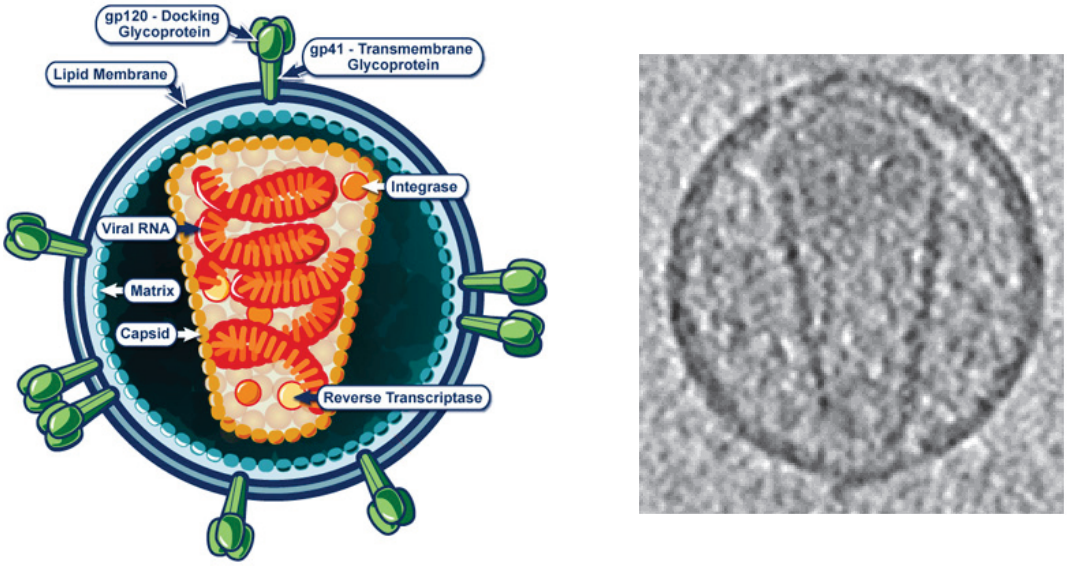


Figure 3.1.: **Complete HIV genome**
The figure illustrates the sequence of genes of the HIV-1 genome and the location of the V3 loop within the env gene.
(image source: image adapted from [5])

Upon the creation of virions, two (usually identical) strands of the full-length HIV RNA, together with the proteins RT, protease, and integrase, and the accessory proteins Nef, Vif, and Vpr are packed together into the cone-shaped viral core formed by the p24 Gag capsid protein. The viral core is coated by a lipid membrane, carrying 10 to 15 protein spikes. Each spike consists of a heterodimeric trimer of the external surface glycoprotein (gp) gp120 and the transmembrane protein gp41. These membrane-protruding spikes play an important role in the cell entry of HIV (see Section 3.1.2).
Mature HI virions are roughly spherical with a diameter of $100 - 120$ nm. Figure 3.2 illustrates the organisation of the viral components of a mature particle.



(a) (b)

Figure 3.2.: **Illustration of a mature HI virion**
a) A schematic sketch of a mature HI virion (`www.niaid.nih.gov/SiteCollectionImages/topics/hivaids`).
b) Electron microscopic image of a mature HI virion (Yu *et al.* [166]).

The HIV-1 infection cycle

The HIV-1 infection cycle can be distinguished into five coarse-grained phases that are depicted in Figure 3.3. In the following description, we will focus on some key aspects of HIV biology [55] which are relevant for this work.
An HIV-1 infection is initiated by the process of receptor binding and cell entry (step 1 in Figure 3.3). After the release of the viral core into the host cell, the viral proteins and the RNA are distributed within the cell. The RT transcribes the viral RNA into cDNA (step 2), which becomes integrated into the host cell DNA by the integrase (step 3). Initiated by some start signals, the replication of the integrated pro-viral DNA starts (step 4). The viral transcripts are translated and processed into viral proteins. They assemble with two full-length viral RNA molecules and form immature virions, which maturate upon budding from the infected cell (step 5). The mature virus particles start a new infection cycle.
In the following paragraphs, the main aspects of these five phases are described. If no other references are given, the content of the following paragraphs is based on the work of Frankel *et al.* [55].

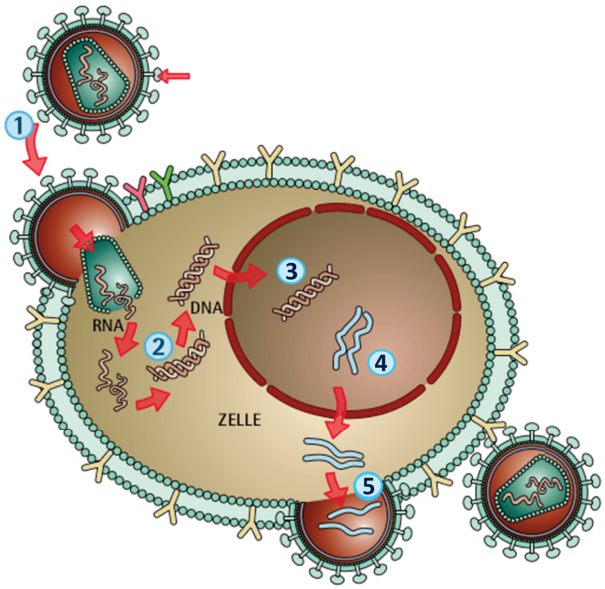


Figure 3.3.: **HIV infection cycle**
The image depicts the sequence of steps of the viral infection cycle. (image adapted from [128])

1. Receptor binding and cell entry
2. Reverse transcription
3. DNA integration
4. Replication
5. Translation, assembly and budding

Step 1: Receptor binding, membrane fusion and cell entry In 2000, Doms and Trono [37] described and illustrated the process of receptor binding and cell entry (see Figure 3.4). The surface protein gp120 and the transmembrane protein gp41 are the most important HIV proteins that are involved into the virus-to-cell contact. Heterotrimers of both proteins are non-covalently associated to form membrane protruding spikes. These spikes are anchored in the lipid membrane of the virion by the gp41 domain, and the tip of the spikes is formed by the gp120 moieties.
When the viral and the cellular membrane proteins are in close proximity, the gp120 region at the tip of the spikes binds to the primary CD4 receptor on the host cell surface. This interaction is a necessary prerequisite for viral cell entry since it initiates the first structural rearrangements of the gp120 moiety and the CD4 receptor. The conformational changes lead to an exposure of the third variable loop (V3 loop), a part of the gp120 subunit, to the host cell [37, 98].
Revealed at the tip of the spike, the V3 loop structure is a high affinity binding site that interacts with a chemokine co-receptor on the target host cell surface. Different chemokine receptors from a seven-transmembrane G-protein-coupled receptor family are known to serve as co-receptors for HIV binding and cell entry, for example the receptors CCR1 to CCR5, CCR8, CXCR4, BOB, or Bonzo [16, 44]. Of those, the chemokine receptors CCR5 and CXCR4 (prior known as fusin) are the most important co-receptors to facilitate HIV cell entry [36, 38, 167].
The co-receptor tropism has been shown to influence significantly the disease progression

[29, 62, 125]. More details on the co-receptor usage are given in Section 3.1.2.
Upon binding of the V3 loop to the cellular co-receptor, further structural rearrangements of the spike are induced. These conformational changes occur predominantly in the gp41 moiety and lead to the formation of a hairpin structure which is exposed to the host cell membrane. This hairpin structure of gp41 triggers the fusion of the membranes of the virus particle and the host cell (possibly including an interaction with another specific host cell receptor) [37, 98]. As a result of the fusion process, the viral nucleocapsid is released into the target cell. In the cytoplasm, the capsid is uncoated and the complex of the viral RNA and the viral proteins is transported into the cell nucleus.

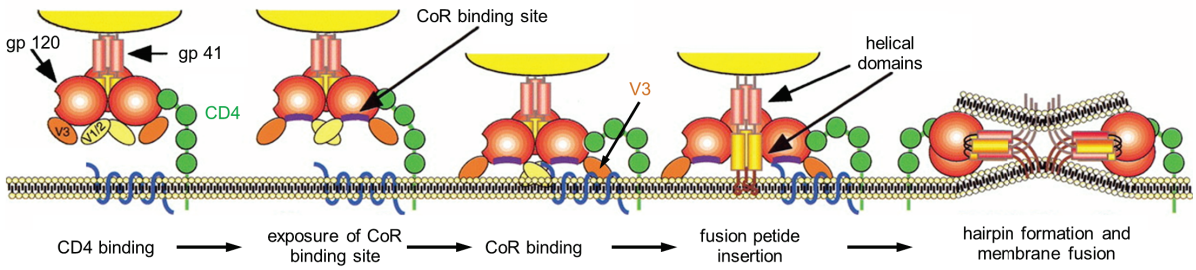


Figure 3.4.: **Binding and cell entry of HIV**
The image describes the sequence of steps of HIV cell entry, from CD4 binding to membrane fusion and the release of the viral nucleocapsid into the host cell.
(image adapted from Doms *et al.* [37])

Step 2: Reverse transcriptase Once the complex of the viral RNA and protein molecules reaches the nucleus, the viral reverse transcriptase (RT) molecules start to transcribe the viral RNA into proviral DNA. The transcription needs specific tRNA primers for initiation. Usually, one Lysin tRNA is packed into the viral capsid and this molecule is used to start the transcription on one of the two RNA strands. During the formation of the first minus-strand of HIV DNA, the viral plus-strand RNA template is degraded. After completion of the minus-strand DNA copy, the RT jumps onto this newly produced DNA strand and uses it as template to transcribe a complementary plus-strand. The two newly synthesised DNA strands hybridise into a double-stranded viral copy DNA (cDNA). It is important to note that the HIV RT lacks an editing function and thus the transcription of the viral RNA into DNA is highly error prone. Estimates of the error rate vary in the range of $5.0 \cdot 10^{-4}$ [126] to $3.4 \cdot 10^{-5}$ [102]. Typically, one error in every 1,000 bp is assumed. Applied to an HIV genome of ∼10,000 bp, this results in ten nucleotide changes in every round of reverse transcription.
This high mutation rate of the HIV RT is the main obstacle to HIV therapy as well as drug and vaccine development.

Step 3: Integration into host genome After reverse transcription, the viral integrase processes the newly formed viral cDNA. By detaching two nucleotides to both sides of the blunt, double-stranded ends, it creates so-called *sticky ends*, single-stranded DNA extensions at both ends of the viral DNA genome. Such prepared, the viral cDNA is integrated into the host genome, being site specific only with respect to the sticky end extensions.

Upon cleavage of the host DNA, the integrase joins the sticky ends of the cDNA to the host cell DNA and ligates them. Upon DNA repair after ligation, host cell enzymes create short repeats at the flanking ends of the integrated cDNA.

Step 4: Replication After integration into the host cell genome, the hosts RNA polymerase II transcribes the provirus and produces HIV RNA and mRNA whenever the DNA of the infected host cells is transcribed. The transcription is regulated by the binding of host factors and viral proteins to the long terminal repeat. The viral proteins Tat and Rev are important for transcription control. While the Tat protein stimulates transcription and facilitates elongation of the transcript, Rev controls the splicing of the transcript and mediates the transport of the fully and partially spliced messenger RNA (mRNA) from the nucleus into the cytoplasm [86]. In the cytoplasm, some of the mRNAs are directly translated into HIV proteins or into chains of multiple protein precursors, which are further processed by the viral protease. Other mRNAs are delivered to different cell locations for translation and processing − the env mRNA for example is translated at the endoplasmatic reticulum.
In contrast to the productive state of activated infected host cells, infected cells can also go to resting state. Then the infection becomes latent and non-productive until the cell is re-activated again [142].

Step 5: Viral assembly and budding After the translation and the processing of the viral proteins, the viral components assemble to form new virions. The structural proteins derived from the gag gene form the new viral cores. Each viral core comprises a complex of two full-length RNA transcripts, some host tRNAs, the viral proteins RT, protease, and integrase, and the accessory proteins Nef, Vif, and Vpr.
Upon budding of the new viral particles from the host cell, the cores are coated with a lipid bilayer taken off the cell membrane. This newly formed lipid envelope becomes spiked with complexes of the viral Env proteins gp41 and gp120. Shortly after budding, the maturation of the new viral particles is completed and the released virions can infect new host cells and restart the transcription cycle.

Co-receptor usage of HIV-1

The binding of the V3 loop to the cellular co-receptor is an important step in cell entry. The co-receptor usage is highly specific and mainly depends on the protein sequence of the V3 loop [24]. Visualisations of the loop structure are given in part two of the work (Figure 4.13).
The amino acid sequence of the V3 loop determines the cell tropism and thereby influences the progression of the disease [29, 62, 125].
As stated earlier, the chemokine receptors CCR5 and CXCR4 are the most abundant co-receptors for cell entry [36, 38]. While the CCR5 receptor is predominately used in early HIV infection, about 50 % of all patients face a co-receptor switch from CCR5 towards CXCR4 in later stages of the disease [35]. The co-receptor switch is associated with disease progression and with a worse prognosis for the patient. CXCR4-tropic strains show an increased growth rate *in vitro* and induce the formation of multi-nucleated cells (syncytia), which have a cytopathic effect on uninfected cells (bystander effect).

The occurrence of CXCR4-tropic strains often marks the manifestation of AIDS related symptoms [29, 135, 138].
Amino acid changes due to non-synonymous mutations in the nucleotide sequence of the V3 loop region are the reason for the switch of the co-receptor tropism from CCR5 towards CXCR4. Xiao *et al.* [165] explored these sequence changes and found conserved uncharged amino acids at position 11 of the V3 loop, mostly serine and glycine, and the negatively charged amino acids glutamic acid and aspartic acid at position 25 of CCR5-tropic strains. Mutational studies showed that a substitution with the positive amino acids arginine or glutamine at both positions altered the co-receptor usage in favour of the CXCR4 receptor. From this observation they derived the amino acid consensus motif `S/GXXXGPGXXXXXXXE/D` for positions 11 to 25 of the V3 loop as a determinant of CCR5 tropism. The notation `S/G` and `E/D` indicates the alternatives of the dominant amino acids at positions 11 and 25 of CCR5-tropic sequences.
Based on these observations, the first sequence derived decision rule to determine the co-receptor phenotype of a V3 loop sequence was established, the so-called 11/25 rule.
The 11/25 rule still is an important determinant of recent *in silico* co-receptor prediction tools. In Section 3.3.3 we describe two co-receptor prediction tools that are used as valuable instruments in scientific research and in therapy optimisation.
The reasons that lead to the co-receptor switch in about 50% of all patients are still unclear, but a number of hypotheses tries to explain the change in cell tropism. The most prominent hypotheses were discussed by Regoes and Bonnhoefer in 2005 [123]. A short description is given in the following paragraphs.

Transmission mutation hypothesis According to Regoes and Bonnhoefer [123], the transmission mutation hypothesis relies on the assumption that CCR5-tropic strains are favoured upon virus transmission. In consequence, CCR5 using viruses are found more often in early stages of disease. This hypothesis further states that the co-receptor switch happens by chance at some time during infection as a result of the high mutation rate of the reverse transcriptase of HIV.

Target cell life time hypothesis Rodrigo [127] attributes the co-receptor switch to a competition of CXCR4- and CCR5-tropic strains. He argues that the more cytopathic CXCR4 viruses are outcompeted since they destroy their replication reservoirs faster than the CCR5-tropic viruses. This disadvantage leads to a shorter life time of CXCR4-infected host cells and thus CCR5-tropic strains persist longer and dominate the infection.

Target cell based hypothesis The target cell based hypothesis formulated by Davenport *et al.* [34] and modelled by Ribeiro *et al.* [124] argues that the co-receptor switch is a consequence of the different rates of available CCR5 and CXCR4 positive cells within the host.
Following this hypothesis, the number of activated

CCR5 positive memory T cells is increased early during the infection as a result of the hosts immune defence. The availability of activated CCR5 cells favours CCR5-tropic viral strains early in the infection. In the late stage of the disease, when the memory T cells are mainly depleted, the rate of the naïve CXCR4 positive T cells increases, and thus the CXCR4-tropic strains dominate in the late stage of the infection.

Immune-control hypothesis

The immune-control hypothesis of Pastore *et al.* [113] explains the predominance of CCR5-tropic viruses in early stages of the disease with a reduced replicative fitness of CXCR4-tropic strains on the one hand and a better immune recognition and control of CXCR4 strains by the human immune system on the other hand. They state that CXCR4-tropic strains are successfully fought by the immune system and are not able to replicate efficiently early in infection, when the number of immune cells is high and the immune system is strong.
With advanced T cell depletion due to continuous infection of T cells by HIV, the immune system is weakened and the immune pressure fades. In the late stage of disease, the exhausted immune system is no longer able to effectively fight HIV and under this condition the CXCR4 strains are able to sustain and become prevalent.

So far, none of the hypotheses could be confirmed or disproved and an explanation for the co-receptor switch is still missing. Since the occurrence of CXCR4-tropic strains marks an advanced stage of the infection and an accelerated disease progression, an understanding of the mechanisms of the co-receptor switch is essential to prevent the disease progression and the development of AIDS.
The knowledge is also necessary in the field of rational drug design to develop new drugs and to avoid unwanted side effects, for example regarding the new drug class of co-receptor blockers, with Maraviroc as the first licensed drug [49, 67]. Since Maraviroc impedes the usage of the CCR5 receptor, it was suspected to accelerate the co-receptor switch by establishing an additional pressure to HIV towards CXCR4 co-receptor usage.
These aspects indicate the relevance of the co-receptor usage and pronounce the importance of an understanding of the mechanisms of HIV evolution.

3.1.3. Phases of an untreated HIV-1 infection

The course of an untreated HIV-1 infection can be separated into three phases, either based on clinical observations or on evolutionary sequence parameters. In both classification systems, the co-receptor switch occurs in an progressed stage of infection.

In this section we describe the three phases of the disease and the different classification methods.

Clinical classification

At the Clinical Staff Conference held on the 27th of June 1990, Anthony S. Fauci and Giuseppe Pantaleo *et al.* [51] for the first time showed an illustration summarising the typical clinical course of an HIV-1 infection. Fauci and Pantaleo *et al.* refined the image in their subsequent publications [64, 50]. This image, shown in Figure 3.5, describes the general course of the $CD4^+$ cells and of the viral load (VL) during the three phases of the disease in untreated patients.
During the initial phase of infection (phase 1), the viral load rapidly increases, leading to an acute retroviral syndrome three to six weeks after primary infection [64]. During this phase, the symptoms are described like the symptoms of an acute seasonal influenza infection. Clinically, a pronounced drop of the number of $CD4^+$ cells to about 500 cells per microlitre of plasma is described.
After one week to three months of high level viremia, the immune system is capable to effectively chase the virus [64]. The viral load remarkably decreases and stabilises at a patient specific virus level known as the *viral set point*. The viral set point is an individual amount of virus that can be controlled by the hosts immune system. This level of viral load is quite stable throughout the chronic phase of the HIV infection (phase 2). Though the immune system is able to control the level of the virus during the chronic phase, the number of $CD4^+$ cells decreases slowly but permanently due to a continuous infection of the activated cells by the replicating virus.
In most patients, the number of $CD4^+$ cells drops below 200 cells per microlitre of plasma during the infection (phase 3). The time span to this clinical observation varies between patients and averages to about ten years [51]. At that time, the immune system of the patients is exhausted and the $CD4^+$ cells are widely depleted. Clinically, a dramatic weight loss is described for the majority of the patients and they suffer from various opportunistic infections. Characteristic examples are respiratory tract infections, herpes, hepatitis, and cancer. This third stage of disease is known as acquired immunodeficiency syndrome (AIDS). The patients die due to multiple opportunistic infections which the immune system no longer can control.

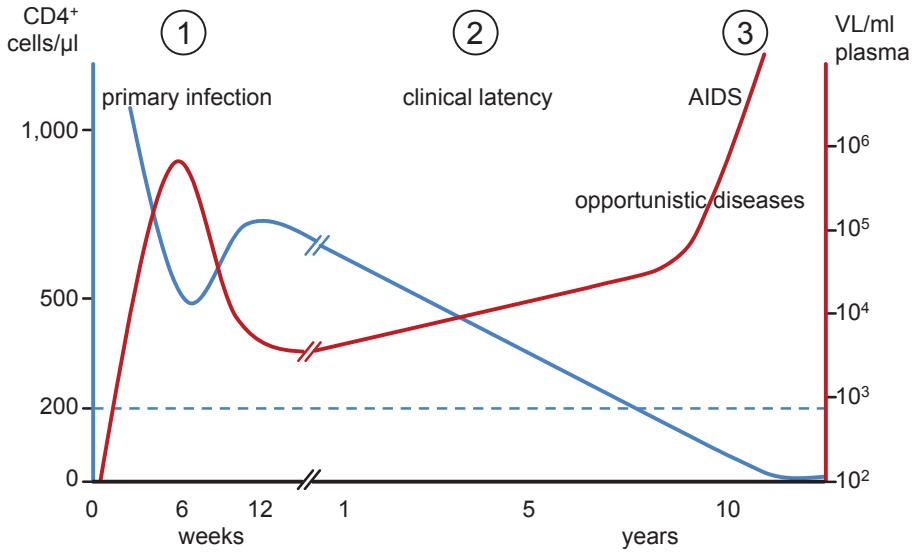


Figure 3.5.: **Clinical phases of an untreated HIV-1 infection**
The illustration describes the clinical course of an HIV-1 infection in a patient without therapy. The blue curve marks the number of $CD4^+$ cells and the red curve describes the course of the viral load.
(image based on Fauci and Pantaleo *et al.* [51, 50, 64])

In 1986, the Centers for Disease Control and Prevention (CDC) [21], Atlanta, USA introduced a classification system for HIV-1 infections. The classification is based on a laboratory category, which is the lowest documented $CD4^+$ cell count of the patient, called *nadir*, and a clinical category based on the presence or absence of specific HIV-related conditions. This staging system is still used in hospital. The most recent version was revised by the CDC in 2008 [133] and is presented in Table 3.1

Table 3.1.: **Classification of the HIV-1 infection stage by the CDC** [133]

$CD4^+$ cell count categories	Clinical categories		
	category A: Asymptomatic, acute HIV, or persistent generalised lymphadenopathy	category B: Symptomatic conditions, not A or C	category C: AIDS-indicator conditions
stage 1: $\geq$ 500 cells/μL	A1	B1	C1
stage 2: 200 - 499 cells/μL	A2	B2	C2
stage 3: $<$ 200 cells/μL	A3	B3	C3

Besides the nadir of the patient, which defines the coarse-grained numerical classification of the disease stage, the CDC provides a list [133] of symptomatic conditions and AIDS-indicator conditions to assist the physicians to determine the disease stage.
In the case of a missing diagnosis, the classification system is supplemented by *stage unknown* (e.g. Ax) and *category unknown* (e.g. x2).

Evolutionary classification

In 1999, nine years after Fauci and Pantaleo *et al.* [51, 50, 64] presented their illustration of the clinical course of the disease, Shankarappa *et al.* [138] described the course of an untreated HIV-1 infection based on evolutionary parameters. They originally discriminated five different phases of the infection, which can be consolidated into three phases comparable to the phases observed by Fauci and Pantaleo *et al.* (compare Figure 3.6).
They found that during the acute initial phase of the disease the genetic diversity within the viral population and the evolutionary distance (termed divergence) of the viral population to the founder strain steadily increase. Furthermore, the transition from acute to chronic phase is marked by the emergence of CXCR4-tropic strains.
During the following chronic stage of disease, the error-prone reverse transcriptase still generates escape mutants and the viral population further diverges from the founder strain. In parallel, the level of diversity stabilises, since the diminishing number of $CD4^+$ cells results in a decreasing immune pressure and an increasing amount of the virus.
In patients with a co-receptor switch, a peak of the CXCR4-tropic viral population marks the breakdown of the immune system and the begin of the last stage of the disease. In that phase, also the viral divergence stabilises. The immune system is exhausted and the depleted population of immune cells is no longer able to fight the virus. High amounts of virus particles are produced.
During the present work we recognised that the description by Fauci and Pantaleo *et al.* [51, 50, 64] mainly coincides with the findings by Shankarappa *et al.* [138]. Both groups distinguished three phases of the disease in the course of an HIV-1 infection. While the older illustration of Fauci and Pantaleo *et al.* focussed on the clinical parameters that were determined routinely by physicians during the regular visits of the patients, Shankarappa *et al.* described the evolutionary measures that were available as a result of the beginning sequencing approaches in later days of biological research.
For the evaluation of the clinical and evolutionary patient data presented in this work, we have to keep in mind that both the publications by Fauci and Pantaleo *et al.* [51, 50, 64] from the 1990th and the publication by Shankarappa *et al.* [138] from 1999 are based on data from the early days of HIV research. The participating patients were therapy naïve or treated only by a single reverse transcriptase inhibitor. In contrast, the present study analysed data from HAART treated patients.

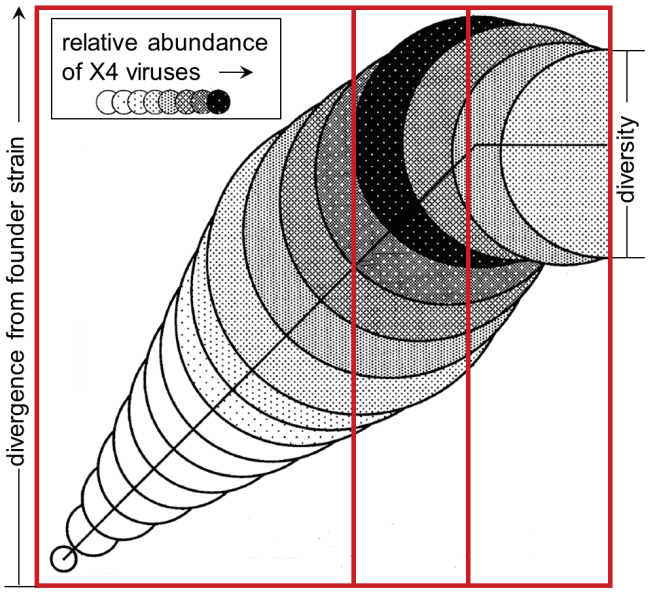


Figure 3.6.: **Phases of an untreated HIV-1 infection** The illustration describes the course of an untreated HIV-1 infection over time (x-axis), based on evolutionary parameters. The divergence of the population from the founder strain is described by the hight of the curve (y-axis), whereas the diversity of the sequences within a sample is described by the size of the circles. (image adapted from Shankarappa *et al.* [138])

3.2. Methods

In this section we introduce necessary terms and definitions as well as mathematical measures and software that were used to describe and analyse the evolutionary interrelations.

3.2.1. Evolutionary and mathematical definitions

Phylogenetics and phylogeny

The study of evolutionary relationships of populations is termed *phylogenetics*. Based on sequencing data and by the computation of multiple sequence alignments (MSA), evolutionary relations are reconstructed. A phylogenetic study yields a hypothesis about the evolutionary history of the population, which is in general described by a genealogic tree, termed *phylogeny* of the population [20].

Coalescence and coalescence time

The term *coalescence* describes the merging of genetic lineages backwards in time towards the most recent common ancestor (MRCA). In this context, the *coalescence time* is the predicted amount of time that passed between the introduction of a mutation and the observation of a particular distribution of the mutation in a population [91].

Effective population size N_e

Following Sewall Wright [164, 163], the *effective population size* N_e is "*the number of breeding individuals in an idealised population that would show the same amount of dispersion of allele frequencies under random genetic drift or the same amount of inbreeding as the population under consideration*". The concept is used to determine the rate of the evolutionary change that results from the effect of sampling in a finite population [23]. N_e is estimated empirically with respect to the coalescence time.
In real populations, N_e is neither constant nor undergoes a regular (e.g. linear or exponential) change [155]. Therefore models developed to estimate the N_e of real populations based on predictions from artificial evolutionary models assuming a regular change have to be used with care.

Bayesian skyline

The *Bayesian skyline* [42] estimates the effective population size N_e over time. It describes a possible course of the number of individuals over time that would result in the observed phylogeny, reflecting the allele frequencies over time. The details of the Bayesian skyline reconstruction we used in the present work are described in the recent BEAST publications [40, 42, 43].

Hamming distance

The Hamming distance is a distance measure for strings that was introduced by Hamming [71, 72]. For two sequences of symbols of equal length, the Hamming distance is defined

as the number of differing sequence positions.
In biological terms, the Hamming distance is the least number of point mutations that are necessary to transform one nucleotide sequence into another. For the analysis of amino acid sequences, it is defined as the number of differing amino acids between two sequences. For two sequences $X = x_1, ..., x_n$ and $Y = y_1, ..., y_n$ of equal length n, the Hamming distance $H(X, Y)$ is the sum of the different sequence positions i, e.g. for two identical positions $x_i == y_i$, $h_i(x_i, y_i) = 0$, and for two differing positions $x_i \neq y_i$, $h_i(x_i, y_i) = 1$:

$$H(X, Y) = \sum_{i=1}^{n} h_i(x_i, y_i) \tag{3.1}$$

Example:
Sequence X = ATGCATGC
Sequence Y = ATGGATCC

$h_1(x_1, y_1) = h_2(x_2, y_2) = h_3(x_3, y_3) = h_5(x_5, y_5) = h_6(x_6, y_6) = h_8(x_8, y_8) = 0$
$h_4(x_4, y_4) = h_7(x_7, y_7) = 1$
$H(X, Y) = \sum_{i=1}^{n} h_i(x_i, y_i) = 2$

Diversity and Divergence

Diversity and divergence are two measures that describe the genetic distance within or between samples of sequences. According to Shankarappa *et al.* [138], the genetic diversity within a group of sequences can be estimated as the mean and the standard deviation of the pairwise nucleotide distances between all pairs of sequences from that group.
The divergence determines the evolutionary distance between two groups of sequences at different time points. In terms of viral infection, the viral sequences from the earlier time point are supposed to be the founder sequence or the founder population (i.e. the population of the first virus-positive sample). The second sample is a more recent sample obtained at a later time point. To determine the viral divergence, all pairwise Hamming distances between the founder sequence and every sequence of the later sample are calculated. The divergence of the later sample from the founder strain is then calculated as the mean and standard deviation of all pairwise distances.
In biological samples, a single founder sequence can often not be determined. In the case of a founder population of differing sequences, Shankarappa *et al.* [138] either defined one random sequence from the first virus-positive sample as the founder sequence or approximated the founder sequence by a consensus sequence of all sequences of the founder population.
In the present work, we adapted the definitions of Shankarappa *et al.* [138] to fit our needs. We computed the diversity of any pair of sequences as the pairwise nucleotide distance, i.e. the Hamming distance, of this pair of sequences. We defined the diversity of a sample S of m sequences as the average of the Hamming distance of all $m(m-1)$ pairs of sequences

(X_j, X_k), $(k \neq j)$ of that sample.

$$diversity(S) = \frac{1}{m(m-1)} \sum_{\substack{j=1 \\ k \neq j}}^{m} \sum_{k=1}^{m} H(X_j, X_k) \tag{3.2}$$

with $H(X_j, X_k)$ being the Hamming distance of sequence X_j and sequence X_k.
The divergence estimates the genetic distance of a sample of sequences taken at time t $(t > 0)$ to the viral founder strain (at time t_0). Since many of the patients of our study were in a progressed state of the disease, we were not able to determine the founder strain of the viral population or to estimate it as Shankarappa *et al.* [138] did. Thus, we adapted the definition of the divergence. We defined the divergence of the first HIV-positive blood sample E of each patient, consisting of l sequences Y_o, to be zero. The divergence of any more recent sample S of m sequences X_j is defined as the average Hamming distance of all ml pairs of a sequence X_j of S to a sequence Y_o of E of that patient.

$$\begin{aligned} divergence(E) &= 0 \\ divergence(S) &= \frac{1}{ml} \sum_{j=1}^{m} \sum_{o=1}^{l} H(X_j, Y_o) \end{aligned} \tag{3.3}$$

with $H(X_j, Y_o)$ being the Hamming distance of sequence X_j and sequence Y_o.

Bayes factor

The *Bayes factor* (BF) [87, 110, 61] is a method for hypothesis testing. The test is a modification of the classical likelihood ratio test described by Neyman and Pearson [111] and can be used to decide between two alternative models M_1 (e.g. the null model) and M_2 (an alternative model) to describe some observed data X. To decide whether model M_1 with parameters θ_1 or model M_2 with parameters θ_2 fits the data X best, the BF [40] is calculated:

$$BF = \frac{p(X|M_1)}{p(X|M_2)} = \frac{\int p(X|\theta_1, M_1) p(\theta_1|M_1)\, d\theta_1}{\int p(X|\theta_2, M_2) p(\theta_2|M_2)\, d\theta_2} \tag{3.4}$$

where the conditional probability $p(X|M_i)$ is called the *marginal likelihood* for model i and the θ_i are vectors of the model parameters. Since the Bayesian model comparison integrates over all parameters θ_i in each model M_i, the method does not depend on a single set of model parameters.
A problem of model fitting is the danger of overfitting. In general, a model can be adapted to reproduce any data set by an increase of the number of model parameters, but with an increasing number of parameters, the model loses the ability to describe new, unknown data. The difficulty is to find the best trade-off between the model complexity and the quality of the data fitting.
According to Kass *et al.* [87], the problem of overfitting is avoided by the Bayesian model comparison, since the usage of large numbers of parameters is intrinsically penalised during the calculation of the BF.
In this work, the BF was used for the evaluation and the selection of the best phylogenetic model. The exact implementation of the BF is described in the recent BEAST literature

[40, 42, 43].

Mathematical correlations

In general, correlations or dependencies are mathematical relations between two or more features of a data set or between two or more random variables. By definition, mathematical correlations are in the range of $[-1, 1]$. While a positive correlation value of two variables x and y describes a *the more of x, the more of y* relation (e.g. *the more $CD4^+$ cells, the higher the immune pressure*), a negative correlation of x and y means a *the more of x, the less of y* association (e.g. *the higher the viral load, the less $CD4^+$ cells survive*).
The absence of a significant correlation between two sets of observations does not necessarily mean that there is no causal relation. This was formulated by Carl Sagan, an US astronomer and popularizer of astronomy (1934 - 1996), in the famous and often cited expression "*The absence of evidence is not the evidence of absence.*"
For example the application of a weak or wrong method of analysis or an inadequate sample size could be reasons for a missing significant correlation [4]. On the other hand, the presence of a significant correlation does not guarantee a causal interrelation of the respective data [2]. The causation has to be ensured by the design of the experiment.

Statistical significance

In statistical hypothesis testing it is indispensable to check for the *statistical significance* of a correlation [53]. A test on the statistical significance analyses whether an observation is only a rare event that arose by chance (null hypothesis) or whether an observation reflects a real pattern among the observed data (alternative hypothesis).
The p-value describes the level of significance. A p-value of 0.05 states that the probability to gain the observed result (or an even extremer one) by chance is 5.0%. If the p-value of a correlation is smaller than the defined significance level, than the observed pattern is defined to be statistically significant with respect to the selected significance level.
To avoid misinterpretations, the desired significance level has to be defined prior to testing. In general it is not admissible to adjust the level afterwards when the result is already known.
In the present work, we used a threshold for the p-value of 0.05, if not stated otherwise.

Confidence interval

Closely related to the statistical significance is the *confidence interval* (CI). The CI is calculated from the observed data and estimates an interval that contains an unknown parameter of interest with a given probability.
Similar to the significance level, the level of confidence describes the reliability of a model, or to be more concise, it describes the probability that the estimated parameter of interest is comprised in the calculated interval. A CI of 95% states that one is 95% confident that the true value of the parameter of interest can be found within the interval (i.e. we face a risk of 5.0% that we have made the wrong decision). A CI of 95% defines a significance

level of $\alpha = 5\%$.
If not stated otherwise, we used a CI of 95% throughout this work.

Correlation coefficients

Many tests on mathematical relations are parametric tests that rely on specific assumptions about the data of interest, for example the assumption that the data follow a specific distribution. It is important to know and to be aware of the underlying assumptions of an applied test method and it is crucial to critically check whether the assumptions suit the specific data that are analysed.
In the following paragraphs we describe the parametric Pearson correlation coefficient that relies on the assumption that the observed data follow intrinsically a linear dependence and the non-parametric rank correlation coefficients of Spearman and Kendall.

Pearson correlation coefficient The parametric *Pearson correlation coefficient* r can be used to measure the strength of the linear dependence of samples of paired values x_i and y_i. Thus, the Pearson correlation coefficient relies on the assumption that the data intrinsically follow a linear dependence.
A perfectly positive correlation results in a value of $r = 1$, e.g. for the correlation x and $y = 2x$, while a perfectly negative linear relation results in $r = -1$, e.g. the correlation between x and $y = -2x$. A value $r = 0$ indicates that the two variables are not linearly dependent.

Spearman rank correlation coefficient The non-parametric *Spearman rank correlation coefficient* ρ estimates how well the relation between samples of paired values x_i and y_i can be described by a monotonic function. The Spearman rank correlation coefficient is related to the Pearson correlation coefficient and is mathematically defined as the Pearson correlation between the *ranked* values. For the computation of ρ, the x_i and y_i values are ranked and the coefficient of the ranks is calculated.
A Spearman rank correlation coefficient $\rho = 1$ states that the association between the x_i and y_i values is a monotonically increasing function, and $\rho = -1$ states that the association is monotonically decreasing.
If the samples contain many identical values, so called *tied values*, the computation of ρ has to be performed with care. Identical values within a sample are ranked as the mean of all their ranks, e.g. if $x_k == x_l$ occupy both rank 2 and 3, than x_k and x_l are both ranked as 2.5.

Kendall rank correlation coefficient The *Kendall rank correlation coefficient* τ is a third measure to quantify the extent of the statistical dependence between pairs of observations x_i and y_i. Similar as the Spearman rank correlation coefficient ρ, τ is a non-parametric rank correlation measure. The value of τ describes how well the rankings of the x_i's and y_i's coincide given that the spacing is non-equidistant. If $x_k > x_l$ and $y_k > y_l$, the ranks of both variables are said to be concordant, if $x_k > x_l$ and $y_k < y_l$, the ranks are discordant, and if $x_k == x_l$ and $y_k == y_l$ they are neither concordant nor discordant.

Since the Kendall rank correlation coefficient relies on the counts of concordant and discordant data pairs, pairs that are neither concordant nor discordant weaken the method, comparable to the difficulties caused by tied values upon computation of the Spearman rank correlation coefficient.
Details on the described mathematical concepts can be found in the classical textbooks of Cramer [30] and Bronstein *et al.* [17].

3.2.2. Software tools and programming languages

R scripting

We used the freely available *R* [121] software environment and the R package *MASS* [152] for standard statistics, e.g. to compute the Pearson, Kendall, and Spearman correlation coefficients and to provide significance testing. In the second part of the project, we used the BioPhysConnectoR [77] package of R for the calculation of the mutual information and for the structural coupling analysis.

Perl scripting

In addition to the Perl [26] standard routines, we used the *BioPerl* [56] package and the Perl *Statistics* [105] package. BioPerl is a developers project that comprises routines and code snippets that are useful for biological applications, and the Statistics package is a collection of a number of basic statistics methods.
Based on these packages, we developed a number of Perl scripts for sequence collection and handling, data renaming and statistics calculations.

BALLView software suite

BALLView [107, 106] is a Bioinformatics software designed for molecular modelling and visualisation. BALLView supplements the functionality of the Biochemical Algorithms Library (*BALL*) [73] with an integrated graphical user interface. In combination, BALL and BALLView provide tools and methods for the visualisation and manipulation of molecular structures, e.g methods for energy minimisation and molecular dynamics simulations with different force fields or the calculation and visualisation of the electrostatic properties of molecules.
We used the software for the modification and visualisation of different 3D crystal structures of the V3 loop of HIV.

geno2pheno[coreceptor]

Due to a prior cooperation with the scientific research group at the *Max Planck Institute for Computer Science, Saarbrücken*, we decided to use the Bioinformatics tool *geno2pheno[coreceptor]* [97] for the *in silico* predictions. geno2pheno is a classification tool based on a machine learning approach using so-called support vector machines (SVM) for two-class decisions.
Prosperi *et al.* [119] validated the geno2pheno[coreceptor] predictions in 2010. They found

a prediction accuracy of 71.4% for plasma RNA and of 70.6% for whole-blood DNA [151].

FSSM

We used *FSSM* [82, 81, 118] as an additional co-receptor prediction tool. Similar to geno2pheno, the tool first aligns the amino acid sequence of a V3 loop against a reference sequence. In the next step, the co-receptor tropism of a specific sequence is calculated based on a prediction score derived from position specific scoring matrices (PSSM) of the amino acid content of the FSSM R5 and X4 reference data set.
We used the HIV-1 subtype B X4/R5 scoring matrix for our predictions and selected the option to exclude degenerated sequences from the analysis. Regarding the publication of Low *et al.* [101], we predicted sequences with a sequence score above -8.12 as X4- or dual-tropic, while scores equal or below -8.12 indicate the use of the R5 co-receptor. The FSSM web server allows the upload of datasets of up to 200kb in size, which conforms to a maximum of 1,300 V3 loop sequences in .fasta format.

Clustal software family

Clustal [25, 149] is a software family that comprises a number of widely used tools for the computation of multiple sequence alignments (MSA). In MSAs, biologically related amino acid sequences (or less frequent nucleotide sequences) are organised in sequence tables in a way such that corresponding sequence positions of the individual sequences are contained in the same alignment column. The Clustal software suite is designed to handle large sets of sequences.
The Clustal software family consists of ClustalX, ClustalW, and most recently ClustalO. We used ClustalW with a command line interface. For our data, the most important alignment parameters were the gap associated parameters. We chose a gap opening penalty of 10 and a gap extension penalty of 0.1 for the pairwise alignments. During the extension of the pairwise alignments into an MSA, the gap extension penalty was increased to 0.2, since a value of 0.1 introduced several large gaps that massively prolonged the sequence length of the resulting MSA.
In detail, we used the following parameter settings for the computation of the multiple sequence alignments:

- Gap opening penalty: 10
- Gap extension penalty: 0.2 (0.1 for pairwise alignments)
- Protein Weight Matrix: BLOSUM
- Residue specific penalties: ON
- Hydrophilic penalties: ON
- Gap Separation distance: 4
- End gap separation: OFF
- Use negative matrix: OFF
- Delay divergent cutoff: 30%
- Keep predefined Gaps: NO

MEGA software

For the illustration and the visual inspection of the sequences, we used ClusalW [149] in the *MEGA* [95] (Molecular Evolutionary Genetics Analysis) environment. MEGA comprises a collection of tools for DNA and protein sequence analyses. We used the software version MEGA 4.0 [147] which was released in 2007. Among other applications, MEGA 4.0 enables the user to calculate multiple sequence alignments, to reconstruct phylogenetic trees, and to estimate rates of molecular evolution.
We chose the MEGA environment mainly due to the nice MSA illustration and the graphical sequence modification properties, which enabled us to do visual checks and experimental modifications of our sequence alignments.

BEAST software package

The *BEAST* [40] (Bayesian Evolutionary Analysis by Sampling Trees) cross-platform software package is a Bioinformatics tool that can be used for a number of phylogenetic inferences, for example for the reconstruction of evolutionary trees and for estimates of population measures. Mandatory input for the calculations are files containing MSAs.
The BEAST software family comprises a number of integrated tools with multiple functions. We only give a short impression of those properties of BEAST that we used to analyse our data. For details on the variety of possibilities provided by the BEAST software package, we recommend to consult the BEAST literature [40, 39, 42, 41, 43].
BEAST provides methods for the reconstruction of phylogenetic trees. The computations are not restricted to the reconstruction of single phylogenetic trees, but create a forest of independently calculated trees. The number of trees can be predefined by the user. The result is a cross-section of the most frequently created tree topologies.
BEAST uses Markov Chain Monte Carlo (MCMC) [10] as core algorithm. MCMC is used to average over tree space and thus weights each tree proportional to its posterior probability.
Based on the reconstructed phylogenetic trees, BEAST enables the calculation of important evolutionary measures, e.g. the coalescent time and the effective population size N_e. The user can decide between different molecular clock models and different options for the prior distribution and the integration of a priori knowledge.
We used the following tools of the BEAST software distribution (version 1.6.1):

BEAUti	(Bayesian Evolutionary Analysis Utility) provides a graphical user interface to prepare .xml input files for BEAST based on sequence alignment files in .nex format.
BEAST	reconstructs phylogenies using an MCMC algorithm based on a tree weighting process [40, 39].
LogCombiner	combines the output log and tree files from different BEAST reconstruction runs.
LogAnalyser	enables analyses of the result and the quality of the reconstruction runs.
Tracer	allows a visual inspection of the BEAST log files, especially of the reconstruction traces created by the MCMC calculations, and enables the calculation of an estimated time course of N_e [42, 41].

TreeAnnotator consolidates the information of a forest of phylogenetic trees into one consensus tree [43].

Figtree enables the visualisation and the modification of the reconstructed phylogenetic trees [122].

BEAST workflow We combined the BEAST tools to a sequential workflow to perform phylogenetic reconstructions of the viral sequences of the study patients. Using ClustalW in the MEGA environment, we translated the available nucleotide (nt) sequences of the V3 loop into amino acid (aa) sequences, calculated the aa MSA for each patient, and back-translated the aa MSAs into nt MSAs. The resulting patient specific nt MSAs were converted into .nex file format and separately uploaded into BEAUTi to create the .xml input files for the phylogenetic calculations of BEAST.
After a predefined number of BEAST reconstruction runs, LogCombiner was used to combine the multiple .log files into one summary file. The combined result was inspected with LogAnalyser and Tracer to evaluate the reconstruction quality and to decide about the so called *burn-in*, the number of initial BEAST runs necessary to calibrate the parameters for each specific data set. In general, the burn-in comprised the first ten percent of the runs, the exact value for each data set was individually determined.
Excluding the burn-in, Tracer was then used to the determine the effective population size N_e over time, also known as Bayesian Skyline reconstruction. In parallel, the forest of reconstructed phylogenetic trees was consolidated into one summary tree by TreeAnnotator. Finally, FigTree was used to illustrate and inspect the consensus trees.
Due to the close association with the study data, the model and parameter selection is described in detail in Subsection 3.4.1 within the data Section 3.3.

RAxML

In addition to the BEAST software family, we used the *RAxML*-VI-HPC [144, 145] software package for the reconstruction of evolutionary phylogenies. RAxML-VI-HPC is an acronym for Randomized Axelerated Maximum Likelihood for High Performance Computing, version VI. The software was developed to enable fast and parallel calculations of multiple runs on distinct starting trees. Details can be found in the RAxML publications [144, 145].
Using a bootstrapping approach, RAxML creates a set of independent random starting trees for the phylogenetic reconstruction. After the independent reconstruction runs, the bootstrapping results are summarised on the tree with the best likelihood. In the last step, the pairwise distances are extracted.
RAxML can directly be operated from the command line, is less complex than the bundle of BEAST tools, and enables fast phylogenetic reconstructions, thus we used the software to get a fast estimate of the viral phylogenies and to cross-check the BEAST results.

3.3. Data

The first part of our project is based on longitudinal patient data that were collected in the course of the joint project *Monitoring of resistant HIV in newly and chronically infected HIV patients in Germany - Evolution of HIV-genotype and phenotypes during antiretroviral therapy*. Starting in 1999, it was initially planned to collect sequential blood samples of HIV-1-infected patients at the *Universitätsklinikum Frankfurt am Main* during regular visits of the patients at the hospital to monitor the intra-host evolution of the HIV genome during infection. In addition, we received selected information from the patient records of the participants of the study.
The study was authorized by an ethical commission and informed consent was obtained from all participants before they were included in the study.

3.3.1. Patient records

Our cooperating physicians at the hospital provided us with a data base of selected data from the patient records. Among others, the files comprised the sex, the date of the first HIV positive test, and the putative way of infection. In some rare cases also the date of the last HIV negative test was given. The patients were identified by an unique, patient-specific number.
During the sequential visits of the patients at the hospital, the number of $CD4^+$ cells and the viral load were documented. Occasionally, also the number of $CD8^+$ cells was determined. Based on these data, the stage of the disease with respect to the CDC classification system [21, 133] was regularly ascertained. In addition, the drug therapy and the reason for an eventual therapy change was reported in most cases. Less frequently, the drug related side effects or patient-induced therapy discontinuations were documented in the patient records.
These general disease-related data were complemented by information on occasional tests of the patients for co-infections (e.g. human cytomegalo virus, syphilis, and hepatitis A, B, and C). In the case of stationary visits of the patients at the hospital, the reason and the duration of the stay at the hospital are also noted in the patient records.

3.3.2. Blood samples

According to the study protocol, it was planned to include 300 patients into the study and to collect blood samples of the participants every three to six month up to ten years. Without further preparation, the refrigerated full blood samples were directly sent to the *Paul-Ehrlich-Institut* in Langen, where the blood samples were processed for sequencing. The elaborative isolation of the virus from the blood was performed in the wet lab of the division of Virology in the section *AIDS, New and Emerging Pathogens* of the institute.
During the process of virus isolation, it is important to ensure the absence of a selective pressure on the virus. Therefore, a multi-step protocol was developed for sequence extraction. According to Werner *et al.* [159] breeding was successful in about 87 % of all cases. For a detailed description of the sequencing process we recommend the respective publication [159].

3.3.3. *In silico* processing of sequence data

We were provided with the nucleotide (nt) V3 loop sequences of 106 patients. Collected during more than years of research, the digital data were organised in multiple directories and stored in hundreds of files of different data types. Unfortunately, neither a uniform system of file and folder naming or folder structure was used, nor a unified file (.doc, .txt, .xls) and sequence (.ab1, .seq) format or consistent information content of the files. The sequencing protocols contained the nt sequences as well as some additional information not relevant for this study. It was very challenging to ensure an automated data analysis of the complete data.
We developed a Perl script and searched all directories and folders for sequence data. Upon data collection, an unique identifier was created for every sequence, consisting of:

- the unique patient number,
- the number of the blood sample,
- an enumerator counting the sequences within one blood sample,
- the number of days that passed since the first HIV-positive test of the patient

As an example, the identifier 004_P411_01_d5 describes sequence 01 of patient 004 from sample P411 that has been collected five days after the first HIV-1 positive test. The identifiers were later supplemented by the predicted co-receptor (compare Section 3.3.3).
In the first approach, we extracted all sequences available in any kind of text format (.txt, .doc, .xls) and ended up with 2,349 nucleotide sequences, on average 42 sequences of four visits at the hospital. Some of the sequences contained stop codons or were extremely shortened (in the range of 100 nt instead of 300 nt). A further inspection of the data revealed that the sequences covered only half of all blood samples that were documented to be collected in the course of the project. Therefore, we decided to go one step back in the protocol and to have a closer look onto the raw sequencing data.
We checked all directories for .seq and .ab1 files, manually inspected the data, and exported the nt sequences. In a subsequent refinement step, fragmented sequences of a length of less than 200 nt (i.e. less then two thirds of the target region) were removed. On average, the remaining sequences consisted of 300 to 330 nt. Using ClustalW [149] in the MEGA environment [95], the nt sequences were translated into amino acid (aa) sequences and an MSA was calculated. Finally, we excluded aa sequences that contained stop codons or large gaps within the V3 loop.
This extended data collection resulted in a data set of 3,132 sequences, about 800 sequences richer than the original set of 2,349 sequences. The sequences were extracted from the blood of 47 male and eight female patients.
Since we planned to examine the sequence evolution over time, we determined a subset of all patients with at least four sequential blood samples and remained with 2,224 sequences of 34 male and two female patients - the final data set contained an average of 60 sequences per six visits per patient. Figures 3.7 and 3.8 illustrate the distribution of sequences for each patient and for each blood sample.

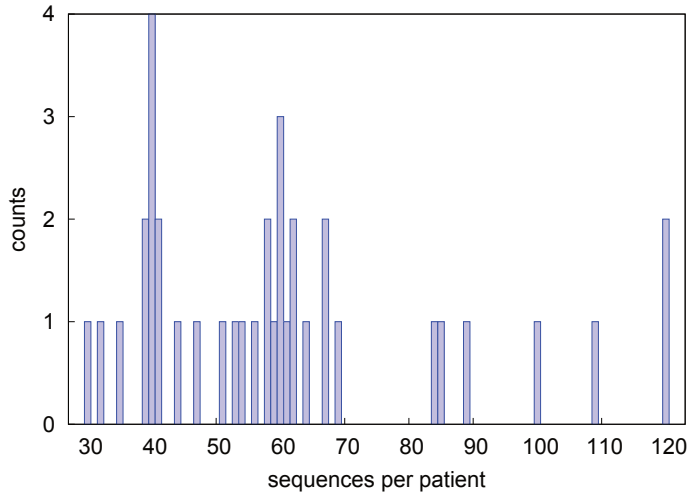


Figure 3.7.: **Distribution of the number of sequences per patient**
The illustration shows the number of sequences per patient. 30 to 120 sequences per patient were extracted during the course of the study.

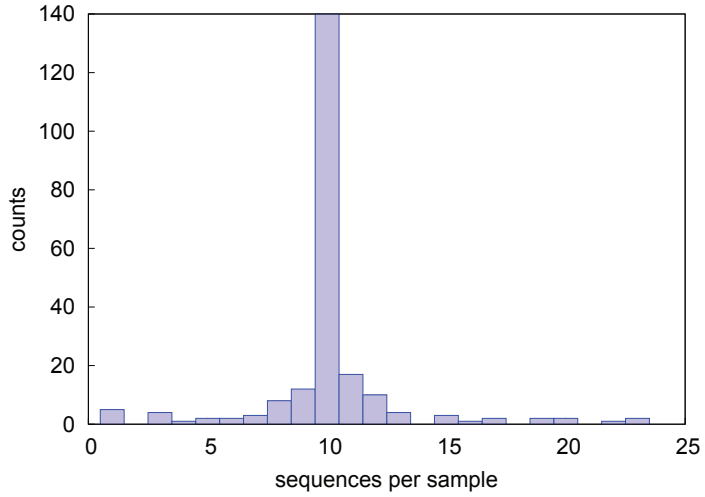


Figure 3.8.: **Distribution of the number of sequences per blood sample**
The distribution of sequences per blood sample shows that the intended number of ten sequences per sample could be achieved for the majority of samples. Operating the sequencing process at full capacity resulted in more than the envisaged ten sequences for some samples.

In silico co-receptor determination

We started the data analyses with an *in silico* determination of the co-receptor. Using geno2pheno[coreceptor] [97] for the classification, we applied a FPR of 10% according to the recommendations of the European Consensus Group on clinical management of HIV-1 tropism testing [151] (i.e. sequences for which the geno2pheno classifier reported a FPR $< 10\%$ were labelled as X4- or dual-tropic, while sequences with a reported FPR of the classifier $\geq 10\%$ were labelled as R5-tropic). Of the 2,224 viral sequences, 1,557 sequences

were predicted to be R5-tropic and 677 sequences were predicted to be X4-tropic.
We used ClustalW to calculate two independent multiple sequence alignments (MSA) for the R5-and the X4-tropic data set to get an impression of the genotypic differences. Based on the MSAs, we derived an R5 and X4 consensus sequence by the determination of the most probable amino acid at each position. The resulting consensus sequences are:
R5: `CTRPNNNTRK S--IHIGPGR--AF YATGDIIGDI RQAHC`
X4: `CTRPNNNTRK R--IHIGPGR--AF YTTGAIIGDI RKAHC`
We excluded the gap positions 12/13 and 21/22 from the sequences to adapt the sequence positions to meet the 11/25 motif for co-receptor description:
R5: `CTRPNNNTRK SIHIGPGRAF YATGDIIGDI RQAHC`
X4: `CTRPNNNTRK RIHIGPGRAF YTTGAIIGDI RKAHC`
To illustrate the position specific amino acid probabilities of the MSAs, we used WebLogo [31, 134] to create the sequence logos presented in Figure 3.9.

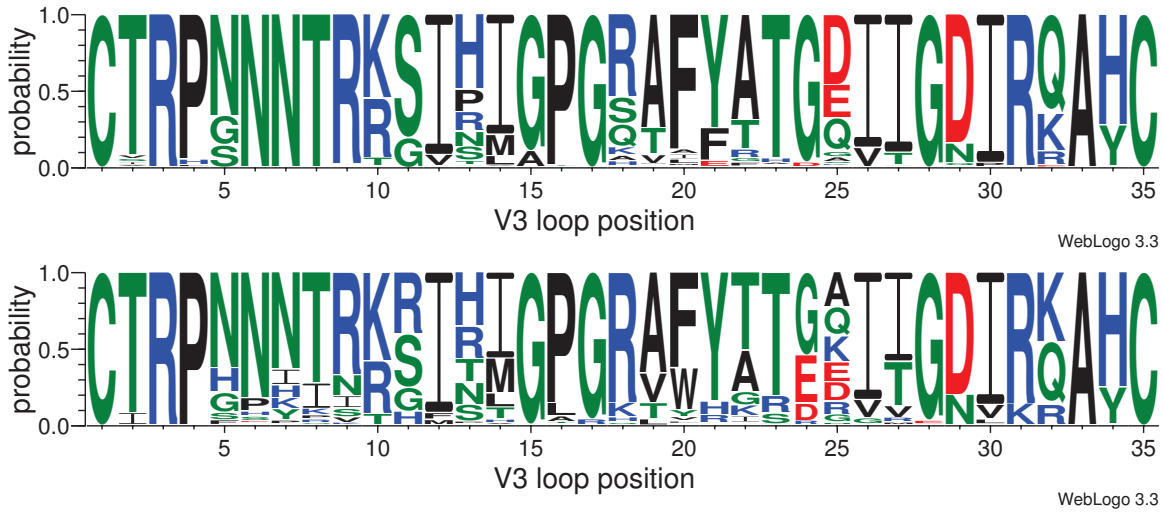


Figure 3.9.: **Sequence logos of patient sequences with predicted R5 and X4 phenotype** The sequence logos illustrate the MSA of the 1,557 R5 predicted sequences and the MSA of the 677 X4 predicted sequences.
The hight of the letters (y-axis) describes the position specific amino acid probability. The colours illustrate chemical amino acid properties: blue: basic (K,R,H), red: acidic (D,E), green: polar (C,G,N,Q,S,T,Y), and black: nonpolar/hydrophobic (A,F,I,L,M,P,V,W).

Comparing the R5 and X4 MSA and the respective consensus sequences, we found differences in the tropism defining positions of the V3 loop, namely in position 11 (R5: S, X4: R) and position 25 (R5: D, X4: A) of the sequence. Furthermore, positions 22 (R5: A, X4: T) and 32 (R5: Q, X4: K) showed differing consensus amino acids. The exchange of the negatively charged aspartic acid (D) in position 25 of the R5 consensus sequence by the non-polar amino acid alanine (A) of the X4 consensus sequence, as well as the observation of the positively charged amino acids arginine (R) and lysine (K) in positions 11 and 32 of the X4 consensus sequence confirmed the well-known switch of X4-tropic sequences towards a positive net charge [165, 7, 22, 54, 141]. Thus, the observed differences between the R5 and X4 data set indicated the quality of our data.
For a further validation of the data, we calculated the Hamming distance of any pair sequences of our data sets. In general, we expected lower Hamming distances between

any two sequences of the more conserved R5 population (R5-R5), and higher Hamming distances for sequence pairs of the more heterogeneous X4 population (X4-X4). A calculation of the Hamming distances of mixed sequence pairs (X4-R5 or R5-X4) was expected to result in Hamming distances in between the two regimes. Figure 3.10 depicts the frequencies of the respective Hamming distances. The curves confirmed our expectations. While most R5-R5 sequence pairs showed an amino acid distance of nine, the most frequent X4-X4 distance was 13, and most mixed R5-X4 pairs differed 12 by amino acids.

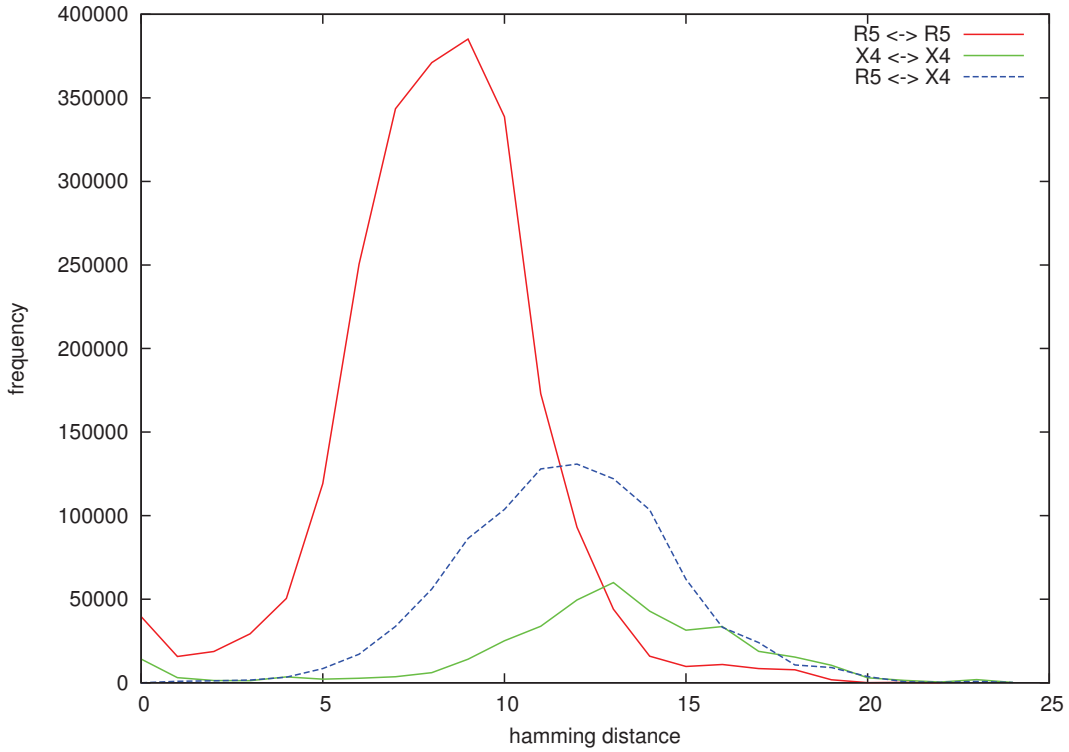


Figure 3.10.: **Hamming distances of sequences of selected patients**
The figure illustrates the Hamming distance distribution of all R5-R5, X4-X4, and R5-X4 sequence pairs of the patient sequence data.
Both the Kolmogorow-Smirnov [92, 143] and the χ^2 [114] test confirmed that the Hamming distance distributions of the data sets were significantly different:
R5 vs. X4 pairs: $p - value =< 0.0001$, R5 vs. mixed pairs:$p - value < 0.0001$, X4 vs. mixed pairs: $p - value < 0.0001$.

Interestingly, the illustration of the X4 distances indicated a second peak at a Hamming distance of 16. Upon literature search, we found a publication of Bozek *et al.* [15] (belonging to the group of geno2pheno developers) that also showed this double peak. Bozek *et al.* analysed the distribution of Blosum62 distances of an R5 and X4 population, but they did not further address this phenomenon in their study.
A closer inspection of our data showed that the second peak resulted from three sequence positions of the X4 MSA that were occupied by more than two predominant amino acids, namely positions 11, 13, and 32. The increase in the Hamming distance from 13 to 16 resulted from these three positions.
Thus, the observed Hamming distance distribution of our data was in good agreement with

the prior described V3 sequence population study of Bozek *et al.* [15] and gave further evidence for the validity of our R5 and X4 patient data set.

3.3.4. Data processing of clinical measurements

Subsequent to the processing and the phenotypic classification of the sequence data, we inspected the electronic patient records. We found that the data of one male patient was missing, thus we excluded this patient from further studies. In addition, we corrected some entries of the clinical measurements and some data on therapy and co-infections for misspellings (i.e. the format of the date and the use of the wrong line separator), and removed duplicated entries.
To build a better basis for the following analyses, all dates of the samples, of the clinical measurements, and of the drug therapy were uniformly transformed into *days since the first HIV-positive test*.

3.4. Results

In this section we first present the phylogenetic reconstructions of the patients sequence data. Since the process of model selection and parameter fitting strongly influences the result of the phylogenetic reconstructions, we describe the search for a suitable phylogenetic model in this section.
Based on the phylogenetic reconstructions, we determined the correlations between the evolutionary data and the clinical measurements. An analysis of the course of the infection of the patients under successful HAART and the possibility of a back-switch of the viral population from X4 towards R5 tropism are addressed at the end of this section.

3.4.1. Model selection using Bayes factor analysis

The reconstruction of phylogenies and the estimation of population sizes were performed with the BEAST software package [40, 39, 42, 41, 43]. The the initial model selection process and the parameter setting were tied up to the prior study of Kamp *et al.* [85] in the course of the project. Based on their work, we did not consider the HKY or TN93 nucleotide substitution model for the phylogenetic reconstructions, but restricted our analyses to the general time-reversible (GTR) models [96, 130].
Furthermore, we applied coalescent models [91, 90], since the coalescent approach in general relies on the assumption that only a limited sample of a population is studied, instead of the whole population, which is valid for our data. For the reconstruction of the viral phylogenies, the empirical base frequencies, the gamma distribution of among-site rate variation, and the rate of invariant sites (+I) [65, 154] were derived from the sequence data. The analyses were initiated from randomly generated starting trees.
We started the model selection process from the following subset of evolutionary models:

- Site heterogeneity model
  - gamma
  - gamma+I
- Molecular Clock Model
  - Relaxed uncorrelated exponential
  - Relaxed uncorrelated lognormal
  - Strict Clock
- Tree prior
  - Coalescent: Constant Size
  - Coalescent: Exponential Growth
  - Coalescent: Logistic Growth
  - Coalescent: Expansion Growth
  - Coalescent: Bayesian Skyline
  - Coalescent: Extended Bayesian Skyline

In addition to the data of patient 265 that was analysed extensively by Kamp *et al.* [85] in an earlier study, comprehensive reconstructions of three additional patients served as a decision basis to test a variety of model parameters. The patients 041, 107, and 132

were selected since their data resembled different courses of disease. We used Bayes Factor statistics [87, 110] to analyse the suitability of the different phylogenetic models and to measure their capability to perform phylogenetic reconstructions. The Bayes factors (BF) of the different models for the selected patients are summarised in Table 3.2.

Table 3.2.: **Bayes factor of four phylogenetic models and three patients**
The table summarises the Bayes factor of four general time-reversible (GTR) Bayesian skyline coalescence models (BSC) for three different data sets and 10^6 reconstruction steps. The empirical base frequencies, the gamma distribution of among-site rate variation, and the rate of invariant sites (+I) were derived from the sequence data.

patient 041	ln $p(model\|data)$
GTR gamma+I constant BSC	-822,04
GTR gamma+I linear BSC	-824,85
GTR gamma linear BSC	-829,23
GTR gamma BSC expansion	-829,56
GTR gamma extended linear BSC	-831,27
GTR gamma constant BSC	-831,83
GTR gamma extended exponential BSC	-832,37
GTR gamma constant BSC	-905,43
GTR gamma extended exponential BSC	-907,73
GTR gamma BSC expansion	-4707,58
patient 107	
GTR gamma+I linear BSC	-1211,86
GTR gamma linear BSC	-1214,44
GTR gamma+I constant BSC	-1214,52
GTR gamma+I extended linear BSC	-1215,59
GTR gamma extended linear BSC	-1216,41
GTR gamma extended exponential BSC	-1216,80
GTR gamma constant BSC	-1217,03
GTR gamma+I extended exponential BSC	-1251,60
patient 132	
GTR gamma linear BSC	-1533,88
GTR gamma+I constant BSC	-1536,35
GTR gamma constant BSC	-1539,95
GTR gamma+I linear BSC	-1545,88
GTR gamma BSC expansion	-1546,88
GTR gamma+I extended exponential BSC	-1550,45
GTR gamma+I extended linear BSC	-1554,92
GTR gamma extended exponential BSC	-1556,44
GTR gamma extended linear BSC	-1559,63
GTR gamma+I BSC expansion	-1573,95

During these first analyses, we found some of the models to be less suitable to reconstruct our sequence data. The Bayesian skyline expansion model as well as the extended linear and exponential models did not fit our data well. Furthermore, the reconstruction runs for some data sets finished without being able to calculate any phylogenetic reconstruction.

Thus we reduced the analyses to the four best-ranked models: the GTR Bayesian skyline model with constant or linear coalescence and with gamma distribution of among-site rate variation, with (+I) or without invariant sites.
Since we recognised a danger of overfitting of the models to the individual patient data, we included further patients into the model selection process. The decision about the suitable evolutionary model for the phylogenetic reconstruction finally was based on reconstruction runs results of the four preselected models for 14 different patient data sets. The ranked BF are listed in Table 3.3 (with best model ranked 1, least model ranked 4).

Table 3.3.: **Bayes factor of four phylogenetic models and 14 patients**
Comparison of the Bayes factor of four general time-reversible (GTR) Bayesian skyline coalescence models (BSC) for 14 patients (ID):
(top) $1 \cdot 10^6$ reconstruction steps, (bottom) $2 \cdot 10^8$ reconstruction steps.
The empirical base frequencies, the gamma distribution of among-site rate variation, and the rate of invariant sites (+I) were derived from the sequence data. The number in parentheses gives the rank of the respective model among the four analysed models.

ID	linear model + I		constant model + I		linear model		constant model	
041	-824,85	(2)	-822,04	(1)	-829,23	(3)	-831,83	(4)
107	-1211,86	(1)	-1214,52	(3)	-1214,44	(2)	-1217,03	(4)
132	-1545,88	(4)	-1536,35	(2)	-1533,88	(1)	-1539,95	(3)
005	-1220,86	(1)	-1223,85	(2)	-1229,43	(4)	-1226,47	(3)
007	-1571,46	(4)	-1571,12	(3)	-1569,10	(2)	-1559,84	(1)
013	-709,44	(2)	-707,91	(1)	-710,26	(3)	-712,32	(4)
040	-776,82	(2)	-776,71	(1)	-780,62	(4)	-779,49	(3)
085	-1621,86	(4)	-1621,32	(3)	-1613,31	(2)	-1610,40	(1)
127	-1266,96	(1)	-1269,40	(4)	-1267,68	(2)	-1268,31	(3)
180	-1289,11	(4)	-1285,82	(2)	-1286,08	(3)	-1284,01	(1)
190	-1078,23	(1)	-1083,53	(3)	-1084,18	(4)	-1078,26	(2)
194	-1043,69	(2)	-1043,50	(1)	-1051,83	(4)	-1050,29	(3)
196	-771,34	(1)	-772,35	(2)	-775,96	(4)	-775,03	(3)
265	-687,83	(4)	-686,49	(3)	-683,72	(2)	-682,29	(1)

41	-823,56	(3)	-822,99	(1)	-823,12	(2)	-830,29	(4)
107	-1216,00	(4)	-1215,62	(2)	-1215,94	(3)	-1215,09	(1)
132	-1554,20	(4)	-1545,02	(1)	-1551,22	(2)	-1553,33	(3)
190	-1075,83	(1)	-1084,01	(4)	-1083,60	(3)	-1078,06	(2)

Considering the BF, we preferred the performance of the linear or constant Bayesian skyline model with codon handling using the GTR Gamma model with invariant sites (+I), but we were aware that the differences of the best rankings were very small.
To gain further information about the suitability of the preferred model, we increased the number of calculations to $2 \cdot 10^8$ reconstruction steps per data set, and reduced the number of data sets again to the four patients resembling a different course of the disease. The respective BF and rankings, subsumed in the lower part of Table 3.3, showed only subtle differences in the BF statistics, but the analyses again confirmed the danger of overfitting when using a proportion of invariant sites.
We decided to perform the phylogenetic reconstructions of all patient data sets with one

single model and parameter set. A fixation of the reconstruction parameters yielded at least some comparability of the results, since the data of the individual patients already showed a broad variability. We finally selected the linear GTR Bayesian skyline model with relaxed clock coalescence and uncorrelated lognormal distribution of among-site rate variation for the reconstruction of the phylogenies of the study patients. To decrease the danger of overfitting, we restricted the use of invariant sites.

Parameter setting

Following the model selection process, we determined the parameters for the phylogenetic reconstruction. Using the empirical base frequencies of the MSA, we selected four gamma categories (A,G,T,C) and three unlinked codon positions, thus substitution rate parameters, rate heterogeneity, and base frequencies were set to be unlinked across codon positions. Using the relaxed clock coalescence model with uncorrelated lognormal distribution and linear Bayesian skyline estimate, we started the calculations from randomly generated starting trees.
Since BEAST can integrate information about the timely course of sequences, we included the age of the sequences, defined as the number of days since the first HIV-positive test. According to the study of Kamp *et al.* [85], we mainly stuck to the default BEAST model priors and exclusively adjusted the following parameters:

- CPx.mu
  - start 0.5
  - min 0
  - max 1,000
- ucld.mean
  - start 0.5
  - min 0
  - max 1,000

With these parameters, we performed 10^9 reconstruction steps for each set of sequences. Saving the trees to file every 50,000 steps yielded 20,000 phylogenetic trees per patient. We selected a posterior probability limit of 0.5 with maximum clade credibility and median heights to summarise the trees with TreeAnnotator [43]. For the skyline reconstruction with Tracer [41], the date of the latest sample served as starting point. All other Tracer parameters were extracted from the BEAUTi .xml files.

3.4.2. BEAST phylogenetic reconstruction

After the determination of the reconstruction model and the phylogenetic parameters, we performed the reconstruction runs for the sequence data of each individual patient. The N_e was extracted from the Bayesian skyline reconstructions and is presented in Section 3.4.5.

3.4.3. RAxML phylogenetic reconstruction

RAxML was used to form a second view on the phylogenetic reconstructions. The reconstruction model and the parameter settings were adopted from the BEAST selection process. We found that RAxML in general yielded comparable phylogenetic reconstructions

and successfully confirmed the BEAST results.

3.4.4. Associations between clinical parameters

We started the data analysis with a determination of the correlations between the clinical parameters HAART, viral load, $CD4^+$ cell count, and co-receptor usage.

Association of HAART and the viral load

Analysing the correlations between the clinical measurements, we focussed on the question whether the early illustration of the clinical course of an HIV-1 infection presented by Pantaleo *et al.* [51] fits the data of recent HAART treated patients.
To get a first impression of the clinical course of disease of the study patients, we illustrated the clinical data in Figures 3.11 and 3.12. While most patients were under long term therapy, 8 patients (013, 040, 051, 062, 100, 127, 197, and 212) were included into the study upon HAART initiation. HAART treatment of patient 004 was initiated at the day of the last clinical measurement.
As the illustrations indicate, the HIV therapy in general successfully suppressed the viral growth. The rare cases of therapy failure (patients 005, 010, 026, 041, 072, 107, and 196) showed a dramatical increase of the viral load within a short period of time, with a beginning decline of the number of $CD4^+$ cells. In most cases, a subsequent change of therapy immediately suppressed the viral load again below the limit of detection (compare data of patients 041, 072, 107, and 196).
Only five patients (005, 010, 026, 109, and 180), all of them therapy experienced, showed a high viral load persisting for more than a few days. Patient 005 had a transient increase in viral load due to therapy discontinuation, while patient 010 was documented to take the drugs irregularly. For patient 026, a therapy failure in the late stage of the disease was documented, and for patient 109, the uptake of several drugs in parallel led to drug interactions. The therapy changes of patient 180 mainly were a consequence of side effects and of frequent co-infections.
Despite these patient specific exceptions, the clinical data reported an immediate decrease of the viral load upon a therapy initiation or a therapy change, in most cases suppressing the viral load below the detection limit within a few days. Contemporaneously with the initiation of the (new) therapy, the number of $CD4^+$ cells started to recover, but at a lower speed than the observed drop of the viral load.
In summary, the visual inspection of the clinical data gave first evidence for a negative correlation between the viral load and the number of $CD4^+$ cells in HAART treated patients — a fact, that is already known for untreated patients [50, 64, 104]. In addition, the data indicated that successful HAART might be able to pause the course of the infection or to even turn back the disease progression in time.

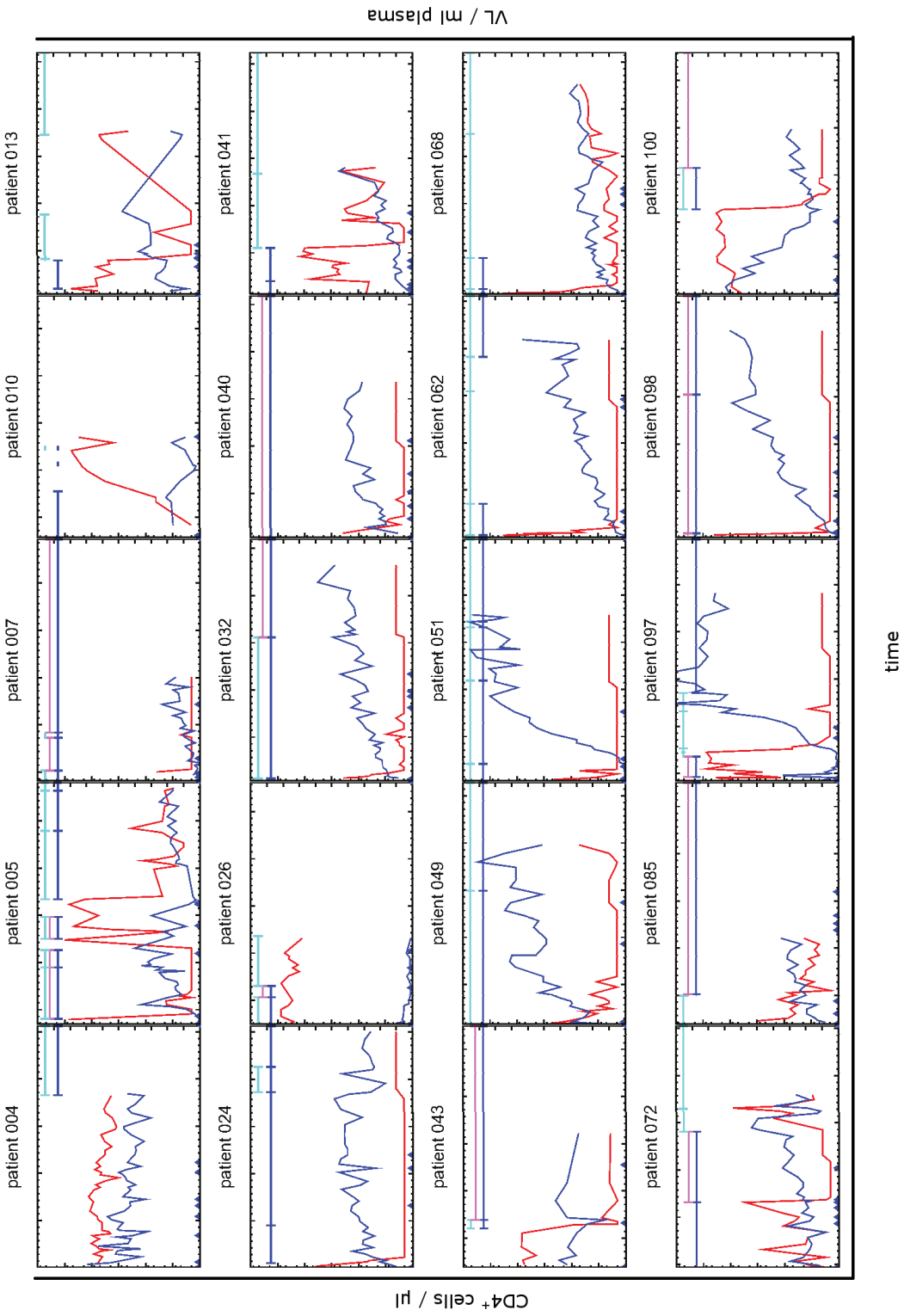


Figure 3.11.: **Course of the disease of 20 patients**
The figure illustrates the course of the CD4$^+$ cells (green) and the viral load (red) of 20 study patients on an uniform time interval of seven years (starting at the time of the first patient specific measurement). Blue triangles along the x-axis indicate the time points of the sequenced blood samples, and the HIV therapy is depicted as horizontal lines in the upper part of each plot. Blue lines mark nucleosidic (NRTI) and violet lines non-nucleosidic (NNRTI) reverse transcriptase inhibitors, cyan lines mark protease inhibitors (PI) and red lines mark integrase inhibitors (II). An arrow on the therapy lines indicates the initiation of the respective therapy.

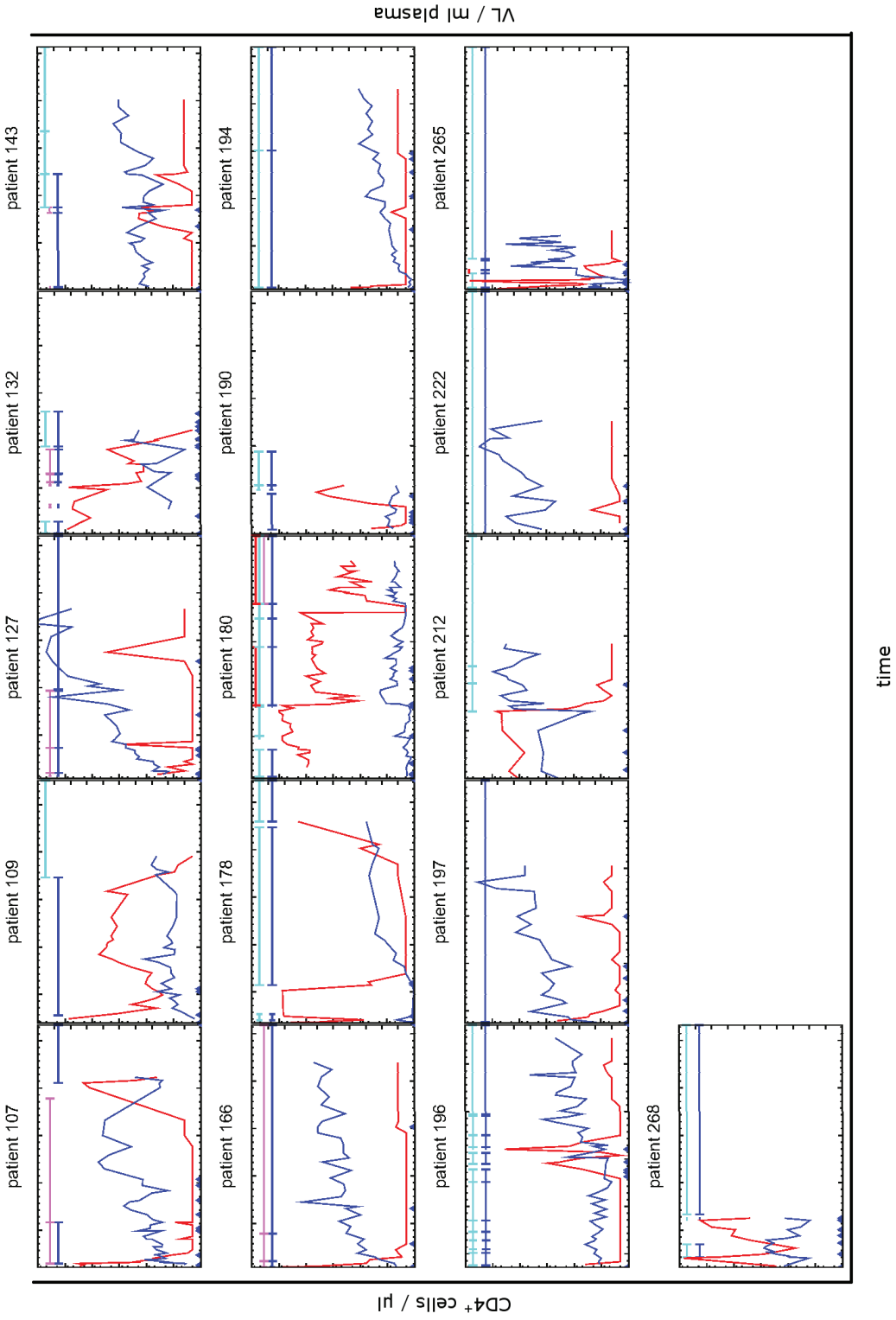


Figure 3.12.: **Course of the disease of 16 patients**
The figure illustrates the course of the CD4$^+$ cells (green) and the viral load (red) of 16 study patients on an uniform time interval of seven years (starting at the time of the first patient specific measurement). Blue triangles along the x-axis indicate the time points of the sequenced blood samples, and the HIV therapy is depicted as horizontal lines in the upper part of each plot. Blue lines mark nucleosidic (NRTI) and violet lines non-nucleosidic (NNRTI) reverse transcriptase inhibitors, cyan lines mark protease inhibitors (PI) and red lines mark integrase inhibitors (II). An arrow on the therapy lines indicates the initiation of the respective therapy.

Viral load and the number of CD4$^+$ cells

Based on the indications derived from the visual inspection of the clinical data, we hypothesised to find negative correlation between the number of the CD4$^+$ cells and the viral load of HAART treated patients. We observed that successful therapy suppressed viral load and enabled the number of CD4$^+$ cells to recover, while therapy discontinuation or failure facilitated an increase in viral growth and a decrease of the number of CD4$^+$ cells. From these findings, we expected the course of the disease to step forth and back in progression, in contrast to the linear course of the disease described by Pantaleo *et al.* for untreated patients [51, 50, 64].
To address this question, we computed the correlations between the clinical measurements. We extracted the numbers of CD4$^+$ cells and the viral load from the patient records and determined the Pearson correlation coefficient separately for the data of each patient (compare Table 3.4).

Table 3.4.: **Pearson correlation of the CD4$^+$ cell count and the viral load**
The table lists the Pearson correlation coefficient r and the respective p-values for the association of the of the CD4$^+$ cell count and the viral load. 'ID' is the unique patient identifier and the column 'obs.' gives the number of observations.

ID	obs.	r	p-value	ID	obs.	r	p-value
004	41	0.290	0.066	098	33	-0.348	0.047
005	44	-0.397	0.008	100	33	0.467	0.006
007	28	0.064	0.745	107	42	-0.488	0.001
010	9	-0.780	0.013	109	29	-0.417	0.024
013	20	-0.840	<0.001	127	29	0.103	0.595
024	43	-0.606	<0.001	132	15	-0.632	0.011
026	13	0.298	0.323	143	38	-0.225	0.174
032	41	-0.441	0.004	166	35	-0.530	0.001
040	25	-0.355	0.082	178	21	-0.484	0.026
041	26	-0.192	0.348	180	68	-0.497	<0.001
043	11	0.123	0.718	190	12	-0.540	0.070
049	36	-0.475	0.003	194	40	-0.144	0.375
051	34	-0.143	0.418	196	44	-0.162	0.293
062	45	-0.461	0.001	197	22	-0.254	0.254
068	45	-0.381	0.010	212	21	-0.812	<0.001
072	33	-0.061	0.737	222	12	-0.200	0.533
085	38	-0.600	<0.001	265	44	-0.552	<0.001
097	40	-0.839	<0.001	268	11	-0.872	<0.001

We found negative correlations between the number of CD4$^+$ cells and the viral load for 30 of 36 the patients, with a p-value <0.05 in 20 data sets. In contrast, 6 of 36 patients showed a positive Pearson correlation (p-value <0.05 only for patient 100).
Though the Figures 3.11 and 3.12 indicated a negative association of the observed data, the totality of all correlations could not support our hypothesis. An additional calculation of the Spearman and Kendal correlation neither could confirm our idea (data not shown). Thus, based on the present study data, we could not decide whether Patients under

HAART treatment show a comparable association between the number of $CD4^+$ cells and the viral load as therapy naïve patients.

3.4.5. Associations between evolutionary parameters

Following the correlation analyses of the clinical data, we analysed the correlations between the evolutionary measures diversity, divergence, and the effective population size N_e.

Diversity and the effective population size N_e

The internal Bayesian skyline reconstruction method of BEAST estimates the putative course of the effective population size N_e over time that would result in the reconstructed phylogeny of sequences. N_e describes the genetic range of sequences within a population, thus the N_e estimates are comparable to the course of the population diversity over time as defined in Section 3.2.1. Following this idea, we hypothesised to find a positive linear correlation of N_e and the population diversity.
We illustrated the N_e and the diversity of each patient to get a first visual impression of the data (compare Figures 3.13 and 3.14). For ease of computation, we scaled both measures onto the interval $[0, 1]$.
In the next step, we computed the Pearson correlation coefficient r of N_e and the viral diversity for a statistical evaluation. The Pearson correlation values are listed in Table 3.5. We found a positive correlation in 28 of 36 patients, ranging from almost zero (0.02) to strong (0.9) positive correlations. Due to the small number of observations (i.e. four to 12 sequenced blood samples), the level of significance of the results was poor. The envisaged p-value <0.05 was missed in all but one data set (patient 107). Eight of 36 samples showed negative Pearson correlation values ranging from -0.02 to -0.57, but none of the correlations was significant.
A repetitive analysis using the Spearman and Kendal rank coefficients confirmed these findings. We calculated a positive correlation for 28 respective 26 data sets, yielding significant results (p-value <0.05) only for patients 010, 196, and 265 with the Spearman rank coefficient and for patients 010 and 265 with the Kendal rank coefficient (data not shown).
In summary, the we were not able to confirm our hypothesis of a positive linear correlation between N_e and the viral diversity due to missing empirical evidence.

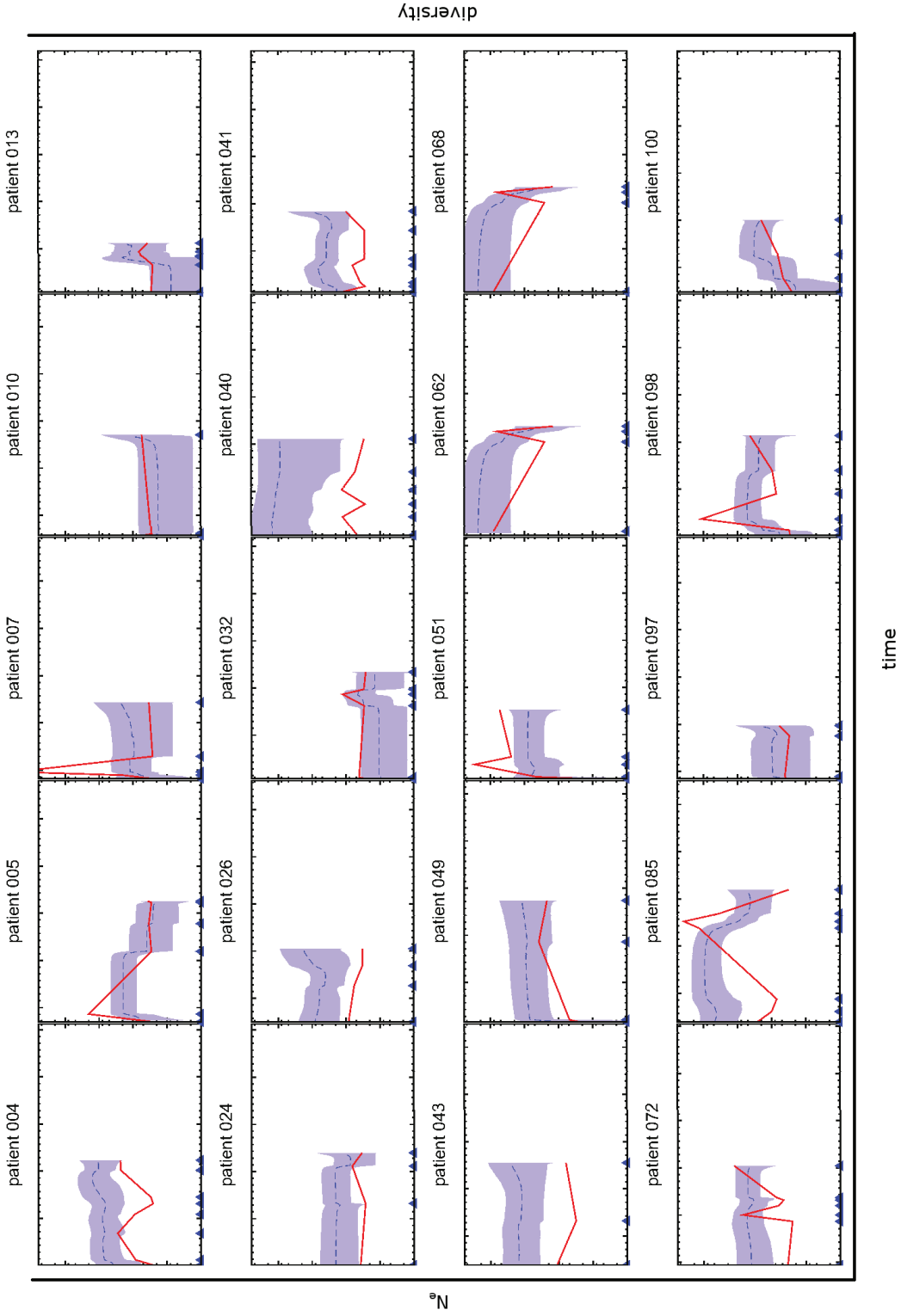


Figure 3.13.: **Course of N_e and diversity of 20 study patients**
The figure illustrates the course of the estimated effective population size N_e (blue line) with 95% confidence interval and the course of diversity (red line) of 20 study patients on an uniform time interval of seven years (starting at the time of the first patient specific measurement).

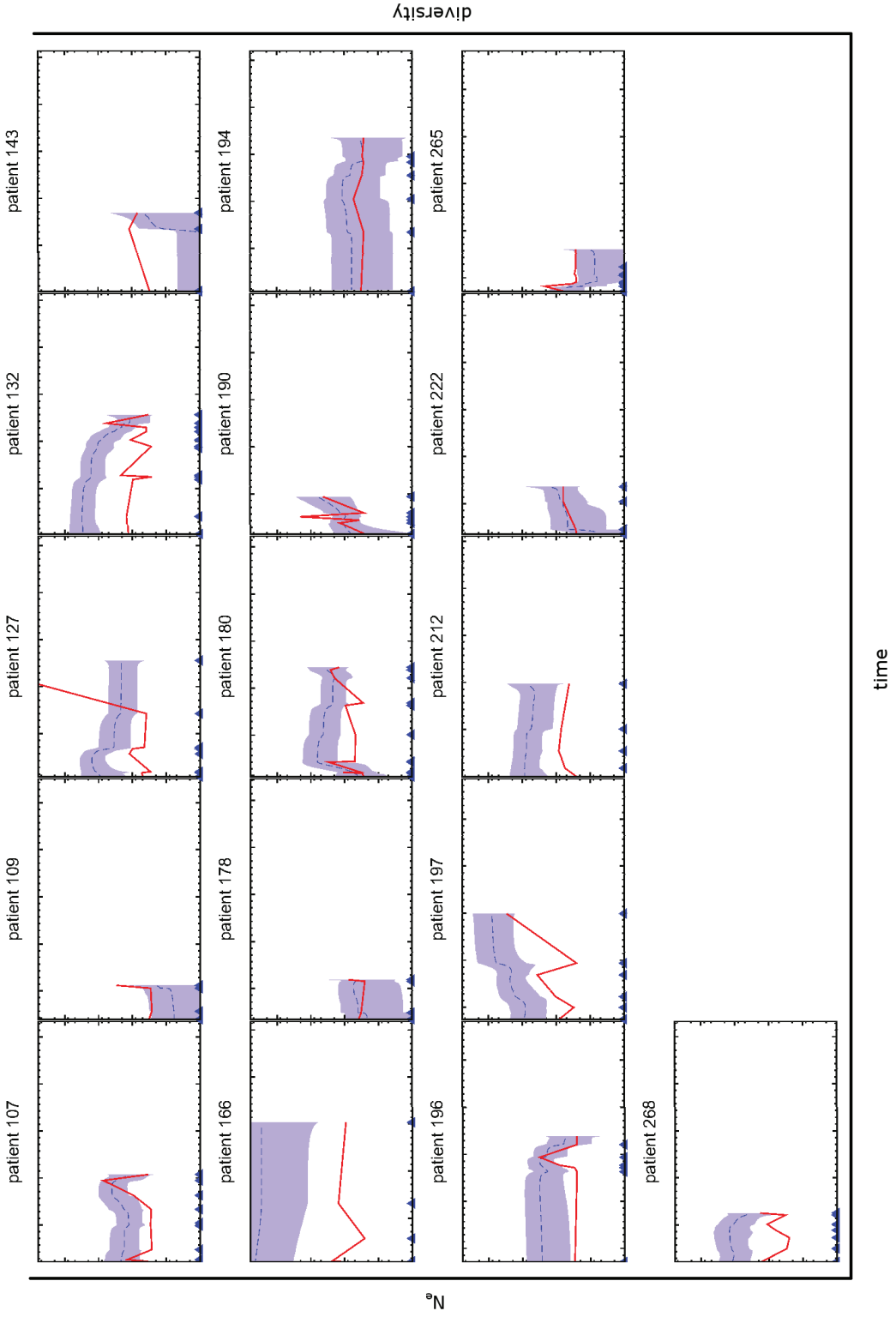


Figure 3.14.: **Course of N_e and diversity of 16 study patients**
The figure illustrates the course of the estimated effective population size N_e (blue line) with 95% confidence interval and the course of diversity (red line) of 16 study patients on an uniform time interval of seven years (starting at the time of the first patient specific measurement).

Table 3.5.: **Pearson correlation of N_e and diversity**
The table lists the Pearson correlation values r and the respective p-values for the association between N_e and the diversity. 'ID' is the unique patient identifier and the column 'obs.' gives the number of observations.

ID	obs.	r	p-value	ID	obs.	r	p-value
004	8	0.609	0.109	098	6	0.587	0.221
005	6	0.795	0.058	100	4	0.765	0.235
007	5	0.172	0.783	107	9	0.686	0.041
010	5	0.754	0.141	109	4	0.783	0.217
013	6	0.788	0.062	127	8	-0.510	0.197
024	4	-0.567	0.433	132	12	0.025	0.939
026	4	-0.545	0.455	143	4	0.827	0.173
032	6	0.629	0.181	166	5	-0.017	0.978
040	6	0.246	0.639	178	4	0.501	0.499
041	7	0.251	0.587	180	12	0.490	0.106
043	3	0.251	0.838	190	6	0.324	0.530
049	4	0.769	0.231	194	7	0.672	0.098
051	5	0.080	0.899	196	7	0.474	0.282
062	5	-0.247	0.689	197	6	0.509	0.302
068	4	0.552	0.448	212	5	-0.366	0.544
072	7	0.276	0.549	222	4	0.782	0.218
085	7	-0.261	0.572	265	9	0.518	0.153
097	4	0.901	0.099	268	7	-0.220	0.636

Diversity and divergence

From the illustration of the evolutionary course of the HIV-1 infection of Shankarappa *et al.* [138], we derived a hypothesis about the sequence evolution. The observation of the parallel increase of the viral diversity and the divergence in the initial phase of the infection and the simultaneous stabilisation of both measures in the third phase led us to the assumption of a positive linear correlation between the measures. Therefore we next analysed the Pearson correlation coefficient r of the viral diversity and the divergence.
We found a positive correlation for 24 of 36 patients, with a p-value <0.05 for eight data sets. The data on the diversity and the divergence showed an almost perfect positive linear correlation ($r >0.95$) for patients 049, 100, 127, and 222, confirmed by an significance level of 0.05 in all four cases. In contrast, 12 patients showed negative correlation values, being significant for patients 041 and 043.
A closer inspection of the data revealed a difficulty regarding the definition of the divergence of the sample at the first time point. The genetic distance of a sample to itself is zero, thus the divergence of the first sample of each patient was defined to be zero. Since the first sample of our patient data does not equal the founder sample, we excluded this time point from the analyses.
A repetitive analysis, excluding the first sample, resulted in 24 of 36 positive and 11 of 36 negative correlations. We could not calculate a correlation for patient 043, since only two data points remained in the respective data set. Three of the four data sets with initial positive correlations $r >0.95$ (patients 049, 127, and 222) increased to significant correlations $r >0.99$, while three additional data sets yielded significant positive correlations

r >0.95 (patient 040, 166, and 265). In summary, we only found a significant positive correlation for six patients (040, 049, 127, 166, 222, and 265), while none of the negative correlations was significant.
All Pearson correlation values for the association of the diversity and the divergence are given in Table 3.6. A calculation of the respective Spearman and Kendal rank correlation coefficients could not improve the results (data not shown).

Table 3.6.: **Pearson correlation of diversity and divergence**
The table lists the Pearson correlation values r and the respective p-values for the association of the diversity and the divergence. 'ID' is the unique patient identifier and the column 'o.' gives the number of observations. The numbers in parentheses resemble the correlation values calculated upon the exclusion of the first data point.

ID	o.	r		p-val.		ID	o.	r		p-val.	
004	8	-0.247	(-0.488)	0.556	(0.267)	098	6	0.648	(0.585)	0.164	(0.300)
005	6	0.576	(0.691)	0.231	(0.196)	100	4	0.977	(0.993)	0.023	(0.073)
007	5	0.075	(-0.255)	0.905	(0.745)	107	9	-0.082	(-0.143)	0.834	(0.735)
010	5	-0.342	(0.739)	0.573	(0.261)	109	4	0.794	(0.987)	0.206	(0.104)
013	6	0.524	(0.259)	0.285	(0.674)	127	8	0.994	(0.994)	<0.001	(<0.001)
024	4	0.725	(0.940)	0.275	(0.221)	132	12	-0.406	(-0.372)	0.191	(0.260)
026	4	-0.825	(0.571)	0.175	(0.613)	143	4	0.724	(0.884)	0.276	(0.309)
032	6	0.315	(0.212)	0.543	(0.733)	166	5	0.774	(0.988)	0.124	(0.012)
040	6	0.839	(0.976)	0.037	(0.004)	178	4	0.342	(0.397)	0.658	(0.740)
041	7	-0.756	(-0.572)	0.049	(0.235)	180	12	0.383	(0.324)	0.219	(0.330)
043	3	-0.998	(-)	0.038	(-)	190	6	-0.251	(-0.773)	0.632	(0.125)
049	4	0.997	(1.000)	0.003	(0.002)	194	7	-0.067	(-0.061)	0.887	(0.909)
051	5	0.887	(0.576)	0.045	(0.424)	196	7	0.016	(-0.163)	0.973	(0.758)
062	5	0.894	(0.926)	0.041	(0.074)	197	6	0.041	(-0.587)	0.939	(0.298)
068	4	-0.480	(0.955)	0.520	(0.193)	212	4	0.805	(0.856)	0.195	(0.346)
072	7	0.198	(0.011)	0.671	(0.983)	222	4	0.999	(0.999)	0.001	(0.026)
085	7	0.457	(0.505)	0.302	(0.307)	265	9	0.893	(0.994)	0.001	(<0.001)
097	4	-0.648	(-0.816)	0.352	(0.392)	268	7	-0.532	(-0.475)	0.219	(0.341)

An further inspection of the data of the 11 patients with a non-significant negative correlation revealed that 8 of these patients were clinically classified as CDC stage A3 or Bx patients, presumably situated in the intermediate phase of the infection. Therefore the observed negative correlations hint towards a discordant course of the diversity and the divergence, as described by Shankarappa *et al.* [138] for the intermediate phase of the infection. Due to the non-significant statistics, we could not confirm this idea.
In summary, the hypothesis about the evolutionary data could not be confirmed. Our data did not enable us to decide whether the evolutionary correlations of HAART treated patients coincided with the observations of Shankarappa *et al.* due to insufficient statistical support.

3.4.6. Analyses of the correlation between evolutionary and clinical parameters

So far, the analyses of the associations between the clinical as well as between the evolutionary parameters in general gave no statistical support to our hypothesis about the course of the disease in HAART treated patients. In the next step, we determined the correlations between the evolutionary data on the one hand and the clinical data on the other hand.

Since we hypothesised the immune system to drive the speed of evolution, we supposed to find an interrelation between the strength of the immune system and the speed of the sequence evolution. In evolutionary terms, we expected to find a positive association between the diversity and the number of $CD4^+$ cells as well as between the divergence and the number of $CD4^+$ cells.

The computation of the Pearson correlation coefficient did not support our idea. Though we found a weak trend for a positive correlation between the viral diversity and the number of $CD4^+$ cells (positive r for 22 and negative r for 12 of 36 patients), only two of the positive and one of the negative correlations were significant (p-value <0.05). In two cases, the calculation of the correlation failed due to missing measurements.

The analysis of the divergence and the number of $CD4^+$ cells yielded comparable results. 24 of 36 correlations were positive, 10 of 36 correlations were negative and for two patients, no correlation could be calculated. Only one positive and one negative correlation were significant (p-value <0.05). The complete list of the Pearson correlation coefficients of all patients is given in the supplementary material in Section A.1.1.

Thus, our idea of the immune system as driving force of the viral evolution could not be confirmed. Our data gave no statistical support for our hypothesis.

We further examined the association of the size of the viral population and the viral diversity. Formulating an idea of a *survival of the fastest* we expected a large viral population to be more homogeneous.

A negative Pearson correlation coefficient between the viral load and the diversity for 18 patients (p-value <0.05 for one patient) and a positive correlation coefficient for 14 patients (p-value <0.05 for one patient) confuted our idea. The calculation of a correlation for four data sets failed due to a lack of data. Thus, the data did not support our hypothesis of a dominance of the fastest replicating sequence.

A number of additional analyses also failed due to missing statistical support (e.g. the analysis of the association between the slope of the clinical and evolutionary measurements or an approach using a joined data set of all patients). Correlations with the putative way of infection, the medication as well as the occurrence of co-infections could not be observed.

Additional calculations using the Kendall and Spearman rank correlation coefficients encountered further difficulties, for example the existence of identical viral load measurements in the data sets (e.g. the value of 50 viral RNA copies per millilitre was multiply observed, indicating a viral load below the detection limit). Therefore, including a ranking of the data points into the calculation of the correlations did not improve the power of the analyses.

Summarising the data analysis, we can not conclude that the illustrations of Pantaleo *et al.* [51] and Shankarappa *et al.* [138] derived in times before HAART resemble the correlations

of patients under modern HIV combination therapy. Due to the weak statistical support, the analyses have to be repeated by a larger data set.

The course of the co-receptor tropism over time

Analysing Figures 3.11 and 3.12 we supposed that the course of disease of patients with successful HAART does not develop one-directional, but can be paused or reverted by successful therapy. Following this idea, we finally asked whether the therapy could invert an X4- or R5/X4-tropic population to use exclusively the R5 co-receptor. We hypothesised that the reversion of the course of infection entails the reversion of the co-receptor usage. Jensen and Shankarappa *et al.* [82] reported the observation of transient X4 strains in a co-receptor prediction study, describing two possible reasons for the back-switch. On the one hand, they observed a switch in the viral production through a shift in the latency and a repression of X4-tropic by viruses by R5-tropic viruses, while on the other hand, they monitored a possible sequence of back-mutations from X4-tropic sequences towards R5-tropic sequences.
Upon the analyses of the clinical course of the infection (compare Figures 3.11 and 3.12), we selected five patients for an extended co-receptor analysis: 007, 051, 062, 098, and 197. Using geno2pheno and FSSM for *in silico* co-receptor predictions, we found the following course of tropism:

- 007: $mixed \rightarrow X4 \rightarrow mixed \rightarrow X4 \rightarrow X4$
- 051: $R5 \rightarrow mixed \rightarrow mixed \rightarrow mixed \rightarrow mixed$
- 062: $X4 \rightarrow R5 \rightarrow X4 \rightarrow X4 \rightarrow X4$
- 098: $X4 \rightarrow mixed \rightarrow mixed \rightarrow X4 \rightarrow mixed \rightarrow mixed$
- 197: $R5 \rightarrow X4 \rightarrow X4 \rightarrow X4 \rightarrow X4 \rightarrow mixed$

This result was very surprising and did not support our idea in any of the five patients. In fact, the observed co-receptor evolution over time seemed to be almost random and provoked a further inspection of the sequence data.
An investigation of the sequencing method and the laboratory records as well as personal discussions with laboratory assistants and biologists revealed an important fact about the sequence data. For patients with viral load below the limit of detection it was in general not possible to extract viral RNA sequences. Therefore pro-viral DNA of the infected host cells was sequenced.
This knowledge led to the realisation that our evolutionary data did not show the recent population at the time of sampling, but constituted a view back into the past of the viral memory of the patient. We will discuss this observation extensively at the end of the chapter, since it has major impact on the presented evolutionary data.
With respect to the co-receptor analysis, we looked for a way to solve this problem and to gain information about the recent viral population at the time point of sampling. The previously described *in vitro* co-receptor prediction method of Binninger-Schinzel *et al.* [14] provided a solution. The researchers predicted the co-receptor usage of some of the study patients based on the isnoR5 cell line.
The isnoR5 reporter cells are highly sensitive for the existence of X4-tropic viruses, as the researchers could show in their study [14]. The cell culture exclusively facilitates the growth of HIV populations that contain at least some replicative-competent X4 virions. Mere R5-tropic populations are not able to infect the cell line due to the absence of the

CCR5 receptor on isnoR5 cells. To avoid misclassification and to confirm the existence of replicative competent HIV virions in the absence of viral growth, control experiments were performed.
Assessing the *in vitro* co-receptor prediction data of Binninger-Schinzel *et al.*, we gained experimental data for four of the five patients. The co-receptor usage of patient 062 was not examined *in vitro* and one sample of patient 197 could not be analysed. We found the following course of co-receptor usage:

- 007: $X4 \rightarrow X4 \rightarrow X4 \rightarrow R5 \rightarrow R5$
- 051: $X4 \rightarrow X4 \rightarrow X4 \rightarrow R5 \rightarrow R5$
- 098: $X4 \rightarrow X4 \rightarrow R5 \rightarrow R5 \rightarrow R5 \rightarrow R5$
- 197: $X4 \rightarrow X4 \rightarrow R5 \rightarrow R5 \rightarrow$ not possible $\rightarrow R5$

The *in vitro* co-receptor analyses finally confirmed our hypothesis. The experiments showed that all patients harboured at least some X4-tropic virions at the beginning of the study period. The observed early X4-tropic strains vanished under successful therapy, leaving mere R5-tropic samples during subsequent *in vitro* analyses. In the case of the patients 098 and 197, the X4-tropic strains were completely lost one year after the introduction of HAART, and in the case of the patients 007 and 051, the X4-tropic subpopulation dispersed about one year after a successful change in therapy.
For clarification we want to stress that the isnoR5 cell line is highly sensitive for the existence of X4-tropic virons in the viral population. Discussions with our cooperation partners revealed that the presence of at least one replicative competent X4 virion suffices to induce viral growth on the isnoR5 cell line. Therefore we can rule out the presence of X4-tropic strains in the case of an absence of viral growth. Due to this property, the method is not feasible to determine the co-receptor usage of the dominant viral strain, since as few as one X4-tropic minority strain induces viral growth on isnoR5 cells.
Thus it is save to say that at least a subpopulation of X4-tropic virions was present under phases of high viral load, and this subpopulation completely vanished from the active population under successful HAART.
Summarising this last part of the analysis, we can say that the emergence of an X4-tropic strain during the course of the infection is no final state of the population We observed that successful therapy can obliterate sequences with X4 phenotype from the population and facilitate the outgrowth of mere R5-tropic populations from populations harbouring at least some X4-tropic strains.

3.5. Discussion

In the first part of this section, we analysed correlations between clinical and evolutionary data of HAART treated patients. Unfortunately, our study data were neither feasible to confirm the well-known negative correlation between the number of $CD4^+$ cells and viral load, nor to find the expected positive correlation between the diversity and the divergence.
At most, the data of some patients suggested a difference in the timely course of the disease between HAART treated HIV-1 infections and infections in the early days of HIV therapy. We suppose that the linear course of the disease that was observed in therapy naïve patient ends with the initiation of HAART. For some patients, we found weak indications that the infection under HAART does not develop linearly from phase one towards phase three, but describes a steady forth-and-back course, in which either the virus or the therapy dominates the infection and determines the speed and the direction of evolution.
Due to the weak statistical support resulting from mainly non-significant correlations, the idea has to be confirmed by further studies.
In the last part of the section we analysed the co-receptor usage over time. Our results indicated that the emergence of X4-tropic strains during an HIV infection is no one-way street towards disease progression. An approach using the isnoR5 cell line [14] for *in vitro* co-receptor predictions showed that successful HAART suppressed an initially observed X4-tropic viral subpopulation and facilitated the subsequent outgrowth of an R5-only population in the analysed patients. Thus, X4-tropic strains can be displaced by R5-tropic strains upon successful therapy.
During the analyses we revealed a weakness of the applied sequencing method and therefore a problem of the sequence data. For patients with a viral load below the detection limit, it was not possible to obtain viral RNA from the blood samples. Therefore, pro-viral DNA sequences were extracted from the infected host cells. In consequence, the evolutionary data does not reflect the recent population at the day of sampling, but comprises a view into the history of the viral evolution of the patient.
Therefore, the analysis of a correlation between evolutionary and clinical measurements is not possible due to the insecure timing of the pro-viral sequences. This finding explains the observation of mainly non-significant and contradictory correlations of our data set.

3.6. Outlook

The presented study is based on longitudinal blood samples of HIV-infected patients that have been collected beginning in the year 1999. During the last ten to 15 years, the sensitivity of the clinical measurement methods as well as the sequencing techniques have widely improved. Due to the progress in the field of sequencing, a recent study using a deep sequencing approach could provide the whole viral population instead of a reduced sample that is mainly restricted to the dominant subpopulation. Also the pro-viral DNA could be examined in detail and could provide a complete history of the infection.
An interesting extension of the present work would be an analysis of the frozen longitudinal blood samples of some patients with modern sequencing methods, with special emphasis on a complete picture of both the active viral population and the viral reservoir. Such an analysis could presumably answer the question whether the viral reservoir contains a more complete image of the evolutionary pathways of the viral sequences by presenting sequences that are normally swept off the population by selection. Furthermore, we could gain insight into the time delay between the active and the latent viral population.
For the design of a new, consecutive study, special attention should be paid to sound and consistent sampling intervals, e.g. in the range of three to six months. As a consequence of successful HAART, the viral population evolves at very low levels of viral load. Due to the small population size, presumably less mutations accumulate among the individuals of the population. Thus, the follow-up times per patient should be increased.
Last but not least, the responsible physicians should seek to include patients in an early stage of the infection to gain a view of the complete viral evolution, since the present study was dominated by patients in CDC stage C3.

4. Fitness function and fitness landscape

4.1. Introduction

In the second part of the work, we used two large datasets of R5- and X4-tropic sequences to derive to fitness functions that enabled us to analyse the differences of the R5 and X4 fitness landscape.
In modern theoretical population genetics, the concept of *fitness landscapes* was introduced by Sewall Wright [164]. The basic idea of the concept is to use a suitable mathematical function, termed *fitness function*, to translate the genotype of an individual into a *fitness value*. The fitness value of an individual for example measures its capability to create offspring.
For means of visualisation, Wright used the method of fitness landscapes to reduce a multi-dimensional field of possible gene combinations of a population onto a two-dimensional landscape map, which he compared to a geological landscape with contour lines that are build around peaks and valleys in the landscape. Sequences with high fitness values tower above their neighbourhood and create peaks in the fitness landscape, similar to mountains. Sequences with low fitness create valleys that pass through the landscape, and sequences with zero replicative fitness are represented by holes in the fitness landscape.
In general, two types of fitness landscapes are used. Dynamic fitness landscapes [161] adapt to changing environmental conditions. These might be internal population properties, for example the competition of the individuals for limited resources (e.g. available host cells), as well as external influences, for example the immune pressure or the administration of drugs.
Dynamic landscapes are difficult to handle. It is not only challenging to determine all relevant environmental factors and to measure their individual influence on the fitness of the population, but also to formulate the correct mathematical representation and to compute the solution for a complete dynamic fitness landscape.
In contrast, static landscapes [57] are a good approximation to describe the fitness of individuals on a fixed population background, e.g. an environmental scenario that converged into a steady state. Furthermore, static fitness landscapes can be used to approximate the fast evolution of a population that evolves on a background of long-term changes.
In our work, we used static fitness landscapes to analyse the fitness landscapes of R5- and X4-tropic V3 loop sequences. We hypothesised that the cross section of data bank sequences we obtained from the Los Alamos database [100] represented a kind of steady state of the viral population. Dissociated from their genetic background and without knowledge of the immune status or the therapy scheme of the individual patients, the

database population served as a tool to describe an average V3 loop population. Therefore, static fitness landscapes are a suitable approximation for our approach.

4.2. Methods

In the first part of this section, evolutionary terms and mathematical methods are explained to describe and examine fitness landscapes and sequence networks. In the second part we introduce software and bioinformatics tools that we used to derive our fitness functions and to analyse the resulting fitness landscapes.

4.2.1. Additional Notation and Measures

Fitness

The term *fitness* describes a trait of an individual of a population in a defined environment. All individuals of a population contribute to the mean fitness of the population. The fitness of an individual measures for example its probability of survival, its growth rate, or its contribution to the next generation in general [57, 33, 47].
In the present work, we defined fitness as the replicative capacity of an individual, in terms of the probability to create offspring. The fitness is determined based on the amino acid sequence (i.e. the phenotype) of the individual.

Fitness landscape

Wright [164] described a *fitness landscape* as a "*representation of the field of gene combinations in two dimensions instead of many thousands.*" By translating a genotype into a fitness value, he created a rugged field of fitness peaks, surrounded by contours built of individuals with similar gene combinations, and separated by fitness valleys.
We used the concept of fitness landscapes to describe the replicative capacity of populations of V3 loop sequences. By a translation of the fitness landscape into a network of sequences, we analysed the underlying properties of the landscape and of the respective viral population.

Mutation

A *mutation* is a change in the nucleotide sequence of a gene that is passed on to descendent generations. Mutations are the basis for genetic variability. In HIV, the main source of mutations are replication errors introduced into the viral genome by the reverse transcriptase.
The majority of all mutations are *point mutations* [75, 156] that exchange one nucleotide of a codon triplet for another. Due to the redundancy of the genetic code, a point mutation can have three possible outcomes upon translation: 1) the amino acid sequence remains unchanged (i.e. a silent or synonymous mutation), 2) the amino acid is replaced by another amino acid (i.e. non-synonymous mutation), or 3) the amino acid is replaced by a stop codon. While silent mutations in general do not alter the protein function, non-synonymous mutations, might lead to sensible protein changes. The introduction of a stop codon often has a deleterious effect on the protein function, since it leads to a termination of the translation and results in a shortened protein [109, 132].
In addition to point mutations that conserve the open reading frame of the nucleotide sequence, *frameshift mutations* can occur. A frameshift mutation or *indel* inserts or deletes a number of n nucleotides from the nucleotide sequence (n not evenly divisible by three).

Indels change the reading frame by prolonging or shortening the sequence by lengths n different from the codon length. Starting at the position of the indel, a frameshift mutation alters all subsequent amino acids downstream of the indel upon translation of the shifted codons. A second indel further downstream of the first could shift the displaced reading frame back into the correct frame. Indels are often deleterious for the protein [156].

Selection

The idea of natural *selection* was formulated by Charles Darwin [33] in 1859: "*This preservation of favourable variations and the rejection of injurious variations, I call Natural Selection.*"
In general, a mutation that changes the amino acid sequence also alters the fitness of an individual. If the individual carrying the mutation has a fitness benefit (i.e. a higher fitness than the parent), it is more capable to reproduce. Thus, mutations that increase the replicative fitness have a selective advantage and the mutated individuals dominate the population. Individuals with a lower fitness create less offspring and are outcompeted by number.
Since individuals with higher fitness are favoured upon replication, the process of selection increases the mean fitness of the population.
Silent nucleotide mutations do not alter the amino acid sequence, thus they do not grant any fitness benefit or disadvantage. They are selectively neutral with respect to the replicative fitness and are not affected by selection.

Quasispecies model

The concept of *quasispecies* was introduced by Eigen [45] and put forward in subsequent publications; among others, also Schuster [47, 46, 136] contributed to the idea. A quasispecies represents a stationary distribution of self-replicating individuals with error-prone replication, e.g. a population of RNA or DNA molecules.
The quasispecies model describes the dynamic of an individual or sequence i in an infinite population by the following equation:

$$\dot{x}_i = \sum_{j=1}^{n} f_j x_j \mu_{ji} - \phi x_i \tag{4.1}$$

x_i is the concentration of i, resulting from a specified replicative fitness f_i and a mutation rate μ_{ji} (i.e. parent j producing offspring i). The model forms a population of n genetically diverse but closely related sequence variants, the so-called quasispecies. The term $\phi = \sum_i f_i x_i$ describes the average fitness of the population and is used as a regulator to keep the population size constant.

Mutual information

The *mutual information* (MI) is a non-parametric measure that is used to describe the information content one data set provides about another data set. The method is transferred from information theory and is also known under the older term of *transinformation*.

Derived from the Shannon entropy [139], the MI measures the expected information content of a message by estimating the content of information that can be derived from one random variable X about a second random variable Y, and vice versa.
Korber *et al.* [93] were the first to use the MI in biological context to examine the co-evolution of sequence positions by an analysis of the information content of two columns i and j of a multiple sequence alignment (MSA). Further approaches to reveal co-evolutionary effects were, among others, described by Gloor and Martin *et al.* [59, 103].
Using the probabilities $p(x_i)$ and $p(x'_j)$ of the amino acids x at position i and x' at position j, the MI of the columns i and j can be defined based on the Shannon entropy $H(i)$ and $H(j)$ that measures the amino acid variability at position i and j of the alignment [93]:

$$\begin{aligned} H(i) &= -\sum_{x=1}^{20} p(x_i) \log p(x_i) \\ H(j) &= -\sum_{x'=1}^{20} p(x'_j) \log p(x'_j) \\ H(i,j) &= -\sum_{x=1}^{20} \sum_{x'=1}^{20} p(x_i, x'_j) \log p(x_i, x'_j) \end{aligned} \tag{4.2}$$

$$MI(i,j) = H(i) + H(j) - H(i,j)$$

with $H(i,j)$ being the joint entropy of columns i and j of the alignment and $p(x_i, x'_j)$ describing the joint probability to find x_i in column i while observing x'_j in column j.
Alternatively, the mutual information can be expressed by [93]:

$$MI(i,j) = -\sum_{x=1}^{20} \sum_{x'=1}^{20} p(x_i, x'_j) \log(\frac{p(x_i, x'_j)}{p(x_i), p(x'_j)}) \tag{4.3}$$

The MI values of two columns i and j of an MSA are positive real numbers in the interval $[0, \log(20)]$ (20 being the size of the amino acid alphabet), describing the magnitude of the mutual dependence of columns i and j.
In our work, we applied the MI to identify potentially co-evolving positions in the MSAs of the V3 loop data sets. We used the R package BioPhysConnectoR [77] developed by Hoffgaard and Weil *et al.* which provides an easy-to-use tool to compute the MI of MSAs. The package comprises four alternative approaches to handle gaps in the MSA. The occurrence of gaps in alignments can cause problems upon the calculation and interpretation of the MI values, e.g. by overestimating the information content of the remaining amino acid symbols in columns with frequent gaps. The different methods provided by the BioPhysConnectoR package are described in the following paragraphs.

ORMI The **OR**riginal **MI** intuitively defines gaps as an additional 21th letter of the amino acid alphabet and calculates the MI values following Equation 4.3.

SUMI The **SU**bset **MI** excludes position pairs (x_i, y_i) with gaps and computes the MI of the reduced alignment subset following Equation 4.3, i.e. if one of the selected columns x or y contains a gap, the respective row i is omitted from the computation of the MI

of the columns x and y. By the exclusion of gaps, the SUMI maximises the information content carried by the amino acid symbols.

ESMI The **E**nhanced **S**ampling **MI** computes the mutual dependence of two columns i and j using their position specific amino acid probabilities (compare Equation 4.3). Gap characters are omitted upon the calculation of the respective probabilities, but in contrast to the SUMI approach, the gap frequency is regarded for the computation of the column probabilities.
The ESMI calculation can lead to artificial results. In columns x with many gap characters and reduced amino acid content, the gap handling process results in very small probabilities $p(x)$ in the denominator. Thus, the resulting ESMI values are very large, overestimating the impact of the low amino acid content in the observed column x.

DEMI The gap handling of the **D**elta **E**ntropy **MI** is related to the calculation of the SUMI. The method is based on the entropy definition of the MI (Equation 4.2) and calculates the entropy $H(i)$ and $H(j)$ of two columns i and j upon the exclusion of gap characters only in the actual column. In consequence, all rows i with gap characters in either position x_i or y_i are excluded from the computation of the joint entropy $H(i,j)$.
Due to this mode of calculation, the DEMI values can become negative, i.e. $H(i,j) \geq H(i) + H(j)$.

A detailed description and analysis of the MI and the different gap handling algorithms is presented by Weil *et al.* [158]. Based on their observations and on first evaluations of our data sets, we decided to use the SUMI method to treat gap characters. We omitted the ESMI from our analyses because the method can result in artificially large ESMI values in columns with many gaps and therefore might price positions with few information with high MI values. The DEMI was excluded because it can violate the general definition of the MI by taking negative values.

Normalisation and significance of MI values For MI normalisation, we also consulted the work of Weil *et al.* [158]. They described different methods to normalise the MI values and to determine the significance of the results.
The MI is capable to estimate the mutual dependence between two random variables, but the mere MI values can not be used to discriminate significant results from random effects of the underlying data or to detect so-called *finite size effects* caused by insufficiently large data sets.
To cope with this problem, Weil *et al.* developed a method to normalise the MI values, the so-called *shuffle null model* of the MI. The idea of the shuffle null model is to randomly shuffle the intra-column letters of an MSA to destroy the dependencies between the sequence positions, but to conserve the amino acid content in the column.
Since this shuffling method does not change the amino acid probabilities within a column, it enables an exact estimation of the random dependencies.
After the column shuffling, the MI of the shuffled MSA is calculated. The process of shuffling and the subsequent calculation of the MI is done n times. The average of the n runs gives an estimation of the random dependencies or basic noise of the observed columns.

Following the noise estimation, the concept of *Z-scores* is used to decide about the significance of the MI values. A Z-score counts the number of standard deviations a value differs from the sample mean. Weil *et al.* calculate the respective Z-scores of the MI for two columns i and j as follows:

$$Z(i,j) = \frac{MI(i,j) - \widetilde{MI}(i,j)}{\sqrt{var(\widetilde{MI}(i,j))}} \tag{4.4}$$

with $\widetilde{MI}(i,j)$ being the mean MI value of the shuffled columns i and j and $var(\widetilde{MI}(i,j))$ being the respective variance.
Significant and non-significant MI values can be discriminated by the a priori definition of a threshold. If the Z-score of the MI value of a specific pair of columns i and j surpasses the defined threshold, the respective MI value is significant, else it is discarded as noise.

Structural coupling

In addition to the computation of the MI, the R BioPhysConnectoR [77] package provides an algorithm to analyse the co-evolution of sequence and structure, termed *structural coupling*. The idea of the method is to combine the knowledge about the sequence conservation analysed by the MI with information on the structural conservation of the respective positions of the protein structure.
The structural coupling approach is based on the work of Brooks *et al.* [18] and Go *et al.* [60] who introduced a method to describe the spatial fluctuations of atoms in molecules. They define the coordinates of an atom in a protein structure as the mean position of the real atom fluctuations. Structural coupling combines this idea with an elastic network models approach described by Bahar *et al.* [6].
Using the structural coupling analysis, the spatial protein representation is reduced to the C_α atoms of the amino acid residues. Neighbouring C_α's of the original protein structure are connected by springs. The result of this reduction process is a simplified representation of the contacts of the original protein structure p_0, which can be represented as a matrix m_0. An entry $m_{0_{ij}}$ in this matrix represents the presence or absence of a contact of the C_α atoms i and j. During analysis, the contacts are switched off one after the other and the protein rearranges, resulting in a modified structure p_1 and the respective matrix representation m_1. In the next step, the spatial differences of the original and the altered structure are compared by the calculation of the Frobenius norm of the contact matrices m_0 and m_1. Large values of the Frobenius norm represent large deviations of the structures p_0 and p_1, and thus the causative contact in the original protein structure is supposed to be highly conserved.
The idea behind the coupling of the amino acid sequence and the protein structure states that an amino acid contact that both shows a high MI value and is essential to conserve the protein structure is biologically highly important in terms of protein function.
A detailed description of the theory was published in [77, 76, 158].

Cross correlation

By definition, MI values are always positive. Though the method is highly suitable to reveal the co-evolving positions, is fails to discriminate positive and negative epistatic effects of the coupled positions. To cope with this problem, we used the measure of *cross correlation* (CC). The method was recently described by Kouyos *et al.* [94] and Dahirel *et al.* [32]. Both groups used a formulation of the cross correlation to analyse epistatic interactions of an MSA of HIV sequences. While Kouyous *et al.* determined properties of the fitness landscape of the HIV protease and the partial reverse transcriptase gene, Dahirel *et al.* analysed the gag, nef, and rt gene to detect coupled regions, so-called *sectors of immunological vulnerability*, which they hypothesised to suppress the viral reproduction upon simultaneous targeting by anti-HIV drugs.
Dahirel *et al.* [32] formulated the following equation to compute the cross correlation c_{ij} of a pair of sequence positions i and j:

$$c_{ij} = \frac{< x_i, x_j > - < x_i >< x_j >}{\sqrt{(< x_i^2 > - < x_i >^2)(< x_j^2 > - < x_j >^2)}} \tag{4.5}$$

The values x_i and x_j indicate the presence or absence of the consensus amino acid at position i and j. Thus, $< x_i >$ and $< x_i >$ describe the counts of a mutation in position i and j in all sequences of the MSA, while $< x_i, x_j >$ describes the number of the mutual occurrences of mutations in both positions. A higher frequency of the mutual occurrence compared to the single mutations is indicative of positive epistasis, while a lower frequency is indicative of negative epistasis. The observed difference of the mutual and single mutations is divided by the variance of the observed mutations.
The evaluation of positions i and j is commutative, thus c_{ij} and c_{ji} are identical, and the resulting cross correlation matrix C is symmetric. Since all entries of C are real-valued, C can be diagonalized and all eigenvalues of C are real.

Validation and noise reduction of the cross correlation Kouyous *et al.* [94] validated the method using measurements of *in vitro* fitness of 70,081 virus samples of HIV-infected patients [74]. They compared the predicted values with the *in vitro* replicative capacity of HIV protease and reverse transcriptase mutants. They found a Spearman rank correlation of 0.33 ($p \leq 10^{-16}$) between the *in vitro* replicative fitness and the *in silco* results.
In contrast to the biological validation of Kouyos *et al.* [94], Dahirel *et al.* [32] focussed on a mathematical method to discriminate significant correlations from background noise. Furthermore, they addressed the problem of finite size effects due to the limited number of sequences.
In detail, Dahirel *et al.* determined the eigenvalues λ_C and the corresponding eigenvectors k_C of the cross correlation matrix C, which entries are defined by Equation 4.5. Subsequently, they randomised the underlying MSA by randomly shuffling each column of the sequence alignment independently, analogue to the shuffle null model Weil *et al.* [158] developed for the noise-reduction of the MI. The column shuffling removed the dependencies between two columns i and j in the original alignment, but conserved the amino acid composition and thus the phylogenetical and statistical properties within each column.

Subsequent to the shuffling, the authors computed the cross correlation matrix C_{shuff} based on Equation 4.5 and determined the eigenvalues λ_{shuff} and the eigenvectors k_{shuff} of the shuffled matrix. This process was done 1,000 times, resulting in 1,000 cross correlation matrices C_{shuff} and their corresponding eigenvalues and eigenvectors.
They determined the largest eigenvalue $\lambda_{shuff_{max}}$ across all 1,000 random matrices and reduced the original cross correlation matrix C to the eigenvalues $\lambda_{C_>} = \lambda_C > \lambda_{shuff_{max}}$ that are larger than the largest eigenvalue $\lambda_{shuff_{max}}$.
A matrix reconstruction using the significant eigenvalues $\lambda_{C_>}$ and corresponding eigenvectors $k_>$ yielded the matrix C_{clean}, which represented the original cross correlation matrix, cleared from noise and random correlations, but preserving the significant correlations of the underlying MSA.
So far, we found the approach to be closely related to the idea of principal component analysis (PCA) first presented by Pearson [115] and recently reviewed by Abdi *et al.* [1]. In their approach, Dahirel *et al.* stated further, that the contribution of the largest eigenvalue $\lambda_{C_{max}}$ of C is a result of the phylogenetic history of the related sequences in the MSA. They claimed that the contribution of $\lambda_{C_{max}}$ merely resembles phylogeny, but contains no relevant co-evolutionary information. Thus, they removed the largest eigenvalue $\lambda_{C_{max}}$ and corresponding eigenvector $k_{C_{max}}$ from C and recomputed the eigenspectrum of C, since they state that the largest eigenvalue $\lambda_{C_{max}}$ influences the remaining eigenvalues. Finally, they reconstruct $C_{clean_{phyl}}$ from the recomputed eigenvalues.
The details and the derivation of this argument are described in the supplementary material of the publication by Dahirel *et al.* [32].

Biological networks

Graph or network theory is in wide-spread use throughout different research areas, since it provides comfortable and well-studied methods to analyse properties of complex interrelated data. The basics were coined by Euler [48] formulating the Königsberger bridge problem in the eighteenth century. Since then, almost any kind of data was transferred onto network structures for analysis. Kauffman [88] was one of the first to present interaction networks to analyse aggregated nets of chemical reactions and the method is used until today to describe the dynamic interactions of cellular metabolic processes [83].
In our work, we used network theory to analyse the fitness landscapes of populations of V3 loop sequences. The analyses were performed using the Python *Networkx* [120, 69] package, which provides a number of fast routines for the calculation of basic network measures.

Basic network terms A simple graph G is mathematically described as a tuple (V, E), where V is the non-empty set of all vertices (or nodes) and $E \subseteq V \times V$ the set of unordered pairs of vertices, the edges of G. In a directed graph, the set of edges E consists of ordered pairs of vertices.
The number of all vertices $|V(G)|$ is called the order $n(G)$ of the graph, while the number of all edges $|E(G)|$ determines the size $e(G)$ of the graph.
Based on these definitions, we introduce a number of measures from graph theory that we used to analyse the networks of the V3 loop sequences. The definitions in this section

are based on Rosens book of Discrete Mathematics and its applications [129], if no other source is given.

Adjacency Two vertices v_i and v_j are called *adjacent* (or neighbours) in G if there exists an edge e_{ij} in G. The edge e_{ij} is called *incident* with the vertices v_i and v_j. While e_{ij} is said to connect v_i and v_j, v_i and v_j are endpoints of e_{ij}.
If a directed graph G contains the edge e_{ij}, v_i is adjacent to v_j and v_j is adjacent from v_j, while v_i is the initial and v_j the terminal vertex of e_{ij}.
Two edges e_{xi} and e_{ix} of an undirected graph are adjacent or consecutive if they are incident with the same vertex v_i. In a directed graph, two edges or arrows e_{ij} and e_{jk} are called adjacent or consecutive if the terminal vertex of the first edge e_{ij} is the start vertex of the succeeding second edge e_{jk}.

Paths A *path* is a sequence of adjacent edges in a graph. The length n of the path in an unweighted graph is the number of edges traversed. A path of length n connecting v_0 and v_n can be described by listing the vertex sequence $v_0, v_1, ..., v_n$.
The *shortest path* between two nodes v_i and v_j is the path with the least number of edges starting in v_i and ending in v_j or vice versa. In a graph G, the average shortest path $l(G)$ is the average of all shortest paths between any possible pair of vertices in the graph.

Node degree In an undirected graph, the *degree* $k_G(v_i)$ of a vertex v_i is the number of all vertices incident with v_i in G.
In a directed graph, the degree of a vertex v_i is discriminated into an in-degree $k_G^-(v_i)$ and an out-degree $k_G^+(v_i)$. While the in-degree is the number of all edges with v_i as final vertex, the out-degree is the number of all edges with v_i as initial vertex. Incoming edges on a vertex of a directed graph G are represented by an arrow head, while outgoing edges are represented as an arrow tail.
To describe a network or to compare the structure of two networks, we can determine the average, minimal, and maximal (in- or out-)degree of all vertices, as well as the degree distribution of all vertices of the networks.

$$\begin{aligned} \text{in-degree } k_G^+(v_i) &= \sum_{j=1}^{n(G)} e_{ji} \\ \text{out-degree } k_G^-(v_i) &= \sum_{j=1}^{n(G)} e_{ij} \\ \text{degree } k_G(v_i) &= k_G^+(v_i) + k_G^-(v_i) \end{aligned} \tag{4.6}$$

with e_{ij} being an edge directed from vertex i towards vertex j.

Connectedness Two vertices v_i and v_j are called *connected*, if there exists a path $v_i, ..., v_j$ between them in the graph.
A graph is called connected, if such a path can be found for every pair of vertices. A directed graph is *strongly connected* if there is a path from v_i to v_j and from v_j to v_i for each pair v_i and v_j of the graph, and it is *weakly connected*, if the underlying undirected graph is connected.

If the graph is not connected, then it can be divided into at least two connected components, which are disjoint and disjunct connected subgraphs of the original graph.
A vertex v_i of a graph G is isolated, if the degree $k_G(v_i) = 0$.

Reachability In a graph G, a vertex v_j is *reachable* from v_i in G, if there exists a path from v_i to v_j. Each vertex is reachable from any other vertex, if an undirected graph is connected or a directed graph is strongly connected.

Betweenness The *betweenness* b_{v_i}, or betweenness centrality, of a vertex v_i in a graph G is calculated as the fraction of the number of all shortest paths p_{v_j,v_i,v_k} in the network that pass through v_i.

$$b_{v_i} = \sum_{j,k} \frac{p_{v_j,v_i,v_k}}{p_{v_j,v_k}} \tag{4.7}$$

for $v_j \neq v_i$, $v_i \neq v_k$, and $v_k \neq v_j$.
The betweenness of an edge e_{ij} can be defined correspondingly as the fraction of all shortest paths that contain e_{ij}.

Closeness The *closeness* c_{v_i} of a vertex v_i is the reciprocal of the average path length of all shortest paths to all other vertices v_j of a graph G. If G consists of n vertices, the closeness c_{v_i} of a vertex v_i is computed as follows:

$$c_{v_i} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} p_{v_i,v_j}} \tag{4.8}$$

for $v_i \neq v_j$.

4.3. Data

The second part of the present work focusses on the determination of two fitness functions to describe the replicative fitness of R5and X4-tropic sequences of the V3 loop of HIV-1. We developed the multi-step approach presented in Figure 4.1 to derive the fitness functions. The detailed description of each task is given in the following sections.

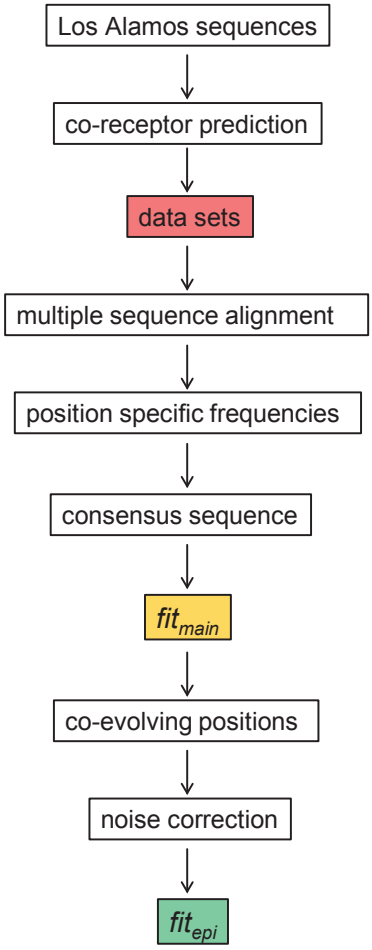


Figure 4.1.: **Sequence of steps to derive main fitness and epistatic interactions**
The figure presents an overview over the sequence of tasks to derive the fitness functions. We separated the Los Alamos V3 loop sequences into an R5 and X4 data set and calculated the respective MSAs. Based on the MSAs, we derived position specific aa frequencies and consensus sequences. The results were used to determine the main and the epistatic contributions of the fitness functions.

4.3.1. Sequence collection

In the first step, we retrieved all HIV-1 subtype B V3 loop sequences from the Los Alamos HIV database [100] that were accessible until 10.05.2012. We used the following parameters for the database search:

- Virus: HIV-1
- Subtype: B
- Genomic Region: V3

All other categories were set to default.
Nucleotide sequences of the V3 loop were the largest group of all HIV-1 sequences in the data base, providing 81,651 loop sequences. For reasons we could not clarify, the database search labelled 1,963 sequences as being problematic. We removed those from the search result and requested the remaining 79,688 nucleotide sequences for download.
44,656 (55%) of the sequences were extracted from blood samples of patients in the United States, whereas only 430 sequences from Germany were contained in the data set. Detailed statistics of the sampling country of the sequences are presented in Figure 4.2.

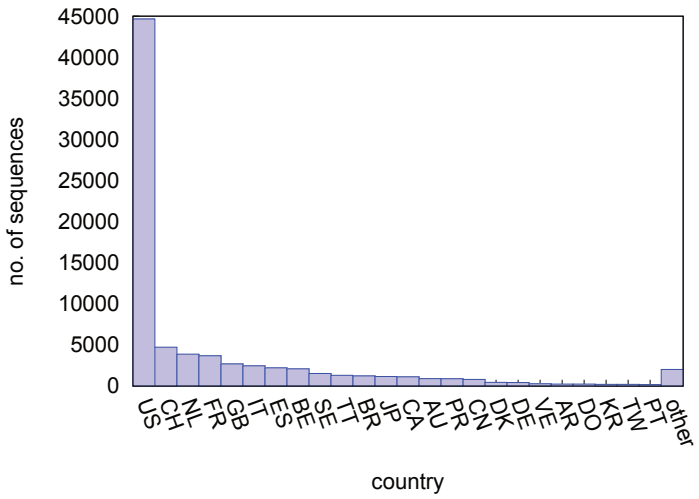


Figure 4.2.: **Statistics of Los Alamos V3 loop sequences: country of sample** The figure distinguishes the sequences based on the sampling country.
(US: United States of America, CH: Switzerland, NL: Netherlands, FR: France, GB: Great Britain, IT: Italy, ES: Spain, BE: Belgium, SE: Sweden, TT: Trinidad and Tobago, BR: Brazil, JP: Japan, CA: Canada, AU: Australia, PR: Puerto Rico, CN: China, DK: Denmark, DE: Germany, VE: Bolivarian Republic of Venezuela, AR: Argentina, DO: Dominican Republic, KR: Republic of Korea, TW: Taiwan, Province of China, PT: Portugal)

The sequences were collected in the years from 1978 to 2011, for 24,415 sequences the year of sampling was missing. Figure 4.3 describes the distribution of sequences over time.

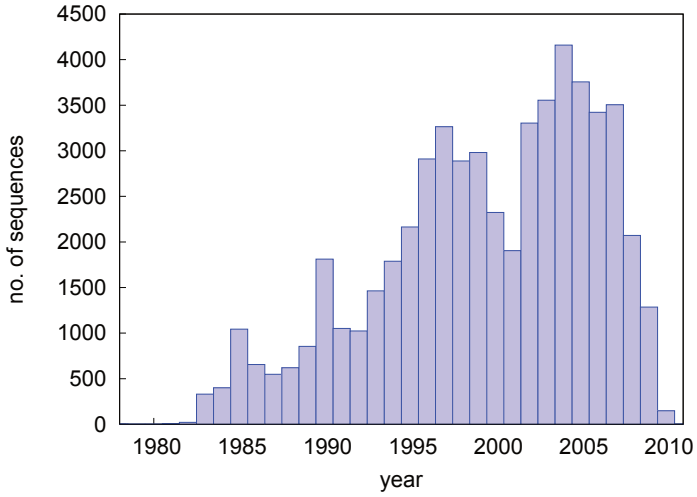


Figure 4.3.: **Statistics of Los Alamos V3 loop sequences: year of sample** The figure distinguishes the sequences based on the sampling year.

For 2,768 patients, at least one sequence was available. Of those, 2,096 patients contributed at least five sequences to the data set. For 7,324 sequences, a patient identification code was missing. More detailed statistics on the sequence distribution per patient are depicted in Figure 4.4.

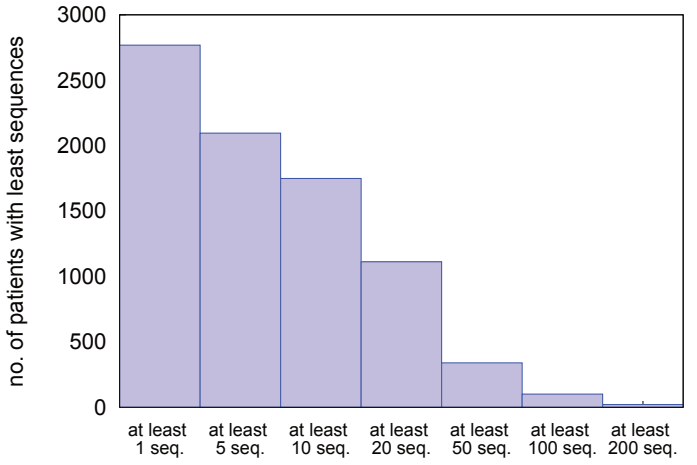


Figure 4.4.: **Statistics of Los Alamos V3 loop sequences: sequences per patient** The figure describes how many patients with at least 1/5/10/20/50/100/200 sequences are included in the data set.

In addition to the sampling country, the sampling year and the unique patient identifier, the database provided a co-receptor phenotype for 4,160 of the 79,688 sequences. Of those, 3,301 sequences were annotated with R5 label, 199 with X4 label, and 500 sequences with both R5 and X4. The remaining 160 sequences were annotated with other known HIV co-receptors.

4.3.2. Separation into R5 and X4 subset

The Los Alamos sequence data served as a basis to determine fitness functions to describe the replicative capacity of R5- and X4-tropic sequences. Since a high quality of the data set is indispensable for the derivation of the fitness functions, we performed several analyses of the sequence data set.

R5 and X4 sequence subsets

Starting from the 79,688 V3 loop sequences, we first performed an *in silico* co-receptor prediction to separate the sequences into an R5 and an X4 sequence subset. Due to the upload restriction of geno2pheno [97] to ≤ 50 sequences, we used FSSM [82, 81, 118] for the first phenotypic predictions.
Following the analyses of Poveda *et al.* [118] and Low *et al.* [101], we used a prediction cutoff value of -8.12: sequences with scores ≤ -8.12 were classified as R5-tropic and sequences with scores > -8.12 were classified as X4- or dual-tropic.
Upon prediction, FSSM highlighted problematic sequences, either due to alignment errors or due to prediction values that lay outside the 95% CI of the distribution. We excluded these sequences from further analyses. FSSM provided valid classifications of 67,770 sequences of 6,403 patients: 47,036 sequences of 4,410 patients were predicted to be R5-tropic and 20,734 sequences of 1,993 patients to be X4-tropic.

Intra-sample duplicates Intra-sample duplicates are known to be rather a result of the sequencing technique used in former days than a result of the replicative fitness of the sequence [146], therefore we next removed sequence duplicates within a patient specific sample from the data set. In consequence, we had to remove also sequences with missing patient identifier. This process removed about two thirds of both the R5 and X4 sequences from the data set, reducing the R5 data set to 15,626 sequences (of 4,410 patients) and the X4 data set to 7,053 sequences (of 1,993 patients). The rate of about one third of all sequences being unique could be confirmed by an analysis of the sequences from the clinical study presented in part one of this work (compare Section 3.3), of which we know the exact sequencing protocol that was used.
It has to be clarified that the reduced R5 and X4 data sets still span the same sequence space as the data set including the intra-sample duplicates, but the data sets miss the information of the frequency of the sequences within a sample. Since the sequence counts are a determinant of the replicative fitness, the exclusion of intra-sample duplicates reduced the sequencing bias at cost of a loss of fitness information. Despite, we expect to find sequences with high fitness not only within one sample but within multiple samples, of the same or of different patients. Therefore the loss of information on replicative fitness was rated less severe in comparison to the estimated sequencing bias.
To evaluate the effect of the exclusion on the position specific aa counts of our data sets, we performed a comparison of the R5 and X4 data set before and after exclusion. Using ClustalW [149], we created MSAs of the reduced and the full R5 and X4 data sets. During the alignment process we observed that no gap columns were introduced into the R5 alignment, it maintained the length of 35 aa. In the case of the X4 data set, eight gap-dominated columns were introduced into the MSA, due to insertions found

in single V3 loop sequences of the X4 data subset. Varying alignment parameters did not further reduce the number of gaps, since some of the unaligned X4-tropic sequences already contained 43 instead of 35 amino acids. Length polymorphisms of X4 sequences have already been observed in earlier studies (e.g. [137]).
Due to a better comparability of the R5 and X4 sequences, we reduced the X4 alignment to a sequence length of 35. Therefore, we aligned the complete MSA of the X4 data set to four V3 template sequences, one being the standard HIV-1 B consensus sequence HXB2 [70] and the others being the V3 loop sequences of the .pdb structures presented in Section 4.4.4. As expected, this step introduced eight gaps into the four V3 template sequences. The supernumerary gap columns of the MSA were marked by the introduction of gaps into the four template sequences. By the removal of those gap-dominated columns, we trimmed the X4 MSA to a sequence length of 35 aa.
After the computation of the MSAs, we created heat maps of the differences of the position specific aa counts (left illustration in Figure 4.5). In the case of the R5 MSA, the exclusion of intra-sample duplicates resulted in small deviations of the position specific conservation values. The maximal change of the aa counts was as small as 2.06% and was observed in position 29 for aspartic acid (D). Though the frequency deviations were more pronounced in the X4 MSA, most changes affected the frequency of position specific gaps, a fact that is also reflected in the maximal deviation of 13.50% for the deviation of the gap count in position 24 (compare right image in Figure 4.5).

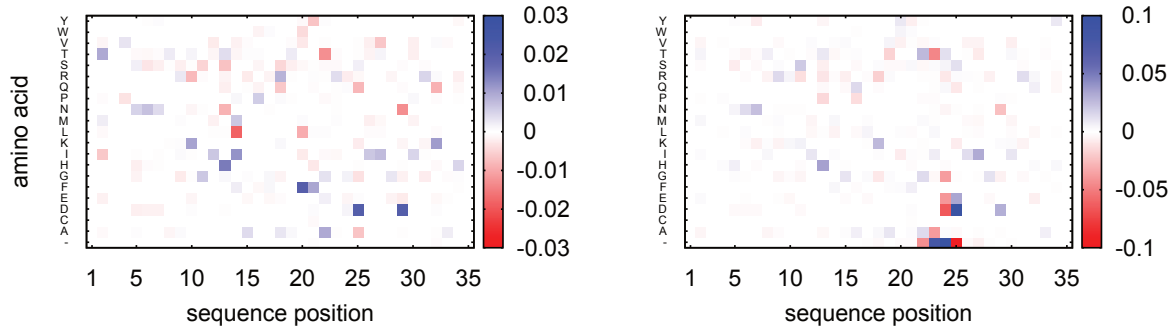


Figure 4.5.: **Differences in position specific amino acid counts of the full and reduced data set**
The heatmaps illustrate the rate of the frequency difference between the full minus the reduced data set.
left: The exclusion of intra-sample duplicates from the R5 data set resulted in small deviations of the position specific aa frequencies.
right: In the X4 data set, the exclusion of intra-sample duplicates resulted in more pronounced deviations of the position specific aa frequencies, but most changes affected the counts of the position specific gaps.

US versus non-US samples Following the exclusion of intra-sample duplicates, we performed additional subset analyses to gain deeper insights into the structure of our data sets. Using Perl scripts we selected subsets based on different parameters, for example sampling country, sampling year, or the number of patient specific sequences in the data set.
We already revealed that 55% of all sequences originate from the US. Therefore we analysed the differences between the US (7,877 R5 and 3,596 X4) and non-US sequences (7,739 R5 and 3,450 X4).
For the R5 predicted sequences, we could hardly find any differences in the position specific aa preferences of the US and non-US data set (compare left image in Figure 4.6). The maximal frequency deviation of 5.46% was observed in positions 13 for amino acid asparagine (N).

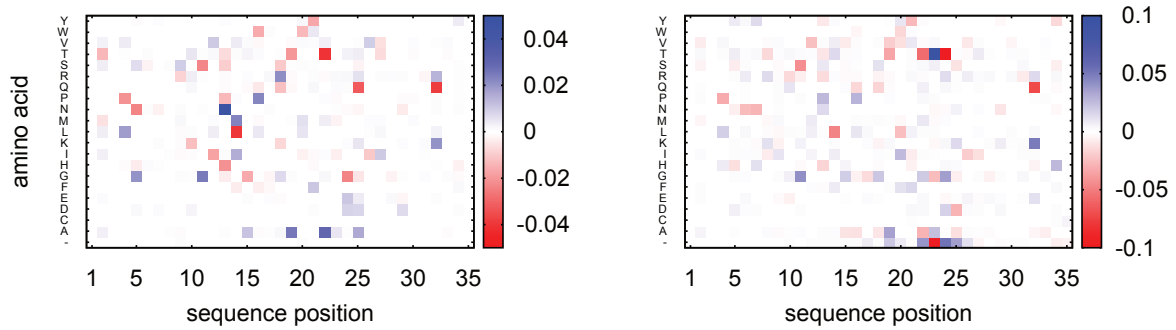


Figure 4.6.: **Differences in position specific amino acid counts of the US and non-US data subset**
The heatmaps illustrate the rate of the frequency differences of the US minus the non-US data subset.
left: In the R5 data set, the maximal deviation of the position specific aa counts was 0.0546 (5.46%), observed in position 13 for asparagine.
right: Positions 13, 14, and 23 of the X4 MSA showed deviations of the position specific aa counts of the US data set compared to the non-US data set, with a maximal deviation of 19.19% for threonine (T) in position 23 in the US subset.

In the case of the X4 data set, the analysis showed increased deviations between the US and non-US sequences (compare right image in Figure 4.6). The largest deviation occurred in position 23, with a frequency shift of 19.19% between the amount of threonine (T) and alanine (A). The other positions showed only marginal differences in the position specific aa probabilities.
For an additional representation of the differences, we created sequence logos of the X4-tropic US and non-US sequences compare Figure 4.7). The deviations of the sequence logos did not alter the aa dominance of the consensus sequence, and only marginally modified the position specific aa probabilities. The largest deviations were found for the less frequent aa in positions 13 (`S,P,R,N`) and 24 (`T,E,D,R`), but histidine (`H`) in position 13 and glycine (`G`) in position 24 were still dominant in both data sets. The differences in the other sequence positions were even smaller, as we have already seen in Figure 4.6.

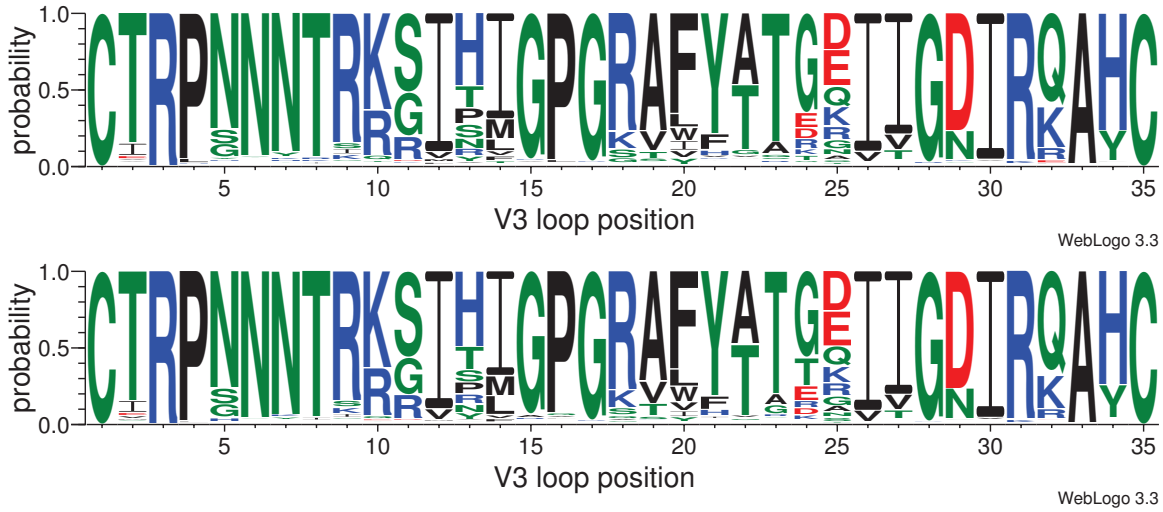


Figure 4.7.: **Sequence logo of X4 predicted sequences of US and non-US subsets** (without duplicates)
The hight of the letters (y-axis) describes the position specific aa probability. The colours illustrate chemical properties: blue: basic (`K,R,H`), red: acidic (`D,E`), green: polar (`C,G,N,Q,S,T,Y`), and black: non-polar/hydrophobic (`A,F,I,L,M,P,V,W`).
The upper sequence logo describes the position specific aa probability of the 3,596 X4 sequences from US, while the lower sequence logo describes the position specific aa probability of the 3,450 non-US X4 sequences. The sequence logos revealed no differences in the dominant position specific aa.

Since the position specific aa counts differed in only a few positions between the US and non-US data set, the respective consensus sequences were not altered by the data separation. Thus, we did not further consider a separation of our sequences into an US and a non-US data set.

R5-only samples versus mixed-tropic samples In another subanalysis, we asked for differences between sequences of patients in early and progressed state of disease. Therefore we discriminated patients with exclusively R5-tropic sequences from patients with X4-tropic sequences. We created an R5 sequence subset consisting solely of sequences of patients without X4 sequences (R5-only). A second data set contained sequences of those patients with both X4 and R5 predicted sequences, but without sequences from R5-only patients (mixed X4/R5). A subset of X4-only patients is difficult to gather, since X4 patients in general also harbour R5 sequences. Thus, we could only perform this weak separation to get an X4 subset of a reliable size.
The comparison of the sequences of R5-only patients and of mixed X4/R5 patients showed no significant differences. Similar to the previous analyses, we mainly observed a slight tendency towards a higher sequence conservation within the R5-only subset, in parallel with a marginal decrease of the conservation within the mixed X4/R5 subset. The differences were even smaller than those observed for the US and non-US sequences (data not shown). Due to the marginal differences between the analysed subsets, we regarded the observed differences as minor random data fluctuations. Therefore we decided to reduce the sequencing bias by the exclusion of the intra-sample duplicates, but to perform no further

subset separation.

Consistent FSSM and geno2pheno co-receptor prediction A correct separation of R5- and X4-tropic sequences is essential for our data sets. Therefore, we used geno2pheno [97] for confirmative predictions of the FSSM [82, 81, 118] co-receptor classification. We first compared the predictions of FSSM (using the HIV-1 subtype B X4/R5 scoring matrix and a threshold of -8.12, compare Section 3.2.2) with the geno2pheno predictions (using a threshold for the FPR of 10%) for a subset of randomly selected V3 loop sequences to get an idea of the concordance of both tools.
Based on geno2pheno predictions of the Los Alamos data set, we selected 580 random R5 and 580 random X4 predicted sequences. Using FSSM, we performed a subsequent prediction of those 1.060 sequences.
Three sequences could not be classified by FSSM due to problems in the alignment process. Of the remaining 1.057 test sequences, FSSM gave concordant predictions for 477 of the 580 R5 sequences (82.24%) and for 399 of the 580 X4 sequences (68.79%). In summary, 281 R5 and X4 predictions (24.29%) were discordant between the tools. The rate of discordant predictions of the random test set was confirmed during the analyses of the complete data set.
Based on discussions with A. Thielen and an additional search in the published literature, we could address this problem mainly to the observed length polymorphism of the X4 data. This phenomenon was earlier analysed in a study by Poveda *et al.* [137]. For sequences of uniform length of 35, the researchers found a concordance up to 88% between the position-specific scoring matrix and geno2pheno algorithms. For V3 loop sequences of other lengths, the concordance was as low as 55%. These numbers were confirmed in our analysis.
The discordance is further influence by the design of geno2pheno, which is aimed to maximize the X4 sensitivity, and by deviations in the underlying consensus sequences of both prediction tools.
In consequence, a restriction to those sequences consistently predicted by FSSM and geno2pheno reduced the 15,626 R5 sequences to 14,008 (89.65%) and the 7,053 X4 sequences to 3,409 (48.33%) sequences.

Final R5 and X4 data set In summary, the final data sets consisted of 14,008 R5 and 3,409 X4 sequences, consistently classified by FSSM and geno2pheno to minimize the prediction bias, and without intra-sample duplicates to reduce the frequency bias resulting from the sequencing method. Based on four template sequences, the length of the X4 sequences was trimmed to a length of 35 aa (as described in Section 4.3.2). We did not use any other restriction.

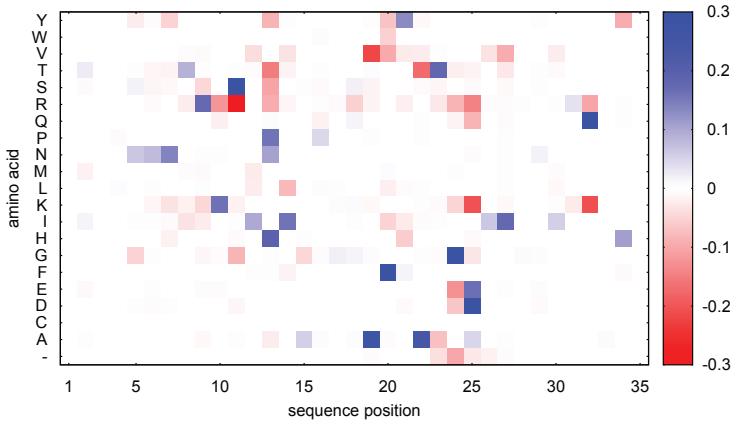


Figure 4.8.: **Differences in sequence conservation of R5 and X4 data set**
The heatmap illustrates the position specific aa differences between the R5 data set (14,008 sequences) and the X4 data set (3,409 sequences). The colours represent the position specific aa probabilities of the R5 data set minus the position specific aa probabilities of the X4 data set. Most differences accumulate close to the co-receptor determining positions 11 and 25.

We analysed the differences between the final R5 and X4 data set. Figure 4.8 illustrates the deviations in the position specific aa probabilities. Differences in the R5 and X4 sequence conservation are especially pronounced in positions 9 to 11 (variation in frequency and conservation of `R`, `K`, and `S`), 19 to 27 (variation in frequency and conservation of `A`, `G`, `F`, `T`, `I`, and `V`), and 32 (`K` and `Q`).
In detail, positions 10 and 11 of the X4 data set show a higher conservation of arginine (`R`), a positively charged amino acid. Furthermore, the content of lysine (`K`), also positively charged, is slightly increased in positions 25 and 32 of the X4 data set. In contrast, the neutral amino acid glycine (`G`) is strongly increased in position 24 of the R5 data set. This reflects also the maximal deviation of 46.44% between the R5 and X4 data set.
It is well-known from literature [165, 7, 22], that an overall positive net charge of the V3 loop is essential to utilise CXCR4 as co-receptor. Therefore these observations gave evidence to evaluate the quality of our data set.
In summary, most position specific aa differences accumulate in the positions flanking position 11 and 25, which are the positions that are known to be causative for the co-receptor tropism. Furthermore, the subsets show a higher position specific aa conservation of the R5-tropic sequences. The observation that the X4 sequences are more diverse was already described in previous studies (e.g. [15]) and confirms the quality of our data sets.
The sequence logos of the final R5 and X4 data set are illustrated in Figure 4.9. Based on the MSAs of the final data sets, we derived the following aa consensus sequences:
R5: `CTRPNNNTRK SIHIGPGRAF YATGDIIGDI RQAHC`
X4: `CTRPNNNTRK SIHIGPGRAF YTTGKIIGDI RQAHC`
The consensus sequences differ in positions 22 (R5: `A`, X4: `T`) and 25 (R5: `D`, X4: `K`), again confirming the preference of positive charges in the co-receptor determining positions of the X4 sequences. An *in silico* co-receptor prediction classified the R5 consensus sequence

as R5-tropic and the X4 consensus sequence as X4-tropic.

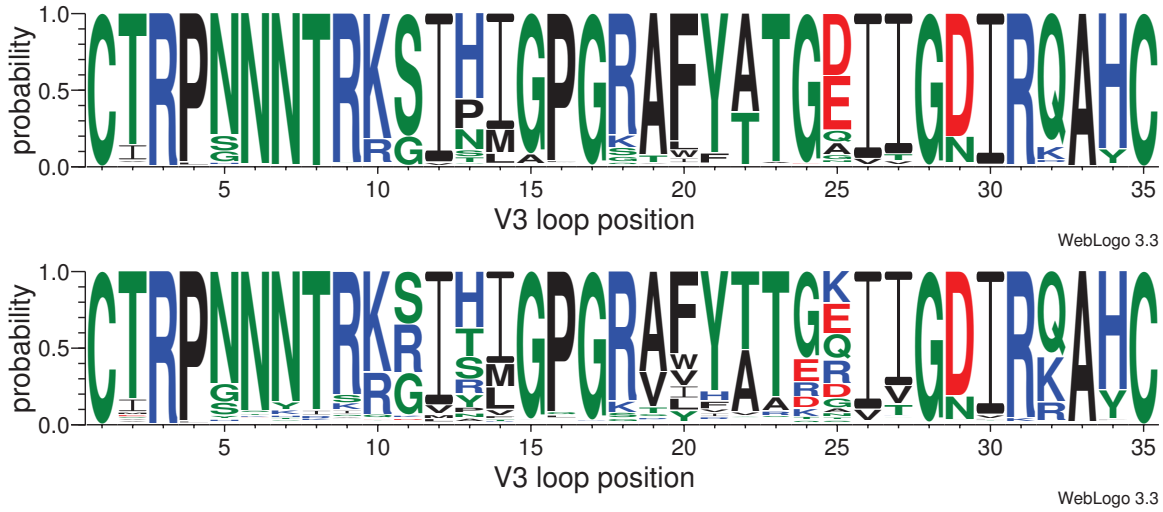


Figure 4.9.: **Amino acid sequence logo of R5 and X4 subset**
The figure illustrates the position specific aa probabilities of
(top) 14,008 R5-tropic sequences, and
(bottom) 3,409 X4-tropic sequences (adjusted to a length of 35 aa).
The consensus sequences differ in positions 22 (R5: `A`, X4: `T`) and 25 (R5: `D`, X4: `K`).
The hight of the letters (y-axis) describes the position specific aa probability. The colours illustrate chemical properties: blue: basic (`K,R,H`), red: acidic (`D,E`), green: polar (`C,G,N,Q,S,T,Y`), and black: non-polar/hydrophobic (`A,F,I,L,M,P,V,W`).

In addition to the amino acid consensus sequences, also the corresponding nucleotide (nt) consensus sequences were determined. Using the original nt sequences as templates, the R5 and X4 aa MSA were back-translated into nt sequences. The dominant nucleotide letter at each sequence position determined the nt consensus sequence. Due to the length adaptation of the X4 MSA, the nucleotide consensus sequences are both of length 105. The sequence logos in Figure 4.10 illustrate the position specific nt probabilities of the R5 and X4 nt MSA.

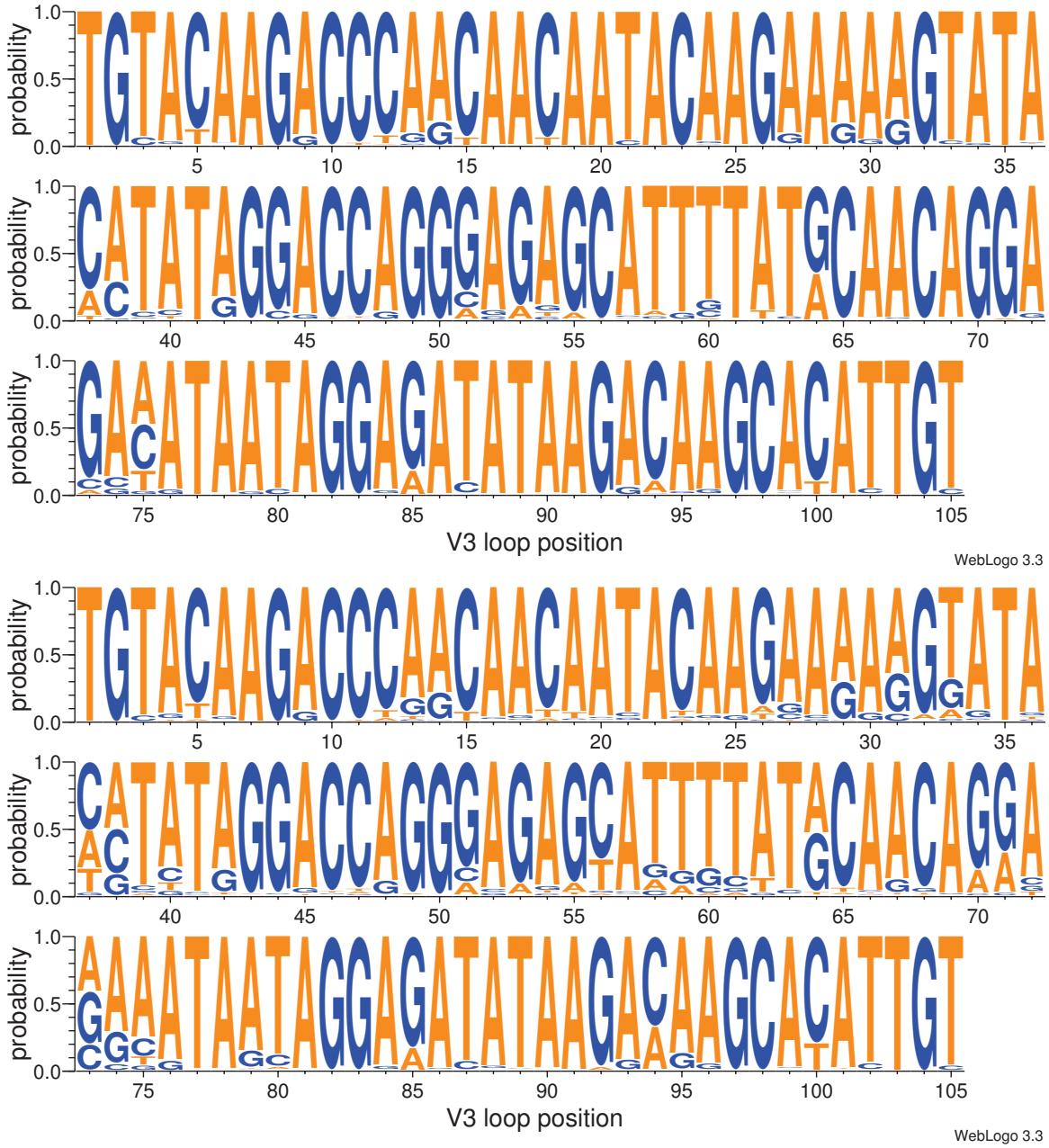


Figure 4.10.: **Nucleotide sequence logo of R5 and X4 subset**
The figure illustrates the position specific nt probabilities of
(top) 14,008 R5-tropic sequences, and
(bottom) 3,409 X4-tropic sequences (adjusted to a length of 105 nt).
The consensus sequences differ in positions 22 (R5: `A`, X4: `T`) and 25 (R5: `D`, X4: `K`).
The hight of the letters (y-axis) describes the position specific amino acid probability. The colours illustrate the nt `A` and `T` in orange and `C` and `G` in blue.

We derived the following R5 nucleotide consensus sequence:

```
TGT ACA AGA CCC AAC AAC AAT ACA AGA AAA AGT ATA CAT ATA GGA CCA GGG AGA
GCA TTT TAT GCA ACA GGA GAA ATA ATA GGA GAT ATA AGA CAA GCA CAT TGT
```

and the respective X4 consensus sequence from the nt MSAs:

```
TGT ACA AGA CCC AAC AAC AAT ACA AGA AAA AGT ATA CAT ATA GGA CCA GGG AGA
```

GCA TTT TAT ACA ACA GGA AAA ATA ATA GGA GAT ATA AGA CAA GCA CAT TGT

Analogue to the aa consensus sequences, the R5 and X4 nt consensus sequences differed only in codons 22 (R5: GCA, X4: ACA, nt positions 64 to 66) and 25 (R5: GAA, X4: AAA, nt positions 73 to 75).

During this analysis, we observed a peculiarity about the nt and aa consensus sequence of the R5 data set. The aa translation of the nt consensus sequence of R5 did not exactly match the aa consensus sequence. Codon 25 (GAA) translates into glutamic acid (E), while we found aspartic acid (D) in the respective aa position 25, which would be encoded by codon GAC.

An examination of our data showed that this peculiarity was not an error in our consensus sequences, but a result of the most weakly conserved nucleotide position 75, the third position in the differing codon. In position 75 of the nucleotide MSA, the nucleotides adenine (A) and cytosine (C) occurred with almost similar probability, but with a slight preference towards nucleotide A, while amino acid D was slightly more frequent in the corresponding position 25 of the aa sequence.

4.4. Results

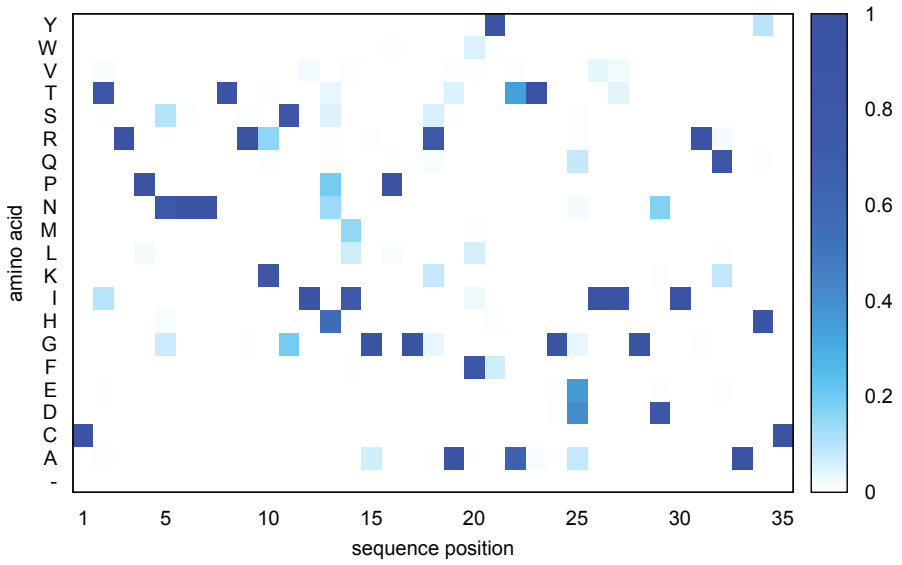
Based on the previously described R5 and X4 sequence data sets, we give a detailed description of the determination of the R5 and X4 fitness function. Additive and multiplicative terms for the fitness functions are introduced and discussed. In the second part of the section, the underlying R5 and X4 fitness landscapes are analysed.

4.4.1. Determination of R5 and X4 fitness function

We derived an independent R5 and X4 fitness functions from the respective R5 and X4 MSA of the V3 loop. Both fitness functions were built of a main fitness term, evaluating the one-dimensional aa sequence, and an epistatic term, accounting for the effects of evolutionary coupled pairs of mutations.

4.4.2. Position specific amino acid counts

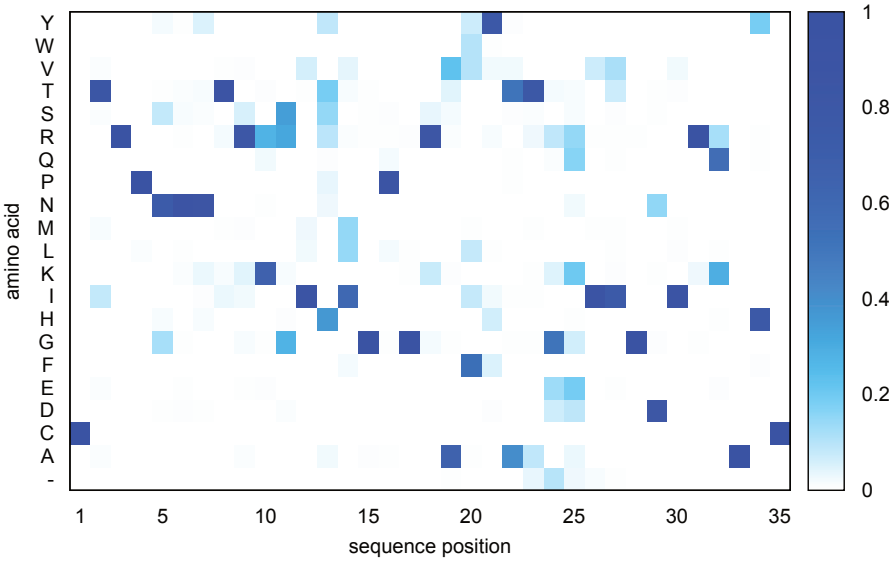
Following the sequence of steps presented in Figure 4.1, we first determined the frequency of each aa at each alignment position, separately for the R5 and X4 data set.
Figure 4.11 illustrates the position specific aa counts of the 14,008 R5 sequences as a heatmap, and Table 4.11 contains the respective numerical values. Since the sequence space of R5 was highly conserved, only one amino acid was dominant in most sequence positions. The few positions with ambiguous aa preferences were found in the regions flanking positions 11 and 25. For example histidine (`H`) is dominant in position 13, but only in half of the sequences (7,888 of 14,008). The remaining sequences mainly contain proline (`P`, 2,723), asparagine (`N`, 1,926), serine (`S`, 658), and threonine (`T`, 546).
A heatmap of the position specific aa counts of the 3,409 X4 sequences is given in Figure 4.12. Though we adapted the X4 MSA to a length of 35 amino acids, it still contained 679 gaps that were scattered throughout the sequences and positions, with most of the gaps accumulated in position 24.
In contrast to the R5 data set, a number of sequence positions of the X4 data set did not show a clear dominance of any aa, an observation that applied in particular for the co-receptor determining positions 11 and 25. Table 4.12, presents the exact position specific aa frequencies and clarifies this observation. In position 11, arginine (R) and serine (`S`) showed almost balanced counts of 1,086 and 1,184. This applied also for threonine (`T`, 1,754) and alanine (`A`, 1,381) in position 22. Most pronounced is the diversity for sequence position 25, with comparable frequencies for the four aa lysine (`K`, 700), glutamic acid (`E`, 661), glutamine (`Q`, 563), and arginine (`R`, 513).



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Y	0	1	0	0	15	7	7	0	0	0	0	0	5	1	0	0	0	0
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	138	4	0
V	0	200	0	42	0	0	0	0	2	0	0	297	0	128	21	12	0	0
T	0	12049	0	8	31	24	6	13958	2	125	2	12	546	26	37	10	0	21
S	0	90	0	0	1446	64	36	9	139	6	11248	0	658	1	3	18	0	793
R	0	3	14008	0	3	0	0	5	13698	2210	0	1	66	4	62	1	40	11133
Q	0	3	0	0	2	2	0	0	1	67	0	0	159	0	36	71	0	245
P	0	12	0	13629	1	0	0	2	0	0	1	0	2723	0	1	13490	0	1
N	0	1	0	0	11177	13835	13925	0	0	29	12	0	1926	1	0	0	8	2
M	0	12	0	0	0	0	0	2	31	3	0	53	0	2151	0	9	2	0
L	0	34	0	292	1	0	0	1	4	3	0	9	9	943	1	188	1	0
K	0	3	0	0	1	2	0	1	25	11475	0	2	1	2	1	0	1	1168
I	0	1407	0	0	5	3	3	2	36	9	0	13633	4	10674	1	0	0	6
H	0	1	0	37	236	21	4	0	0	0	0	0	7888	0	0	0	0	5
G	0	0	0	0	1030	1	0	0	61	10	2717	0	1	0	12910	9	13910	514
F	0	3	0	0	4	0	0	0	0	0	0	1	3	76	0	32	13	0
E	0	55	0	0	0	0	0	0	0	49	0	0	0	0	27	0	21	26
D	0	1	0	0	49	49	27	0	0	20	28	0	8	0	0	0	3	1
C	14008	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0
A	0	133	0	0	6	0	0	28	9	2	0	0	11	1	907	29	5	93
-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
Y	0	65	12798	5	0	1	0	0	0	0	3	0	0	0	0	1365	0
W	0	738	61	0	0	0	0	0	0	1	0	0	0	0	0	0	0
V	160	83	17	55	2	2	6	598	371	1	2	54	0	5	21	0	0
T	714	0	0	4805	13590	24	34	26	652	0	0	46	3	2	20	1	0
S	81	24	9	19	18	1	107	0	3	0	13	0	9	1	7	28	0
R	8	1	3	9	25	27	101	3	4	32	6	1	13883	302	0	49	1
Q	11	0	16	2	11	7	1127	0	0	0	14	0	0	12184	0	60	0
P	6	0	0	13	24	0	0	0	0	0	8	21	0	13	2	7	0
N	0	0	2	1	7	10	280	2	0	0	2429	0	0	5	0	6	0
M	6	126	0	24	39	0	0	37	42	0	0	32	0	0	0	0	0
L	0	832	7	2	1	1	1	12	32	0	0	2	0	80	0	2	0
K	1	0	0	0	0	20	25	4	1	3	85	0	50	1219	0	1	0
I	3	435	10	14	48	2	2	13325	12896	0	1	13852	7	5	0	0	0
H	0	2	112	0	11	0	5	0	0	0	7	0	0	36	0	12437	0
G	4	0	0	21	1	13701	494	0	1	13928	38	0	56	1	6	11	0
F	0	11700	947	0	0	0	0	0	2	0	1	0	0	0	0	30	0
E	5	0	1	5	1	151	5008	0	1	41	90	0	0	131	6	1	0
D	0	0	7	0	0	56	5706	1	0	0	11302	0	0	9	1	7	0
C	0	2	18	0	0	0	0	0	0	0	0	0	0	2	0	3	14007
A	13009	0	0	9033	230	5	1112	0	3	2	9	0	0	13	13945	0	0
-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 4.11.: **Position specific amino acid counts in the R5 data set**
The figure presents the position specific aa counts of the 14,008 R5 sequences and the table lists the respective numerical values. In general, one aa was dominant in each positions of the R5 MSA, but we found a few ambiguous positions in the regions flanking positions 11 and 25.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Y	0	0	0	0	77	20	176	0	0	0	0	3	294	9	0	0	0	1
W	0	0	0	0	0	0	0	0	0	0	2	1	1	0	0	12	2	0
V	0	43	0	5	0	1	1	19	29	0	3	204	3	138	1	0	1	1
T	0	2850	0	7	16	48	56	3100	9	34	12	24	640	59	16	13	1	9
S	0	46	0	0	288	59	41	8	193	5	1184	0	514	1	16	32	3	127
R	0	3	3409	0	3	25	4	74	2748	946	1086	3	330	41	16	15	37	2891
Q	0	2	0	0	0	3	1	1	7	88	0	0	27	0	6	76	0	9
P	0	8	0	3339	0	3	0	4	4	0	6	0	125	0	0	3130	6	2
N	0	0	0	0	2504	3108	2914	0	12	20	8	0	106	0	0	0	0	3
M	0	60	0	0	0	1	0	17	33	0	1	94	0	518	0	1	5	7
L	0	3	0	48	1	15	0	3	12	2	3	80	5	501	0	68	14	0
K	0	6	0	0	4	46	109	61	158	2248	59	0	0	3	0	0	17	259
I	0	287	0	0	10	13	28	110	87	1	8	2986	1	2062	0	3	0	10
H	0	0	0	10	58	4	59	1	0	3	29	0	1260	2	11	1	0	7
G	0	2	0	0	420	14	0	0	57	22	947	0	9	1	3301	23	3316	76
F	0	4	0	0	0	0	2	0	0	0	0	13	10	69	0	10	0	0
E	0	53	0	0	0	14	0	0	20	34	0	0	0	0	5	0	6	1
D	0	0	0	0	23	34	18	0	0	1	49	0	0	0	3	0	0	0
C	3409	0	0	0	2	0	0	0	0	0	3	0	2	0	1	0	0	0
A	0	42	0	0	3	1	0	11	40	5	9	1	82	5	33	25	1	6
-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
Y	11	249	2674	38	1	0	1	0	2	0	9	0	0	0	0	645	0
W	1	358	14	0	0	0	0	0	0	0	0	0	0	1	0	0	0
V	784	355	88	87	13	1	4	253	406	1	2	88	0	1	7	0	0
T	152	4	0	1754	2706	70	60	10	252	5	14	36	5	2	10	2	0
S	70	10	10	34	51	8	57	1	33	0	18	0	10	0	8	12	0
R	52	7	66	6	101	300	513	23	18	19	0	13	3264	418	2	20	0
Q	6	0	1	14	2	53	563	0	18	0	0	0	0	1908	0	26	0
P	10	0	1	16	12	1	0	0	9	0	0	4	0	1	7	3	0
N	4	0	5	1	3	2	90	2	8	0	529	1	0	3	0	2	0
M	0	14	1	7	16	8	0	16	21	0	0	26	0	0	0	0	0
L	0	268	25	12	5	0	0	7	17	0	1	37	0	15	1	6	0
K	37	0	0	9	25	166	700	7	33	5	17	12	94	1001	0	0	0
I	8	274	86	20	25	3	3	3028	2543	0	0	3186	9	8	0	0	0
H	2	3	222	1	2	1	26	0	0	0	0	0	0	15	0	2654	0
G	21	0	0	19	20	1751	220	0	1	3361	33	0	17	1	7	0	0
F	0	1845	175	1	0	0	0	0	0	0	0	0	0	0	0	37	0
E	4	1	1	1	4	464	661	0	14	13	5	0	4	28	0	0	0
D	1	1	31	2	1	233	310	2	5	0	2775	0	0	0	0	0	0
C	1	0	4	1	0	0	0	0	0	0	0	0	0	0	0	2	3409
A	2228	20	3	1381	300	7	107	3	9	2	6	0	0	1	3367	0	0
-	17	0	2	5	122	341	94	57	20	3	0	6	6	6	0	0	0

Figure 4.12.: **Position specific amino acid counts in the X4 data set**
The figure presents the position specific aa counts of the 3,409 X4 sequences and the table lists the respective numerical values. Many sequence positions are highly diverse, some dominated by two aa at similar rates (e.g. `R` and `S` in position 11), others by three to four aa (e.g. position 25).
In some positions, no clear dominance of any aa could be observed, in particular for the co-receptor determining positions 11 and 25.

The derivation of the position specific aa frequencies started from 79,688 V3 loop sequences. During data analysis we concluded that our data reduction method did not lead to relevant changes in the position specific aa probabilities. Despite this fact, we are aware that our data set is not infinite and does not cover the complete V3 loop sequence space. We presumably missed V3 loop sequences with further position specific aa that are not represented in our finite data sets.
This kind of problems, known as regularisation problems, was described by Tikhonov [150]. The method of regularisation tries to find suitable approximations to solve ill-conditioned mathematical problems that arose from finite data or noise contributions.
We addressed the finite size effect of our data by the use of a regularisation parameter γ and adapted the observed aa frequencies using the following equation:

$$freq_{new_{ij}} = freq_{orig_{ij}} + \gamma \tag{4.9}$$

By an addition of $\gamma = 1$ to any frequency value $freq_{orig_{ij}}$ of any amino acid j at any sequence position i, we accounted for amino acids that we possibly have not seen at position i in our data set. This regularisation equation explicitly includes positions with amino acid counts of zero.
The selection of a small γ value guaranteed only marginal alterations of the real amino acid frequencies of our data.

4.4.3. Main fitness contribution fit_{main}

The position specific aa counts of the R5 and X4 data set are the basis for the computation of the main fitness contribution fit_{main}. Given a specific V3 loop sequence, we compute two independent fitness contributions $fit_{main_{R5}}$ and $fit_{main_{X4}}$, using the respective R5 or X4 frequency table (compare Figures 4.11 and 4.12). Based on the frequencies $freq_{ij_c}$, we calculated the position specific aa probabilities $p_{ij_c} = \frac{freq_{ijc}}{N_c}$, with c being the respective co-receptor and $N_{R5} = 14{,}008$ and $N_{X4} = 3{,}409$ being the number of R5 and X4 sequences.

We considered two possible formulations to calculate the replicative fitness, an *additive* and a *multiplicative* fitness function:

$$
\begin{aligned}
addFit_{main_{R5}} &= \sum_{i=1}^{35}\sum_{j=1}^{20} p_{ij_{R5}} r_{ij} \ ; \\
addFit_{main_{X4}} &= \sum_{i=1}^{35}\sum_{j=1}^{20} p_{ij_{X4}} r_{ij} \ ; \\
\\
multFit_{main_{R5}} &= \prod_{i=1}^{35}\sum_{j=1}^{20} p_{ij_{R5}} r_{ij} \ ; \\
multFit_{main_{X4}} &= \prod_{i=1}^{35}\sum_{j=1}^{20} p_{ij_{X4}} r_{ij} \\
\\
\text{with } r_{ij} &= \begin{cases} 1 \text{ if aa } j \text{ is at position } i \\ 0 \text{ else} \end{cases}
\end{aligned}
\tag{4.10}
$$

and the probability $p_{ij_{R5}}$ of aa j at position i in the R5 data set and the probability $p_{ij_{X4}}$ of aa j at position i in the X4 data set.
For any sequence, $r_{ij} = 0$ for all but amino acid j at position i, and thus the inner sum $\sum p_{ij} r_{ij} = 0.0$ for all but one amino acid.
Using the additive fitness function defined in Equation 4.10, we summed up the respective aa 35 probabilities in each position. Maximal fitness values $addFit_{main_{R5}} = 31.26$ and $addFit_{main_{X4}} = 27.26$ were assigned to the respective R5 and X4 consensus sequence. The deviations of the additive main fitness contribution from the absolute maximum of 35.0 reflected the deviations in the conservation of the sequence positions.
To observe the impact of an aa mutation on the replicative fitness, we analysed a number of random sequence examples. For clarification, a small example of two shortened V3 loop sequences a and b (V3 loop positions one to 11) is presented. Sequence a is composed of the consensus amino acids, while position seven of sequence b carries a biologically unfavourable mutation (`E`) that was not observed at this position in both data sets.
We computed the fitness of the example sequences based on the aa frequencies $freq_{ij_{R5}}$ and $freq_{ij_{X4}}$ shown in Tables 4.11 and 4.12, altered by the regularisation Equation 4.9:
sequence a: `CTRPNNNTRKS`

$$addFit_{main_{R5}}(a) = \frac{14009+12050+14009+13630+11178+13836+13926+13958+13699+11476+11249}{14008+20\cdot\gamma} = 10.20$$

$$addFit_{main_{X4}}(a) = \frac{3410+2851+3410+3340+2505+3109+2915+3101+2749+2247+1185}{3409+20\cdot\gamma} = 8.99$$

sequence b: `CTRPNNETRKS`

$$addFit_{main_{R5}}(b) = \frac{14009+12050+14009+13630+11178+13836+1+13958+13699+11476+11249}{14008+20\cdot\gamma} = 9.20$$

$$addFit_{main_{X4}}(b) = \frac{3410+2851+3410+3340+2505+3109+1+3101+2749+2247+1185}{3409+20\cdot\gamma} = 8.14$$

The examples showed deviations of the additive replicative fitness of ≈ 1.0 for the R5 fitness function and of 0.85 for the X4 fitness function, reflecting the reduced replicative fitness of the mutated aa in position seven. Thus, the mutation of a highly conserved aa into an aa not observed in that sequence position of the underlying MSA leads to a fitness reduction of ≈ 1.0.
The regularisation Equation 4.9 with $\gamma = 1$ changed the additive fitness of the consensus sequences in two decimal points (data not shown).
In the alternative formulation of the main fitness contributions, we multiplied the position specific aa probabilities. In this case, maximal fitness values $multFit_{main_{R5}} = 1.149 \cdot 10^{-2}$ and $multFit_{main_{X4}} = 2.779 \cdot 10^{-5}$ were calculated for the consensus R5 and X4 sequence. The consensus sequence maxima differed remarkably from the theoretic fitness maximum of 1.0 (1.0^{35}), which would be assigned to a completely conserved data set. These deviations reflected the variability in the position specific conservation of the aa sequences.
We again demonstrate the calculation of the replicative fitness for sequences a and b, carrying the aa mutation `E` in position seven.

sequence a: `CTRPNNNTRKS`

$$multFit_{main_{R5}}(a) = \frac{14009\cdot12050\cdot14009\cdot13630\cdot11178\cdot13836\cdot13926\cdot13959\cdot13699\cdot11476\cdot11249}{(14008+\gamma)^{11}} = 0.42$$

$$multFit_{main_{X4}}(a) = \frac{3410\cdot2851\cdot3410\cdot3340\cdot2505\cdot3109\cdot2915\cdot3101\cdot2749\cdot2249\cdot1185}{(3409+\gamma)^{11}} = 0.08$$

sequence b: `CTRPNNETRKS`

$$multFit_{main_{R5}}(b) = \frac{14009\cdot12050\cdot14009\cdot13630\cdot11178\cdot13836\cdot1\cdot13959\cdot13699\cdot11476\cdot11249}{(14008+\gamma)^{11}} = 3.02 \cdot 10^{-5}$$

$$multFit_{main_{X4}}(b) = \frac{3410\cdot2851\cdot3410\cdot3340\cdot2505\cdot3109\cdot1\cdot3101\cdot2749\cdot2249\cdot1185}{(3409+\gamma)^{11}} = 2.70 \cdot 10^{-5}$$

In case of the multiplicative fitness function, the mutation of a conserved aa was highly pronounced and resulted in a maximal fitness reduction of $\frac{1}{(14008+\gamma)} \approx 10^4$ for the multiplicative R5 fitness function and in a maximal fitness reduction of $\frac{1}{(3409+\gamma)} \approx 10^3$ for the multiplicative X4 fitness function. This variation of the replicative fitness was in a range we would expect from an aa that is biologically and chemically unfavourable.
An analysis of the influence of the regularisation Equation 4.9 with $\gamma = 1$ revealed that the regularisation altered the fitness value of the consensus sequences in five (R5) respective four (X4) decimal points (data not shown), except for the case when a frequency of 0 is replaced by a frequency of $\gamma = 1$. In that case, the regularisation Equation 4.9 with $\gamma = 1$ avoids the multiplication of a zero probability and restricts a resulting zero fitness.
Based on these observations, in combination with the simulation results we present in Section 5.3.1, we decided to use the multiplicative formulation of the fitness function.

4.4.4. Counts of pairs of coupled mutations

The second part of the fitness function is an epistatic fitness contribution. Epistatic positions are characterised by a coupled evolution, meaning that mutations of the amino acids at two (or more) positions of the V3 loop are linked and the occurrence of an aa mutation in one position influences the aa in the other position.
For the replicative capacity of an individual, the epistatic fitness effect can either be positive or negative. If the mutual occurrence of two aa mutations m_i and m_j in two coupled positions i and j leads to a fitness benefit compared to the fitness effects of both single mutations (e.g. two cysteine residues that build a disulfide bridge), we speak of positive epistasis. In the case of a negative epistatic effect, two contemporary aa mutations lead to a decreased replicative fitness compared to the sole occurrence, for example due to steric hindrance in the case of two large aa in close proximity.
We discriminated positions of positive and negative epistatic based on the sign of the entries of the cross correlation matrix described in Section 4.2.1.

Structural coupling

There are different methods to detect co-evolving positions in sequences. We first used the measure of mutual information complemented by the structural coupling analysis as provided by the R Bioconductor [121, 77] package to determine pairs of co-evolutionary coupled positions in the R5 and X4 MSA.
The structural coupling analysis is based on an MSA and an additional crystal structure of the respective molecule. To obtain a crystal structure of the V3 loop, we used the PDB data base [12, 11], an archive of macromolecular structures. A search for 'env HIV' in the database resulted in 104 structure hits, a search for 'gp120' delivered 153 hits and searching for 'V3' we found 86 structures. The different search results were mainly subsets of each other.
A close inspection of the files revealed that most of the structures contained incomplete V3 loop sequences. Especially the gp120 residues 300 to 320 (V3 loop positions five to 25) were missing in almost all .pdb structure files.
At the date of our search, only the following three .pdb files contained complete HIV-1 V3 loops:

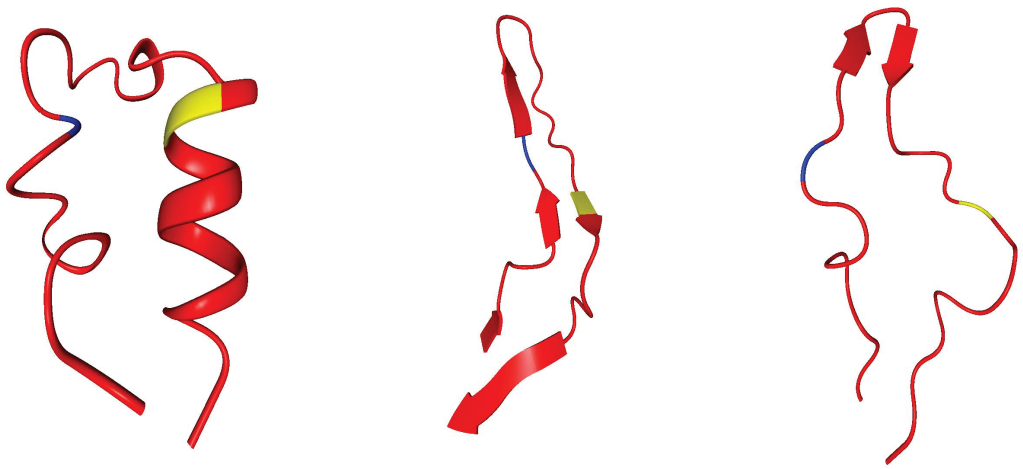
1CE4.pdb The file [153] contained a conformational model for the consensus V3 loop of the HIV-1 envelope protein gp120 published in 1997. The authors presented no unique crystal structure, but described a summary of observations from several structures, and combined them into one consensus model. The resolution of the structure thus could not be determined. In the 1CE4.pdb structure, a synthetically disulfide bridge between cys296 (loop position 1) and cys331 (loop position 35) has been introduced due to crystallisation issues. In consequence, we did not consider this contact during the structural coupling analysis.
We extracted the following aa sequence of the V3 loop from the file:
`CTRPNNNTRK SIHIGPGRAF YTTGEIIGDI RQAHC`

2QAD.pdb In 2007, the group of Kwong *et al.* [78] published the structure file 2QAD.pdb. The file contained the spatial structure of a tyrosine-sulfated antibody,

complexed with the complete HIV-1 envelope protein gp120 and the CD4 receptor. The resolution of the file was 3.3 Å.
We trimmed the molecule to the V3 loop and extracted the following aa sequence:
`CTRPNNNTRK SINIGPGRAL YTTGEIIGDI RQAHC`

2B4C.pdb The third structure was also published from the group of Kwong [79]. The file 2B4C.pdb contained a crystal structure of the gp120 core protein of HIV-1, complexed with CD4 and an X5 antibody. The resolution of the file was specified as 3.3 Å.
We extracted the following sequence, restricted to the V3 loop residues, from the file:
`CTRPNQNTRK SIHIGPGRAF YTTGEIIGDI RQAHC`

Though the research groups of Zolla-Pazner *et al.* and Kwong *et al.* did excessive research in the field of V3 loop structure, no further V3 loop structures were available. The majority of the envelope structures was lacking the V3 region or missed at least the central residues of the loop due to problems upon the crystallisation.
We used the BALLView [107, 106] package to visually inspect the crystal structures and to restrict the sequence to fit the V3 loop region. Figure 4.13 illustrates the three V3 loop structures we extracted from the files.



(a) (b) (c)

Figure 4.13.: **Selected V3 loop structures**
The illustration presents the V3 loop structures extracted from the files 1CE4.pdb (left), 2QAD.pdb (center), and 2B4C.pdb (right). The co-receptor determining position 11 is coloured in blue and position 25 in yellow.

The BALLView software package furthermore enabled us to calculate the root mean square deviation (RMSD), a measure used to determine spatial deviations between protein structures. The best match of the C_α atoms of the protein backbone of the three V3 loop

structures is depicted in Figure 4.14.
Using the backbone mapping, BALLView calculated the following RMSDs:

- 1CE4 - 2QAD: 10.20 Å
- 1CE4 - 2B4C : 7.55 Å
- 2B4C - 2QAD: 5.84 Å

Regarding the resolution of the crystal structures, the RMSD of each pair of structures was small, thus we decided to perform a coupling analysis for all three structures.

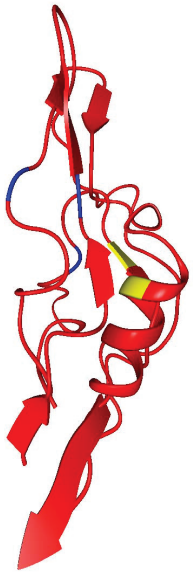


Figure 4.14.: **Optimal match of selected V3 loop structures**
The figure illustrates the best match of the C_α atoms of the V3 loop structures 1CE4, 2QAD, and 2B4C. The amino acid in position 11 is coloured in blue, while position 25 is depicted in yellow.

Despite several independent approaches with varying parameters, the coupling analysis did not lead to any meaningful results. We found a number of co-evolving positions represented by high MI values, but the coupling analyses did not detect positions that showed a significant signal of structural conservation. The spatial V3 loop structure only marginally changed upon the disruption of any protein contact. In consequence, we could not determine positions either in the R5 or in the X4 data set that were both coupled in sequence and in structure.
During private communication with biologists, we addressed this finding to the fact that the bound V3 loop is highly conserved in structure, but less conserved in the aa sequence. The unbound V3 loop in contrast is highly flexible and undergoes large conformational changes upon binding [7, 22], which are still unresolved in detail [140, 78].
Last but not least, the crystallisation of membrane proteins in general is very challenging or often impossible without further modifications. The introduction of a synthetic disulfide bond between positions one and 35 was already described for the structure 1CE4.pdb.
These considerations led us the finding that the structural coupling analysis is not applicable for the V3 loop.

Mutual Information

We next concentrated on methods that are solely sequence based. We used the R Bioconductor [121, 77] package to calculate the mutual information and to detect pairs of co-evolving sequence positions in the MSAs. For an optimal handling of the 679 gaps we observed in the X4 MSA, we used the subset mutual information (SUMI) described in Section 4.2.1. We further applied the shuffle null model of Weil *et al.* [158] to discriminate significant co-evolution signals from background noise.
The SUMI values of the sequence positions pairs are depicted in Figure 4.15. The matrix of the R5 SUMI values showed co-evolutionary signals in positions 13 to 14 linked to positions 18 and 20. Furthermore, positions five and 25 were weakly coupled to a number of additional positions, indicated by the increased SUMI values in column 25.
The X4 data set contained an increased number of coupled positions and also higher SUMI values compared to the R5 data set, but in general, the same regions were detected. The co-evolutionary X4 signal extended from positions 11 to 14, coupled to positions 19 to 25. In addition, positions 19 to 25 of the X4 data set are evolutionary coupled intra-sequentially, indicated by the increased SUMI values in positions 19 to 25.
In summary we found increased co-evolutionary signals for the co-receptor determining positions 11 and 25 and the flanking amino acids.

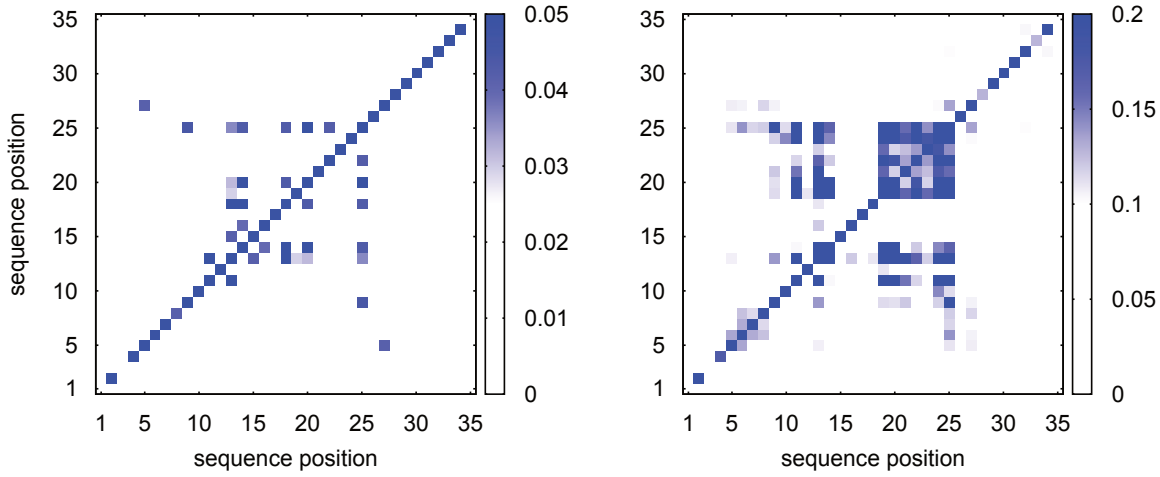


Figure 4.15.: **Mutual information of R5 and X4 sequence alignment**
The heat maps illustrate the R5 SUMI values (left) and the X4 SUMI values (right). In general, both data sets contain a number of coupled positions in the co-receptor determining regions, but the coupling is more pronounced in the X4 data set.

To discriminate the significant SUMI values from the background noise, we used the Z-score normalisation approach described in Section 4.2.1. The resulting Z-scores are illustrated in Figure 4.16.
In general, the Z-scores confirmed the pattern of the co-evolving positions and increased the strength of the signal. The co-evolutionary signal of the R5 MSA was expanded to positions 11−14 coupled to positions 18−20, the pattern of the X4 MSA was conserved Z-score normalisation.

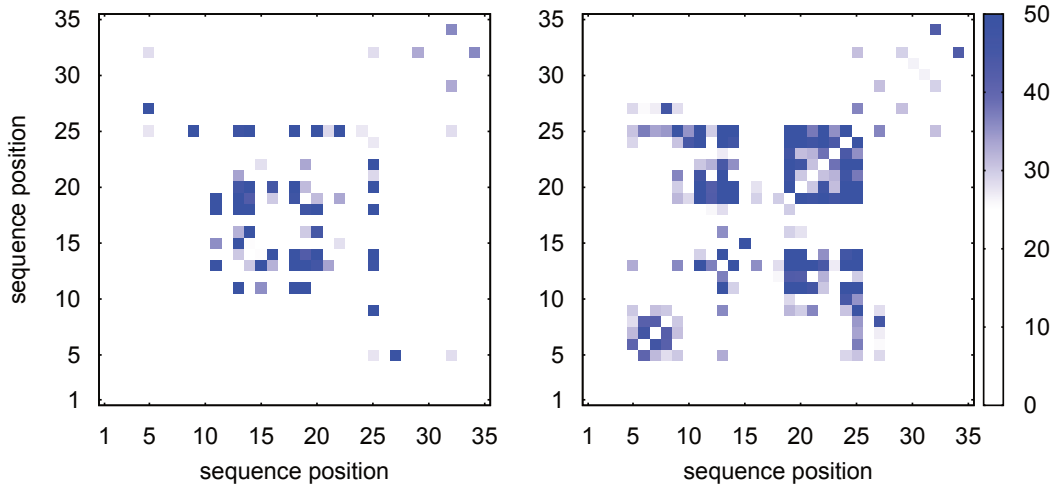


Figure 4.16.: **Z-scores of mutual information of R5 and X4 sequence alignment**
The heat maps illustrate the Z-scores of the R5 (left) and X4 (right) SUMI values. The Z-score normalisation expanded the region of the coupled R5 MSA positions towards positions 11−14 linked to positions 18−20 and confirmed the co-evolutionary signal of the X4 MSA.

Apart from the Z-score normalisation approach, Martin *et al.* [103] presented another idea to discriminate significant values from background noise. They described a method using the entropy values $H(X, Y)$ to normalise the data. To cross-check our findings, we also applied this normalisation method to the SUMI values. Our analyses confirmed the co-evolutionary pattern we found upon Z-score normalisation (data not shown).
In summary, we found an evolutionary coupling of the co-receptor determining positions 11 and 25 and their flanking regions. We know from literature [54] that contemporary mutations of positions 10 to 14 and 25 to 29 of the V3 loop from neutral or negatively charged residues to positively charged residues are indicative of the co-receptor switch. Thus, the observed co-evolutionary pattern in the co-receptor determining regions con-

firmed our expectations about the data and indicated that we found relevant epistatic positions that were cross-linked during evolution.

Cross correlation

Mutual information values are defined to be positive. Though the SUMI was highly suitable to reveal the co-evolving positions, the measure failed to discriminate positive and negative epistatic effects of the coupled positions. To cope with this problem, we applied the cross correlation method described in Section 4.2.1.

Cross correlation of the R5 and X4 data set

Using Perl [26] and additional content of the packages *List* [8] and *Math::MatrixReal* [13], we computed the cross correlation (CC) matrices of the R5 and X4 MSA using Equation 4.5. The resulting matrices C_{R5} and C_{X4} are illustrated in figure 4.17.

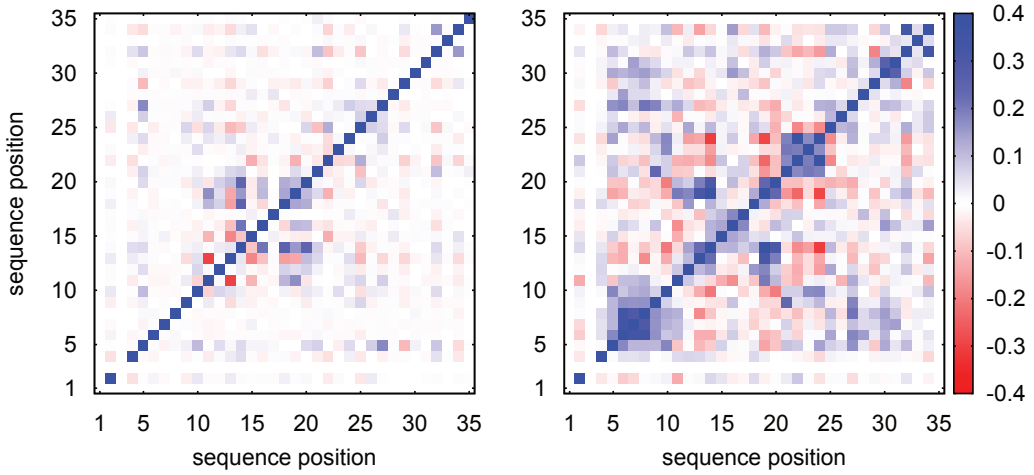


Figure 4.17.: **Cross correlation of R5 and X4 sequence alignment**
The left heatmap illustrates C_{R5}. While pairs (13,18) and (13,20) were highly positively correlated, the matrix showed a strong negative correlation for the pair (11,13).
C_{X4} (right) contained more and increased signals. The pattern of positions 11−14 correlated to positions 19−24 contained the largest positive values of C_{X4} for pairs (13,19) and (14,19), and comprised highly negative correlations. The largest negative correlation was observed for pair (19,24).

Positively correlated positions dominated C_{R5}. The correlations of position 14 coupled to positions 18−20 were pronounced. In contrast, the pair (11,13) showed the largest negative correlation.
Similar to the MI approach, C_{X4} contained a higher frequency of cross linked positions. Furthermore, the CC analysis revealed a high number of negatively correlated pairs in C_{X4} compared to C_{R5}.
In detail, positions 11−14 were positively correlated with positions 19−20, and negatively correlated with positions 22−24. This pattern contained the largest positive values of C_{X4} for pairs (13,19) and (14,19), and also the highly negatively correlated pair (14, 24). The largest negative correlation was calculated for the pair (19,24).
The correlations of positions 1, 3, and 35 were ~ 0.0 in both matrices, since the positions of the respective MSAs were highly conserved.
Next, we applied the noise reduction method to the CC matrices. After a column shuffling of both the R5 and the X4 MSA, we calculated the matrices $C_{R5_{shuff}}$ and $C_{X4_{shuff}}$ and performed eigenvalue decompositions. We found the largest eigenvalue in the sixth shuffling run for the R5 MSA and the 13th for the X4 MSA, therefore we stopped the shuffling after 500 runs for each alignment.
The resulting eigenvalues of the real and column-shuffled MSAs are illustrated in Figure 4.18. We can clearly see the differences between the eigenvalues of the shuffled alignments, which concentrate in the range from 0.7 to 1.2 for both data sets, and the eigenvalues λ_{R5} and λ_{X4} of the original MSAs that are spread over a range of 0.3 to 2.8. The largest eigenvalues of the shuffled MSAs were $\lambda_{R5_{shuffmax}} = 1.177$ and $\lambda_{X4_{shuffmax}} = 1.221$.
The thresholds that we used for separation were determined as half the distance of the largest real eigenvalue that was smaller and the smallest real eigenvalue that was larger than the largest random eigenvalue. Based on these thresholds, we kept four eigenvalues of the R5 MSA and six eigenvalues of the X4 MSA. From these eigenvalues and corresponding eigenvectors, we reconstructed the noise-reduced CC matrices for both data sets. The reconstructed matrices are depicted in Figure 4.19.
An analysis of the differences of the original CC C_{R5} and C_{X4} in Figure 4.17 and the noise-corrected matrices $C_{R5_{clean}}$ and $C_{X4_{clean}}$ in Figure 4.19 revealed that the method of noise-reduction worked as intended. While the previously identified cross-linked pairs were conserved by the method, the signals increased and became more pronounced.

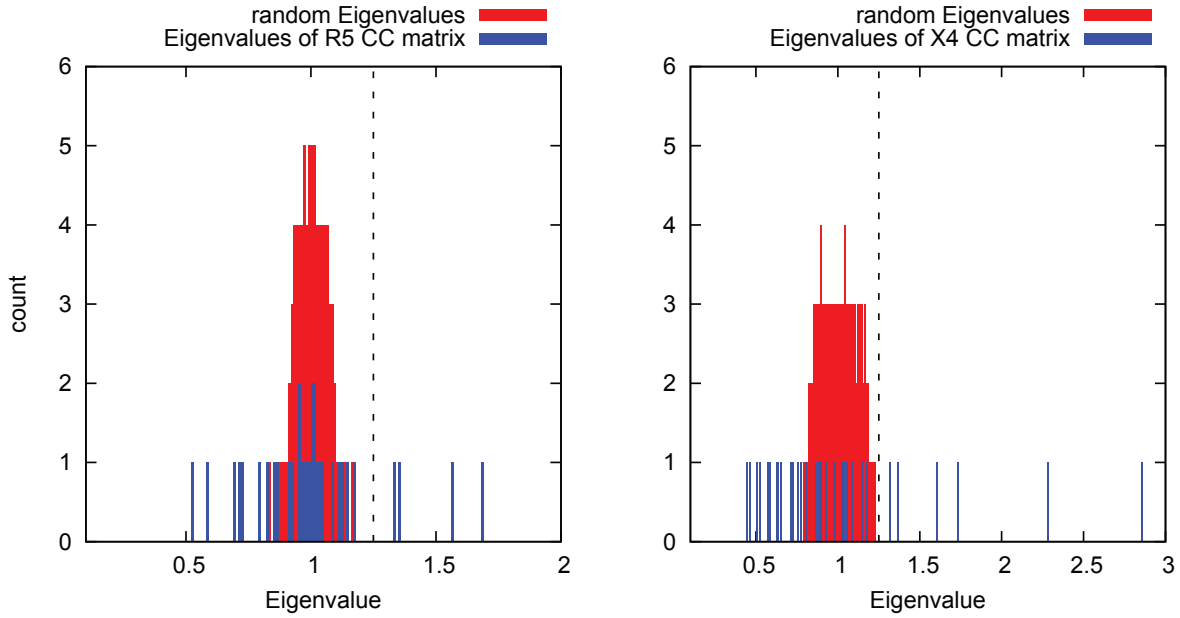


Figure 4.18.: **Eigenvalues of R5 and X4 CC matrices**
The figures illustrate the eigenvalues of the CC matrix of the original MSAs in red, the eigenvalues of the matrices of the shuffled MSAs in blue, and the selected threshold as a black dashed line. Four eigenvalues λ_{R5} of the R5 MSA (left), and six eigenvalues λ_{X4} of the X4 MSA (right) were larger than the threshold.

In their approach, Dahirel *et al.* [32] next removed the influence of the largest eigenvalue $\lambda_{C_{max}}$ and the corresponding eigenvector $k_{C_{max}}$ of C, since they claimed this contribution merely resembled the phylogenetic history of the data and contained no relevant co-evolutionary information. After a recalculation of the eigenspectrum, the researchers reconstructed the cross correlation matrix $C_{clean_{phyl}}$.
Based on the similarity of the presented method to the principal component analysis (PCA), we were not convinced to remove the largest eigenvalue of the matrices, since in a PCA the largest eigenvalue is the one that describes the majority of the data.
For clarification, we computed $C_{clean_{phyl}}$ using a simplified approach. We removed the largest eigenvalues of $C_{R5_{clean}}$ and $C_{X4_{clean}}$ and directly reconstructed the matrices $C_{R5_{clean_{phyl}}}$ and $C_{X4_{clean_{phyl}}}$, without a recalculation of the eigenspectrum.

An analysis of the differences of the matrices before and after removal of the largest eigenvalue $\lambda_{C_{max}}$ showed that the exclusion of the largest eigenvalue in general conserved the co-evolutionary pattern, but decreased the strength of the signal (data not shown).
From this observation, we suggested that our data originated from a consistent phylogenetic background. Due to the small impact of this additional data processing, we skipped this step and used the noise-corrected matrices $C_{R5_{clean}}$ and $C_{X4_{clean}}$ for further computations.

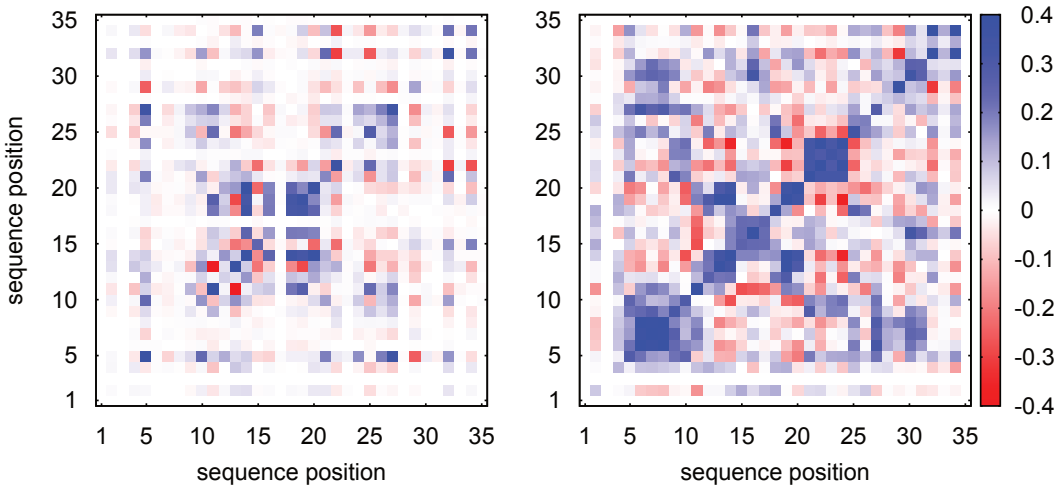


Figure 4.19.: **Noise-reduced CC matrices of R5 and X4 MSA**:
The left figure illustrates the reconstructed matrix $C_{R5_{clean}}$, and the right figure the matrix $C_{X4_{clean}}$. Noise-reduction conserved and pronounced the co-evolutionary signals.

A comparison of the CC matrices depicted in Figure 4.19 on the one hand and the MI Z-score matrices presented in Figure 4.16 on the other hand revealed, that both the CC and the MI method found mainly the same co-evolution pattern. Thus, the well-established and frequently applied MI method confirmed the validity of the recent CC approach for our data.

4.4.5. Epistatic fitness contribution fit_{epi}

The noise-corrected CC matrices $C_{R5_{clean}}$ and $C_{X4_{clean}}$ build the basis for the computation of the epistatic fitness term fit_{epi}. The epistatic fitness contribution is defined by the

following equations:

$$
\begin{aligned}
fit_{epi_{R5}} &= \prod_{i,k=1}^{35} \sum_{j,l=1}^{20} c_{ik_{R5}} s_{ij} s_{kl} \\
fit_{epi_{X4}} &= \prod_{i,k=1}^{35} \sum_{j,l=1}^{20} c_{ik_{X4}} s_{ij} s_{kl} \\
\text{with } s_{ij} &= \begin{cases} 1 \text{ if aa } j \text{ at position } i \text{ is mutated} \\ 0 \text{ else} \end{cases} \\
\text{and } s_{kl} &= \begin{cases} 1 \text{ if aa } l \text{ at position } k \text{ is mutated} \\ 0 \text{ else} \end{cases}
\end{aligned}
\tag{4.11}
$$

and $c_{ik_{R5}}$ and $c_{ik_{X4}}$ being the CC values of position pair (i, k) of the corresponding R5 or X4 MSA.
If both amino acids j and l at the observed positions i and k are mutated, $s_{ij}s_{kl} = 1$, else $s_{ij}s_{kl} = 0$. Thus, for each pair (i, k) the inner sum contributes exactly once.
On basis of the epistasis equations, we used a bootstrapping approach to determine the range of the epistatic fitness values. From the set of all possible aa sequences of length 35, we randomly picked two times $1.6 \cdot 10^6$ sequences and calculated the respective $fit_{epi_{R5}}$ and $fit_{epi_{X4}}$ values. The distributions of the fitness values of both runs were mainly concordant. Thus, we could derive reliable estimates for the range of the epistatic fitness values (data not shown).
Though the main fitness contribution of R5-tropic sequences is generally higher, due to the higher sequence conservation of the R5 data set (compare Section 4.4.3), we observed a larger epistatic fitness contribution for the X4-tropic sequences. A retrospect of the CC matrices revealed that this finding was result of the more frequent epistatic positions and the increased epistatic signals of $C_{X4_{clean}}$.

4.4.6. Complete R5 and X4 fitness function

Based on the R5 and X4 MSA, we determined the position specific aa probabilities to compute the main fitness term fit_{main} (compare Equation 4.10), and we calculated CC matrices to compute the epistatic fitness term fit_{epi} (compare Equation 4.11). We decided to use a multiplicative fitness formulation to ensure fitness values close to zero for sequences with biologically unlikely aa mutations.
First analyses revealed that a combination of the main and epistatic fitness terms resulted in small fitness values, even in the case of the consensus sequences, with a fitness of $1.149 \cdot 10^{-2}$ for the R5 consensus sequence and of $2.779 \cdot 10^{-5}$ for the X4 consensus sequence. Test sequences that deviated in one aa decreased the resulting R5 fitness by a factor of 10^{-4} and the X4 fitness by a factor of 10^{-3} (compare Section 4.4.3). Further tests of the epistatic Equation 4.11 revealed that the fitness of sequences that accumulated mutations could become negative due the multiplication of an odd number of negative epistatic terms. Negative replicative fitness values are biologically contradictory and very small numbers

can lead to numerical instabilities. We resolved these difficulties by the usage of an exponential fitness function. While negative fitness values in the exponent of e resulted in small positive fitness values, the multiplication of the aa probabilities was converted into an addition and thus numerical inconsistencies due to the multiplication of 35 probability values < 1.0 were avoided. The usage of the exponential fitness function was paired with the usage of the inverse logarithmic function. The Taylor expansions of a logarithmic function:

$$\ln(x) = (x-1) - \frac{(x-1)^2}{2} + \frac{(x-1)^3}{3} - \cdots$$

reduces to $\ln(x) \approx (x-1)$ for small x. Since we used a normalisation of our fitness values to an interval of $[0, 1]$, the expression could be reduced further to $\ln(x) \approx x$. Based on these considerations, we used an exponential expression without an inverse logarithmic operation. A subsequent normalisation of the fitness values to the interval of $[0, 1]$ was used to perform the the balancing of the main and epistatic fitness terms.

Since we applied different methods to derive the single fitness contributions from our data, the ratio of the main and the epistatic fitness contribution had to be derived in a sequence of independent steps. In a first step, we approximated the task by scaling both main and epistatic fitness to the interval of $[0, 1]$. Therefore, we determined the maximal fitness values for the main and epistatic term for both the R5 and X4 contribution. The maxima of the main fitness terms equal the main fitness of the consensus sequence, since the consensus sequence is computed based on the most frequent amino acid in each sequence position. The determination of the maxima of the epistatic terms was more challenging. We performed a bootstrapping approach, creating two times 5% of all possible sequences, and determined their epistatic fitness. A comparison of the distribution of the two sets of epistatic fitness values showed similar fitness distributions. Therefore, we could estimate the maximal fitness values by the use of the maximal fitness of the independent random sequence subsets.

Based on the maximal main and epistatic fitness values of both the R5 and X4 data set, we performed a normalisation of each single fitness contribution to the interval $[0, 1]$ by the division of the fitness values by the respective maximal value.

Using a weighting factor β enabled us to further balance the influence of the main and epistatic fitness contribution. Large values of β represent a strong influence of the epistatic contribution, while small values of β decrease the epistatic contribution to the replicative fitness of a sequence. Finally, we combined the individual R5 and X4 fitness functions by an additional weighting factor α. A value $\alpha = 1$ restricted the calculation of the replicative fitness to the R5 fitness term, and $\alpha = 0$ restricted the calculation to the X4 fitness term. A variation of α could for example be used to mimic the availability of CCR5 or CXCR4 positive target cells.

The complete exponential fitness function $eFit$ is defined as follows:

$$
\begin{aligned}
eFit &= \alpha eFit_{R5} + (1-\alpha) eFit_{X4} \\
\text{with } eFit_{R5} &= e^{\sum_{i=1}^{35}\sum_{j=1}^{20} p_{ij_{R5}} r_{ij}} + \beta e^{\sum_{i,k=1}^{35}\sum_{j,l=1}^{35} c_{ik_{R5}} s_{ij} s_{kl}} \\
\text{and } eFit_{X4} &= e^{\sum_{i=1}^{35}\sum_{j=1}^{20} p_{ij_{X4}} r_{ij}} + \beta e^{\sum_{i,k=1}^{35}\sum_{j,l=1}^{35} c_{ik_{X4}} s_{ij} s_{kl}} \\
\text{and } r_{ij} &= \begin{cases} 1 \text{ if aa } j \text{ is at position } i \\ 0 \text{ else} \end{cases} \\
\text{and } s_{ij} &= \begin{cases} 1 \text{ if aa } j \text{ at position } i \text{ is mutated} \\ 0 \text{ else} \end{cases} \\
\text{and } s_{kl} &= \begin{cases} 1 \text{ if aa } l \text{ at position } k \text{ is mutated} \\ 0 \text{ else} \end{cases}
\end{aligned}
\tag{4.12}
$$

and $p_{ij_{R5}}$ being the probability of aa j at position i in the R5 data set and $p_{ij_{X4}}$ being the respective probability in the X4 data set. $c_{ik_{R5}}$ is the cross correlation of position i and position k in the R5 data set and $c_{ik_{X4}}$ the cross correlation of position i and position k in the X4 data set. While α determines the R5 fitness contribution, and $1-\alpha$ the respective X4 fitness contribution, β balances the influence of the epistatic effect on the joined fitness. The normalisation of the main and epistatic term for the R5 and X4, performed by the division by the respective maximal values, is not expressed in the equation for reasons of clarity. When we used Equation 4.12 during subsequent analyses, we used the normalised expression of the exponential fitness equation.

The normalised exponential fitness equation assigns the maximal fitness value of 1.0 to the R5 consensus sequence for $\alpha = 1.0$, while $\alpha = 0.0$ assigns the maximal value of 1.0 to the X4 consensus sequence. The selection of a weighting parameter $\beta < 1.0$ reduces the influence of the epistasis and thus assigns a fitness value smaller than the maximal value to any sequence that deviates from the consensus sequence. A selection of $\beta > 1.0$ pronounces the epistatic effect. Analyses showed that this results in the accumulation of random sequence mutations and destabilises the conservation of the V3 loop sequence. Thus, a selection of $\beta < 1.0$ in Equation 4.12 is recommended.

Mutations that introduce stop codons into the V3 loop typically terminate the protein translation and result in truncated, non-functional proteins. Since the functionality of the loop is essential to HIV, truncated sequences would not enable the process of co-receptor binding and cell entry and inhibit viral replication. Therefore the replicative fitness of sequence variants carrying stop codons was defined to be zero.

It has to be noted that the absolute fitness values defined by Equation 4.12 are only sound for an intra-R5 or intra-X4 population ranking, since the fitness contributions were derived from two independent MSA and we lacked a method to compare the replicative fitness

of the R5 and X4 sequences. Thus, an inter-population comparison of an R5 and an X4 fitness value is not possible with the presented fitness function.

4.4.7. Structure of modelled fitness landscape

We derived fitness functions to describe the replicative fitness V3 loop sequences. Our main concern were the differences between the R5 and X4 population, as well as the evolutionary pathways that transform R5-tropic viruses into X4-tropic viruses. Our fitness functions built the basis to analyse the structural differences of the R5 and X4 fitness landscape.

Local fitness landscape

We created two populations of unique nucleotide (nt) sequences, derived from the R5 and X4 nt consensus sequence by the introduction of a predefined number of mutations. The complete sequence space of the V3 loop covers $4^{105} = 1.6 \cdot 10^{63}$ sequences. Due to this multitude of sequences, it is not possible to analyse the complete sequence space.
A limitation to all five-point mutants of a V3 loop consensus sequences still results in $4\binom{105}{5} \approx 3.9 \cdot 10^8$ sequences. First analyses of two random samples of five-point mutants showed that many of the sequences contained stop codons or yielded very low fitness values due to mutations in highly conserved positions. Following these observations, we stopped the respective analyses and performed a *local* approach - local in the sense that we analysed a reduced fitness landscape of sequences that are evolutionary close to the R5 and X4 consensus sequence. We introduced mutations only in the ten least conserved sequence positions of the R5 and X4 nt MSA.
The following ten nt positions of the R5 and X4 MSA are the weakest conserved positions and were thus selected for mutation (sorted from least to most conserved nt position):

- R5: 75, 64, 38, 37, 51, 31, 85, 14, 42, 29
- X4: 73, 38, 37, 64, 71, 31, 94, 33, 74, 29

Mutations in these positions of the R5 and the X4 MSA were most frequently observed and are thus assumed be biologically most relevant.
We performed analyses of the R5 and X4 populations, consisting of all possible four-, six-, eight-, and ten-point mutants, but with the mutations restricted to the most weakly conserved nt positions, i.e. each of the selected least conserved nt positions was mutated towards every possible nucleotide. This approach built populations of the size $n = 4^m$, with m being the number of mutated sequence positions and 4 being the number of different nucleotides.

- $m = 4 : n = 4^4 = 256$ sequences
- $m = 6 : n = 4^6 = 4,096$ sequences
- $m = 8 : n = 4^8 = 65,536$ sequences
- $m = 10 : n = 4^{10} = 1,048,576$ sequences

The mutated nucleotide sequences were translated into the corresponding amino acid sequences and their fitness was determined based on Equation 4.12.

Upon the inspection of the resulting number of aa changes, we observed first deviations between the R5 and X4 population. The analysis showed that mutations of the ten least conserved nt positions of the X4 nt sequence in general result in a higher number of aa changes, compared to mutations of the least conserved nt positions of the R5 nt MSA. The histograms in Figure 4.20 depict the counts of aa mutations observed in the R5 and X4 populations of all four-, six-, eight-, and ten-point mutants.

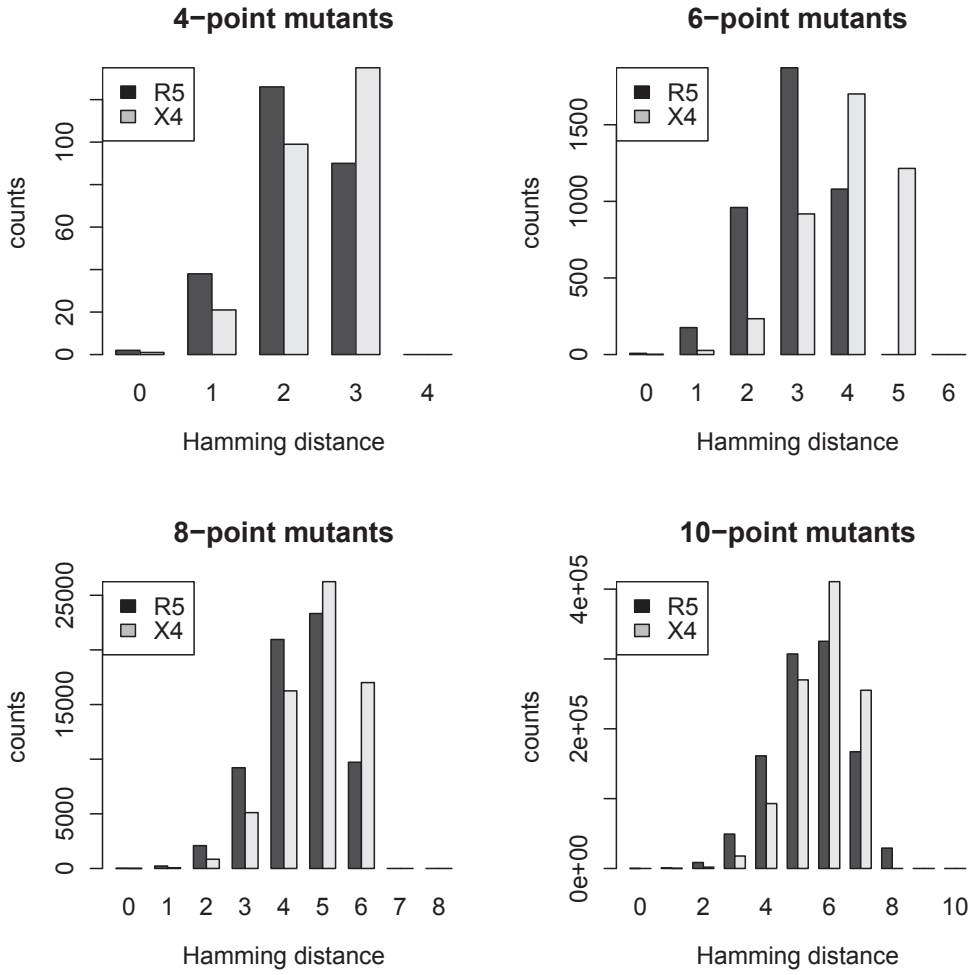


Figure 4.20.: **Histograms of amino acid Hamming distances**
The figures compare the aa Hamming distances of the R5 population (in black) and the X4 population (in grey) of the four-, six-, eight-, and ten-point mutants, compared to the respective consensus sequence. Alterations of the weakly conserved X4 nt positions in general result in higher numbers of aa mutations than alterations of the weakly conserved R5 nt positions.

The differences of the distributions of the number of aa mutations in the R5 and X4 population were analysed in R [121], using the χ^2 test [114]. The test statistics showed that the distribution of the number of aa mutations in the R5 and X4 populations were significantly different for the four-, six-, eight-, and ten-point nucleotide mutants (compare Figure 4.21).

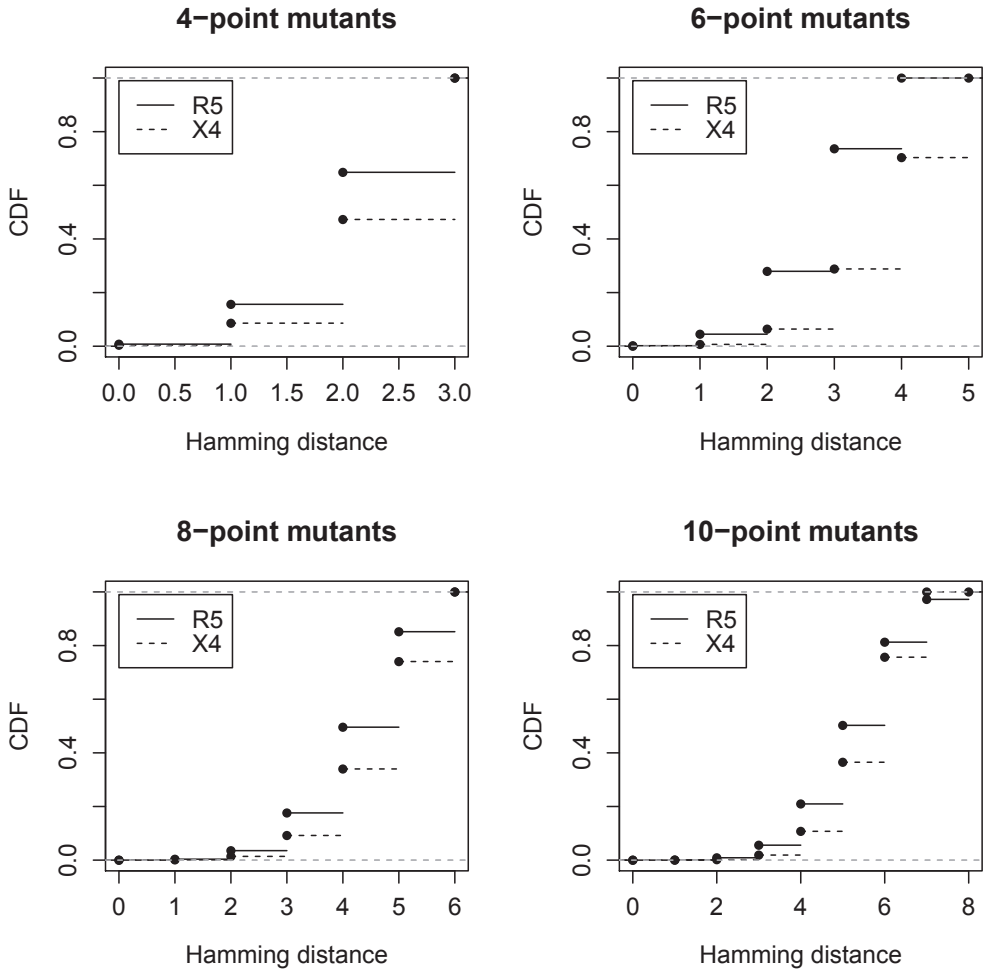


Figure 4.21.: **Cumulative statistics of amino acid Hamming distances**
The figures illustrate the analyses of the differences of the aa Hamming distances of the R5 and X4 sequences. The distributions of the Hamming distances were significantly different, χ^2 tests [114] resulted in the following statistics:
4-point mutants: $p - value = 0.0002$, 6-point mutants: $p - value < 0.0001$,
8-point mutants: $p - value < 0.0001$, 10-point mutants: $p - value < 0.0001$,
CDF: cumulative distribution function, $x \sim f(x)$, $\int_0^x f(y)\,\mathrm{d}y$

The differing numbers of aa mutations were a consequence of the differences of the weakly conserved nt sequence positions, resulting in different mutated codon positions. Though 50% of the less conserved nt positions were concordant between the R5 and X4 MSA (29, 31, 37, 38, 64), a detailed inspection of the remaining discordant positions led us to an interesting finding. Our analyses revealed that the first codon position dominated the ten weakly conserved positions of the X4 MSA, but not of the R5 MSA. The following weakly conserved nt positions occurred at the first codon position (positions sorted by weakness of conservation):

- R5: 64, 31, 85
- X4: 73, 37, 64, 31, 94

Three of the ten most weakly conserved nt R5 positions are first codon positions, in contrast to five of ten positions in the X4 data set. Analysing only the four-point mutants, we found three times the first codon position among the least conserved X4 nt positions (73, 37, and 64), but only one first codon position (64) among the four least conserved R5 nt positions.
A subsequent analysis revealed that respective X4 codons with weakly conserved first codon position in parallel contained weakly conserved second (74, 38) and third (75) codon positions. None of the three R5 codons with a weakly conserved first codon position (64, 31, 85) showed an additional weakly conserved second codon position, and only one of the three showed an additional weakly conserved third codon position (33).
This observation is quite remarkable, since due to the ambiguities of the aa code, a mutation in the first codon position leads to an aa change in all but one of the 61 amino acid codons. In consequence, mutations of the weakly conserved nt positions of the X4 nucleotide alignment led to more aa changes than mutations of the weakly conserved nt positions of the R5 nucleotide alignment.
This finding yielded a first indication that the aa sequence space of R5-tropic sequences is more dense and conserved, while the less conserved X4 sequences are wide-spread in sequence space.
We deepened this analysis and found another interesting aspect of our data. Mutations of the ten most weakly conserved R5 nt positions did not result in any stop codon, while mutations of only the four least conserved nt positions of the X4 nt MSA already introduced stop codons into the X4 aa sequences.
A detailed inspection of the nucleotide consensus sequences is illustrated in Figure 4.22. It revealed that the codon `XXA` is found in the weakly conserved positions 73 and 74 and also in the weakly conserved position 94 of the X4 MSA (an *X* marks the weakly conserved nt position selected for mutation). Furthermore, the combination of the weakly conserved positions 31 and 33 can lead to the codon pattern `XGX`. Matching those codons with the nt sequences of the three stop codons, `TAG`, `TAA`, and `TGA`, we recognised that a mutation of the `X` position of the pattern `XAA` towards `T` creates stop codon `TAA`. In addition, two nt mutations in `XXA` can lead to the stop codons `TAA` and `TGA`, and two mutations in the codon `XGX` generate the stop codon `TGA`.
For clarification, we want to stress that the observed stop codons are not directly present in the X4 consensus sequence, but occur by a mutation of the four to ten least conserved X4 nt alignment positions. The nt sequence of X4 is given below, the five codons with least conserved first codon position were highlighted in red:

```
TGT ACA AGA CCC AAC AAC AAT ACA AGA AAA AGT ATA CAT ATA GGA CCA GGG AGA
GCA TTT TAT ACA ACA GGA AAA ATA ATA GGA GAT ATA AGA CAA GCA CAT TGT
```

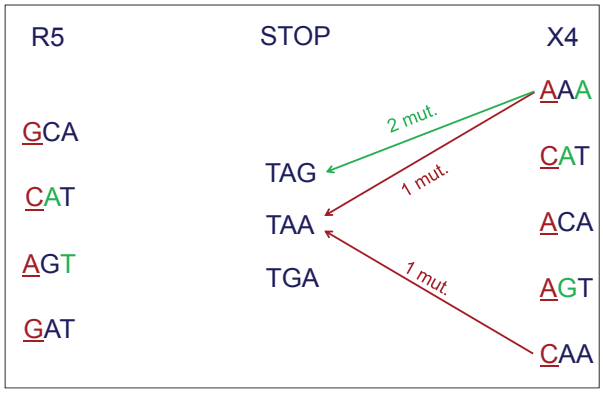
A corresponding analysis of the weakly conserved positions of the R5 nucleotide sequence at the first codon position led to the following pattern:

```
TGT ACA AGA CCC AAC AAC AAT ACA AGA AAA AGT ATA CAT ATA GGA CCA GGG AGA
GCA TTT TAT GCA ACA GGA GAA(C) ATA ATA GGA GAT ATA AGA CAA GCA CAT TGT
```

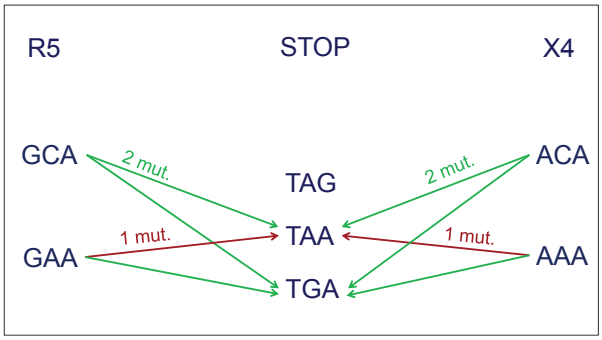
From this analysis, we learned that a weakly conserved first codon position in the R5 nt consensus sequence can not create a stop codon upon a mutation in any case. None of the three codons `XGT`, `XCA`, and `XAT`, which all show a weakly conserved nt in the first codon position, can result in a stop codon upon a mutation in a weakly conserved `X` position.
Though both position one and two of the codon `XAT` (nt positions 37 to 39) were weakly

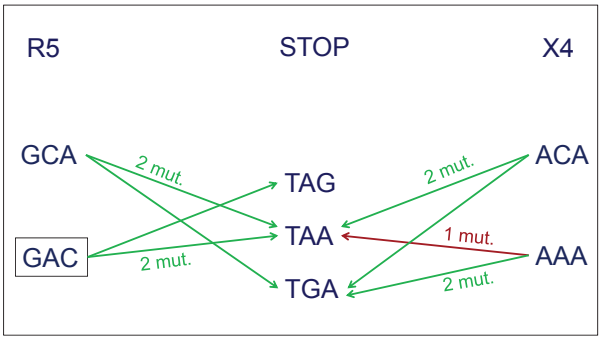
conserved, the nt in the third position, T, restricted the mutation of the codon into a stop codon, since a stop codon with T in the third position does not exists. No other codon of the R5 MSA comprised two weakly conserved nt positions.
Thus, mutations of the least conserved sequence positions frequently create stop codons in sequences derived from the X4 nt consensus sequence, but not in R5 derived sequences.

For a further validation of these results, we made a similar analysis, including all 105 sequence positions instead of focussing on the weakly conserved ones. We previously described in Section 4.3.2 that the R5 and X4 nt consensus sequences differ only in codon 22 (nt 64 to 66) and 25 (nt 73 to 75). Codon 22 of R5 is GCA, of X4 it is ACA. Since there is no stop codon with either G or A in the first position and positions two and three are identical, both codons yield the same probability to mutate into a stop codon. For codon 25, we have GAA in case of R5 and AAA in case of X4. Again, only codon position one differs, and there is no stop codon with either G or A in the first position. Following these observations, the distance to any stop codon is identical for both the R5 and X4 consensus sequence (compare illustration (b) of Figure 4.22).
In consequence, the complete consensus sequences of the R5 and X4 nt MSA did not give further indications whether the R5 or X4 sequence is more or less in danger to be mutated into a sequence with stop codon.
Finally, we did an analogue analysis, but we started from the aa instead of the nt consensus sequences. As described at the end of Section 4.3, a translation of the R5 nt consensus sequence does not exactly match the R5 aa sequence. The codon 25, GAA, translates into glutamic acid (E), while we find aspartic acid (D) at position 25 in the aa consensus sequence of R5, encoded by codon GAC. This differing sequence position is a result of the most weakly conserved nt position 75 (position three of codon 25). If we analyse the codon GAC instead of GAA, and compare the R5 codon GAC in position 25 to the corresponding codon AAA of the X4 consensus sequence, at least two mutations for codon GAC in R5 sequence are necessary to be translated into a stop codon (TAG or TAA), while the codon AAA in the X4 sequence needs only one mutation to be transformed into the stop codon TAA. The distance of both sequences to the third stop codon, TGA, is identical (compare illustration (c) of Figure 4.22).
In summary, the mutation of GAC to TAG and TAA in R5 could occur with the probability $2 \cdot \frac{1}{4} \cdot \frac{1}{4} = \frac{1}{8}$, while a mutation of the codon AAA into TAA in X4 occurs with twice the probability ($\frac{1}{4}$).
Thus, the direct comparison of the complete nt consensus sequences showed no differences, but upon a back-translation of the aa consensus sequences into nt sequences, we found the genetic distance of the X4 consensus sequence to a stop codon to be closer than the distance of the R5 consensus sequence.
In summary, the analysis of the ten most weakly conserved nt sequence positions revealed that the X4 codons are evolutionary closer to stop codons than the R5 sequences and are thus in higher danger to evolve into dead-end sequences. From the observed differences in the intra-codon conservation, we expected to observe an accumulation of stop codons in the X4 mutant population, resulting in many individuals with zero replicative fitness. Furthermore, we hypothesised that the X4 mutant sequences form a rugged and holey fitness landscape. In contrast, we did not expect to see any individuals of zero fitness in the R5 mutant population, provided that we mutate exclusively the ten least conserved



(a)





(b) (c)

Figure 4.22.: **Genetic distance to stop codons**
The figure illustrates the genetic distance of the most weakly conserved R5 and X4 codons to any of the three stop codons.
a) Mutations in the ten most weakly conserved R5 nt positions did not result in any stop codon, while mutations in the ten most weakly conserved X4 nt MSA positions frequently introduced stop codons into X4 sequences.
(colour coding:
red underlined: weakly conserved nucleotides (nts) at the first codon position, green: weakly conserved nts at the second or third codon position, blue: nts that are not among the ten most weakly conserved nts)
b) The R5 and X4 consensus nt sequences differ only in codon 22 and 25. A comparison of the two pairs of differing R5 and X4 codons showed equal evolutionary distances to stop codon sequences.
c) A back-translation of the R5 aa consensus sequence into the respective nt sequence changed the nt position 75 (position three of codon 25, emphasised by square). This *wobble* position of the consensus sequence indicated a farther evolutionary distance of the R5 sequence to stop codons.

sequence positions of the nt MSA. The underlying fitness landscape of the R5 mutant population was expected to be more continuous and without holes built by individuals of zero fitness.

An inspection of the fitness values of the R5 and X4 populations of four-, six-, eight-, and ten-point mutants gave a first hint to confirm our idea. We translated each nt sequence into an aa sequence and determined the corresponding fitness based on Equation 4.12. The fitness values were in the range of $[0.0, 1.0]$, with 0.0 representing sequences with stop codons and 1.0 representing the most fit consensus sequence of each population.
Figure 4.23 depicts histograms of the resulting fitness values of the R5 and X4 mutant populations. The figures show that aa changes decreased the fitness of the R5 mutants to a larger extend than the fitness of the X4 mutants. This was already observed in the sequence examples presented in Section 4.4.3. Due to the larger position specific aa probabilities of the R5 MSA and the higher conservation of the R5 consensus sequence, the R5 sequences are more sensitive to aa changes than the less conserved X4 sequence. A change in the highly conserved R5 consensus sequence decreases the fitness of the mutant to a larger extend.
This effect can also be seen by a direct comparison of the Figure 4.23 to the previous presented distribution of the aa Hamming distances (see Figure 4.20). Though the X4 mutant populations tend to accumulate a higher number of aa mutations upon mutation of the weakly conserved nt positions, the distribution of the fitness values indicates a shift of the X4 mutant fitness towards higher fitness values.
The X4 sequences showed a broader range of fitness values. The fitness values of the X4 four-point mutants for example extended to the complete interval $[0.0, 1.0]$, while the R5 four-point mutants did not drop below a fitness threshold of 0.28. The Figure 4.23 showed further the accumulation of stop codons upon mutations of the most weakly conserved X4 nt positions, resulting in high counts of individuals with zero fitness. The populations of eight- and ten-point mutants furthermore showed that mutations in the less conserved X4 consensus sequence in general lead to a decreased fitness reduction, except in the case that a stop codon is generated.
Using the Kolmogorow-Smirnov test [92, 143], we compared the distributions of the fitness values of the R5 and X4 mutant populations. Our analyses showed that the distributions of the fitness values are significantly different for the four-, six-, eight-, or ten-point mutant populations (see Figure 4.24).
It has to be noted that the absolute fitness values can only be used for a ranking of the replicative capacity within the respective R5 or X4 population, i.e. the deviations between the fitness values are only sound for an intra-population ranking. Since we lacked an *in vitro* method to compare the replicative fitness of the R5 and X4 sequences, an inter-population comparison of the fitness values is not possible with the presented method. The numerical deviations between the R5 fitness values $eFit_{R5}$ and the X4 fitness values $eFit_{X4}$ need a further validation.

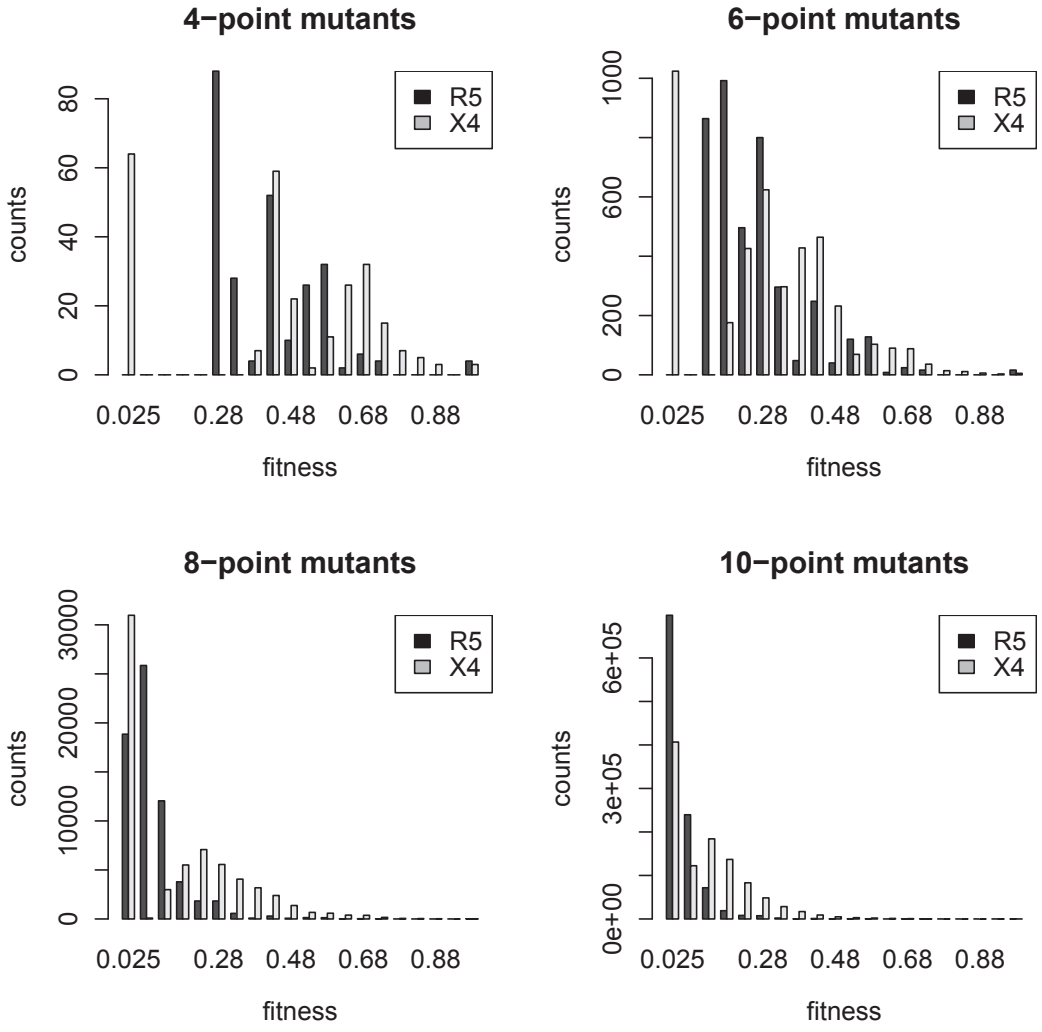


Figure 4.23.: **Histograms of population fitness**
The figures compare the replicative fitness of the R5 population (in black) and the X4 population (in grey) of the four-, six-, eight-, and ten-point mutants. Mutations in the conserved R5 consensus sequence decreased the fitness of the individuals to a larger extend than mutations in the less conserved X4 consensus sequence.

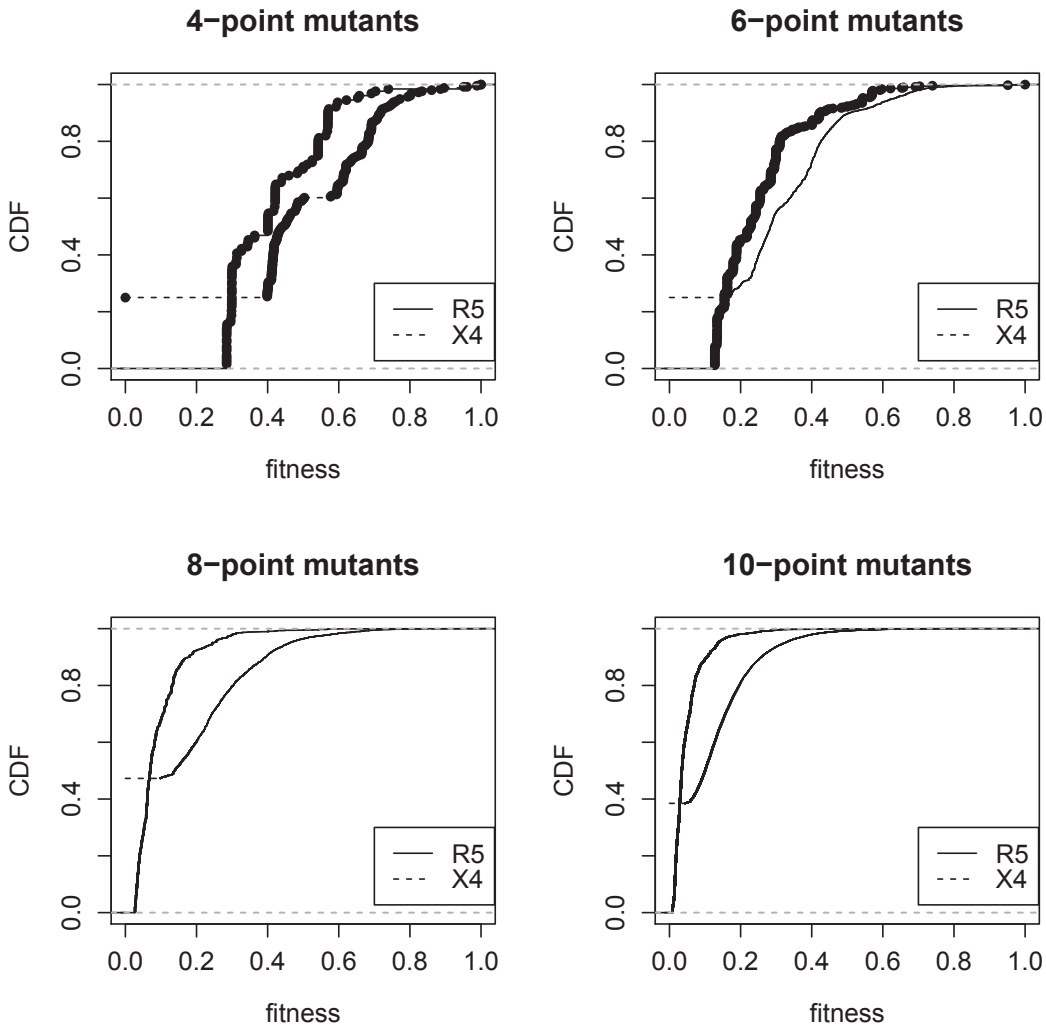


Figure 4.24.: **Cumulative statistics of fitness**
The figures illustrate the differences between the distributions of the fitness values of the R5 and X4 mutant populations. Kolmogorow-Smirnov [92, 143] tests confirmed that the distributions of the fitness values are significantly different:
4-point mutants: $p - value = 0.0001$, 6-point mutants: $p - value < 0.0001$,
8-point mutants: $p - value < 0.0001$, 10-point mutants: $p - value < 0.0001$,
CDF: cumulative distribution function, $x \sim f(x)$, $\int_0^x f(y)\,\mathrm{d}y$

Definition of a graph representation of the fitness landscapes

Analysing the aa Hamming distances and the distribution of the fitness values of the populations of four-, six-, eight-, and ten-point mutants, we found some first evidence for structural differences between the R5 and the X4 fitness landscape. We next examined the underlying local R5 and X4 fitness landscapes. Therefore, we used the Python graph analysis package *networkx* [120, 69] to translate the populations of four-, six-, eight-, and ten-point mutants into a graph representation of the R5 and X4 fitness landscape. The fitness values of the four-, six-, eight-, and ten-point mutants were determined based on equation 4.12.
The mutant networks were created based on the following rules:

- define a maximal nt Hamming distance threshold $max_{Hamming}$
- define a maximal fitness deviation max_{fitDev}
- create a vertex v_i for each sequence i
- add an edge between two vertices v_i and v_j if
  - $H(v_i, v_j) \leq max_{Hamming}$ and
  - $|v_i - v_j| \leq max_{fitDev}$ and
  - $eFit_{v_i} \neq 0.0$ or $eFit_{v_j} \neq 0.0$

Thus, two mutant sequences i and j were connected by an edge, if both the Hamming distance of the two nt sequences and the fitness deviation of the corresponding aa translations did not exceed a given threshold max_{fitDev}. The approach restricted to create an edge between two sequences i and j with zero replicative fitness, since mutants with zero replicative fitness by definition can not produce offspring and thus can not be mutated one into the other. In contrast, a link between a sequence $eFit_i \neq 0.0$ and a sequence $eFit_j = 0.0$ is allowed, since a replicative-competent sequence can be mutated into a sequence of zero fitness.
The size of the resulting networks can be calculated based on the number of sequences respective vertices of the network, $n = 4^m$ (again m being the number of mutated sequence positions). The maximal number of edges e of each network is $\frac{n(n-1)}{2}$. The term $n - 1$ addresses the fact that we do not allow self loops and the denominator of 2 is a consequence of the graphs being undirected. Table 4.1 presents the number of vertices and the maximal number of edges of the graph representations of the four-, six-, eight-, and ten-point mutant populations.

Table 4.1.: **Graph size**
The table gives a overview over the number of nodes and the maximal number of edges of the networks of four-, six-, eight-, and ten-point mutants.

m	$n = 4^m$	$e\frac{n(n-1)}{2}$
4	256	$3.3 \cdot 10^4$
6	4,096	$8.4 \cdot 10^6$
8	65,536	$2.1 \cdot 10^9$
10	1,048,576	$5.5 \cdot 10^{11}$

Fitness landscape of four-point mutants In the following paragraphs, we analysed the local networks of the R5 and X4 mutants with fitness values based on Equation 4.12. The mutated nt positions resembled the four, six, eight, or ten weakest conserved nt sequence positions of the respective data set. We started with a comparison of the R5 and the X4 network of all four-point mutants, using the network definition in Section 4.4.7.
Figure 4.25 illustrates the number of edges that were realised in the R5 and X4 network of four-point mutants. The distinct lines in the illustration represent four nt Hamming distance thresholds. The selected fitness deviation threshold, describing the fitness deviation between connected nodes, is represented on the x-axis. We used the following 28 fitness deviation thresholds: 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, $2.5 \cdot 10^{-2}$, seven subsequent equidistant intervals of $2.5 \cdot 10^{-2}$ up to a maximal fitness deviation of 0.2, and 16 subsequent equidistant intervals of 0.05 up to a fitness deviation of 1.0. Thus, the illustration presented in Figure 4.25 summarises the results of $4 \cdot 28 = 112$ different R5 respective X4 networks. The image presents the number of the realised edges. The maximal possible number of edges (32,640, compare Table 4.3) was realised only in the R5 network with a nt Hamming distance threshold of four and the maximal fitness deviation of 1.0. A Hamming distance of four and a maximal fitness deviation of 0.3 realised the majority of all possible edges in the R5 network (29,876). Using the same same definition, the X4 network contained about half of all possible edges (16,033). Even using the maximal fitness deviation threshold of 1.0 at a maximal Hamming distance of four, the X4 network only contained 30,624 of 32,640 edges. A detailed analysis showed that this is a consequence of the sequences with zero fitness, which were by definition not allowed to interconnect to each other.

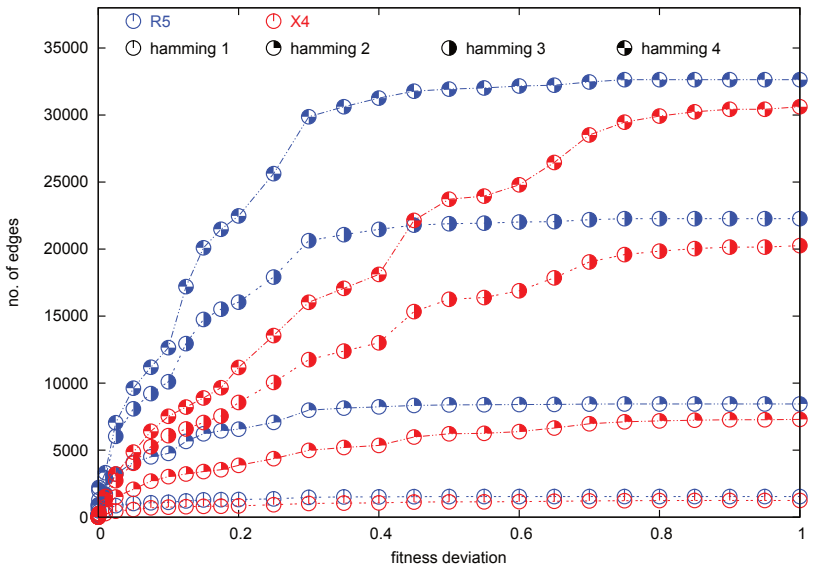


Figure 4.25.: **Number of edges in the R5 and X4 network of four-point-mutants**
The graphic depicts the number of realised edges in the R5 and the X4 network of the four-point-mutants. The distinct lines represent four nt Hamming distance thresholds. The selected fitness deviation threshold, describing the fitness deviation between connected nodes, is represented on the x-axis.
Identical nt Hamming distance and fitness deviation thresholds enabled more connecting edges in the R5 network than in the corresponding X4 network.

Thus, the same number of nt changes created a population of X4 mutants that were more isolated, while the R5 sequences created densely connected regions of homogeneous fitness. The last edges, that were only created upon the definition of the maximal fitness deviation, connected the most fit R5 consensus sequence to the least fit mutants of the population.
We next determined the maximal node degree of the networks. Since self-loops were restricted, the maximal possible node degree was 255. In the R5 network, a fitness deviation of 0.20 and an nt Hamming distance of four almost resulted in the maximal node degree (245). We observed a comparable maximal node degree of 244 in the X4 network at a fitness threshold of 0.4. The corresponding fitness deviation of 0.2 restricted the number of edges in the X4 network to two thirds of all possible edges (180). The exact maximal degree was reached almost in parallel for a fitness deviation of 0.4 for R5 and 0.5 for X4. Using a fitness deviation >0.4 and a consistent nt Hamming distance for both the R5 and X4 network, the R5 and X4 networks had an almost identical maximal degree.
Though mutations of the less conserved X4 consensus sequence in general lead to smaller fitness deviations than mutations of the R5 consensus sequence, the X4 population showed a broader range of fitness values. The X4 fitness values spread along the complete interval $[0.0, 1.0]$, while the fitness of the R5 four-point mutants did not drop below a fitness threshold of 0.28 (though the fitness Equation 4.12 enabled R5 fitness values along the complete interval $[0.0, 1.0]$ as presented in Figure 4.23).

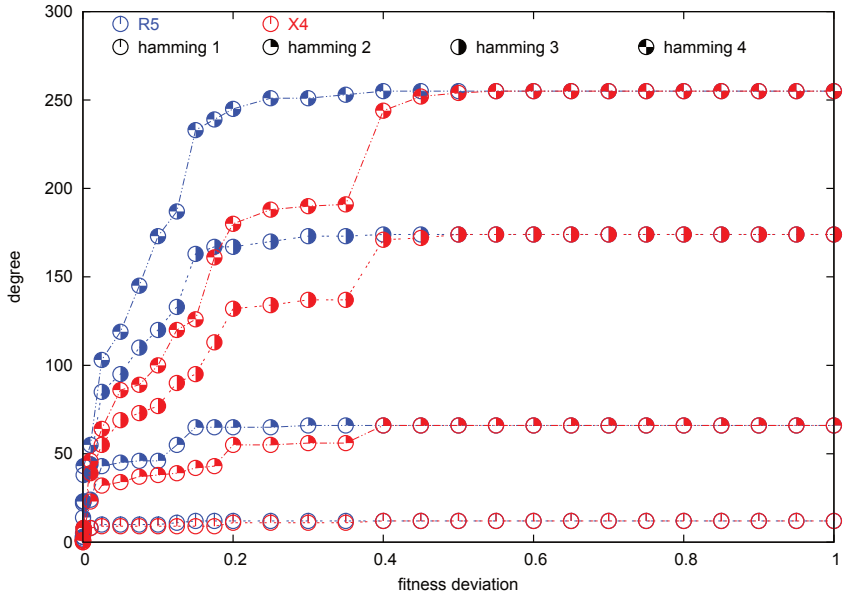


Figure 4.26.: **Maximal node degree in R5 and X4 network of four-point-mutants**
The graphic depicts the maximal node degree that was found in both the R5 and the X4 network of the four-point mutants. The distinct lines represent four nt Hamming distance thresholds. The selected fitness deviation threshold, describing the fitness deviation between connected nodes, is represented on the x-axis.
At low fitness deviations, the maximal node degree of the R5 network was always higher than the maximal node degree of the X4 network at the same Hamming distance. A fitness deviation >0.4 approximated an equal maximal node degree for the R5 and X4 networks.

Further investigations showed that this difference between the R5 and the X4 fitness values was a consequence of the sequences with zero replicative fitness in the X4 population of four-point mutants. An exclusion of those sequences would limit the range of the X4 fitness values to the interval $[0.39, 1.0]$.
An analysis of the number of network components fit into the line of observations (compare Figure 4.27). The maximal number of network components is the number of nodes in the network, which was 256 for the four-point mutants. Independent of the selected nt Hamming distance threshold, the R5 network merged fast into one component. At a fitness deviation of 0.1, one large and one small component were formed. Increasing the fitness deviation threshold to 0.25 joined the small component and the large network component. At the same time, the X4 network consisted of ≥ 65 separated components, and the components were stable until a fitness deviation of 0.35. For fitness deviations >0.35, new edges were created and at an nt Hamming distance ≥ 3, the networks merged into one component. At an nt Hamming distance of two, a fitness deviation of ≥ 0.6 was required to merge the components. Finally, at an nt Hamming distance of one, the maximal fitness deviation of 1.0 (i.e. no fitness restriction) was obligatory to merge the components into one large network.
We already discussed this peculiarity of the X4 network to be a result of the mutant sequences with zero replicative fitness. In the case of the R5 population, the merging network structures again indicated a homogeneous distribution of the R5 fitness values.

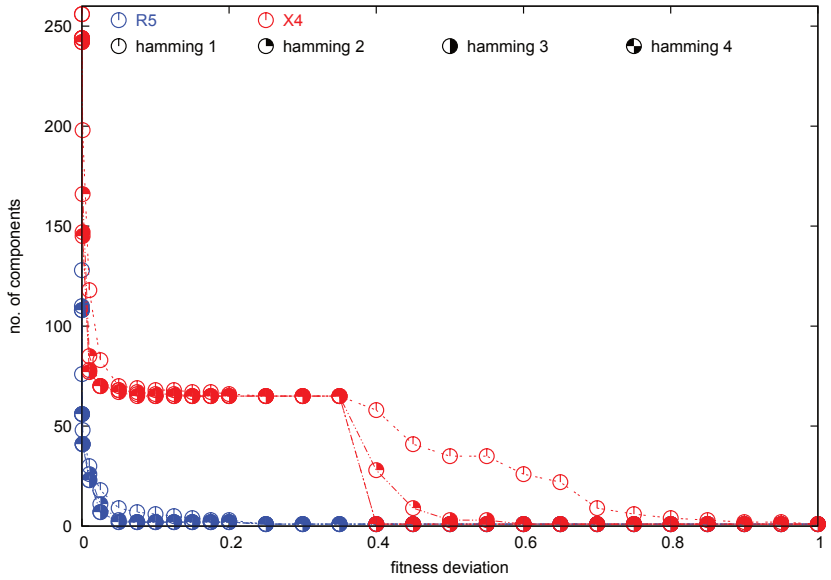


Figure 4.27.: **Number of components in R5 and X4 network of four-point-mutants**
The graphic depicts the number of network components created from the R5 and X4 population of four-point-mutants. Independent of the nt Hamming distance threshold, the R5 networks converged into one large component at low fitness deviations, while the X4 networks stably formed ≥ 65 independent components. For a fitness deviation >0.35, the X4 components started to merge into one component.

The number of network components is highly correlated with the number of isolated nodes of the respective network. A corresponding analysis showed almost identical results (data not shown).

In Figure 4.28, we analysed the degree of the vertex representing the consensus sequence. For an nt Hamming distance of one and for a fitness deviation ≥ 0.75, the R5 consensus sequence connected to all sequences of the population (represented by a node degree of 255). This fitness threshold is consistent for the nt Hamming distances of two and three. The steep increase of the node degree for fitness deviations of 0.4 to 0.75 showed that mutations in the highly conserved R5 consensus sequence remarkably decreased the replicative fitness of the mutant sequences.
In comparison to the previous analysis of the number of network components (compare Figure 4.27), this observation confirmed that the most fit consensus sequence was only connected to a few sequences of similar fitness and was clearly marked off from the majority of the mutated individuals. Thus, the consensus sequence builds a peak in the fitness landscape, while the mutated sequences build a plateau or cloud of sequences of lower fitness. This plateau of sequences of lower fitness is also indicated by Figure 4.26, which showed that a high node degree of 245 in the R5 network was reached at a low fitness deviation of 0.20. Thus, many sequences reside in a region of comparable fitness, separated by fitness deviations ≤ 0.20.
In the X4 network, the drop of the node degree of the consensus sequence was less pronounced, presumably due to the weaker sequence conservation in general. In addition, the X4 sequences of high fitness are more interconnected than the R5 sequences of high fitness and thus have a higher node degree. A close inspection of the X4 MSA showed that this was a consequence of the observed wobble positions of the MSA, populated by almost identical counts of two (positions 11 and 22) or four (position 25) different aa (compare aa counts presented in Section 4.4.2).
In contrast to the higher node degree of the X4 consensus sequence, resulting from a multitude of edges to the neighbouring X4 mutants of high fitness, a biologically unsuitable fitness deviation threshold of 1.0 was required to achieve a maximal node degree of the consensus sequence. The fitness threshold of 1.0 represents a connection of the X4 fitness peak to all sequences of the population, including the holes of zero fitness. A comparison of the average node degree of the R5 and X4 network with the degree of the consensus sequences showed a lower degree than the average node degree for the consensus sequences. This difference in the degree was especially pronounced for fitness deviations ≤ 0.75 for the R5 network and ≤ 0.6 for the X4 network (data not shown). This result confirmed the observation that the fitness peaks are weakly connected to the majority of the sequences of the mutant population.

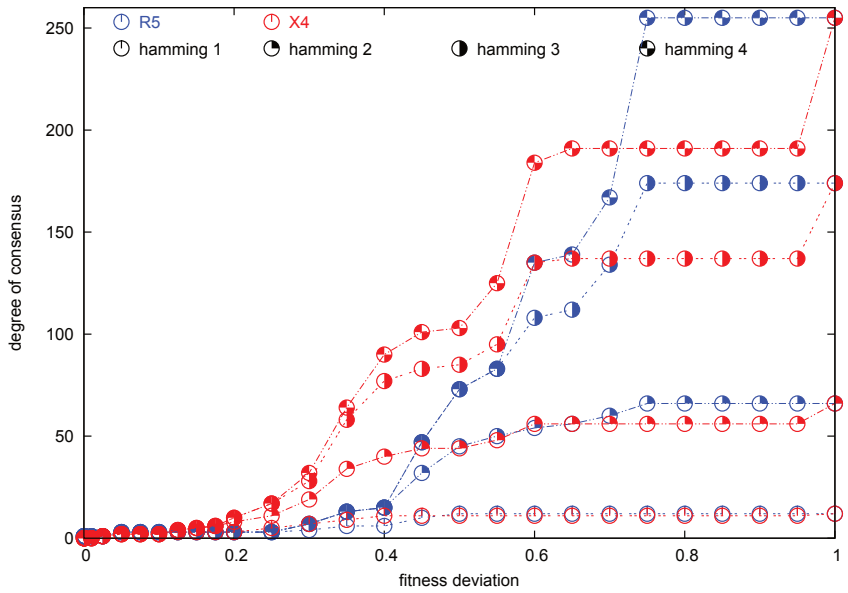


Figure 4.28.: **Degree of the consensus sequence in R5 and X4 network of four-point mutants**
For fitness deviations of ≤ 0.4, the R5 consensus sequence showed a low node degree and was connected to only a few high fit sequences. A fitness threshold of 0.75 connected the R5 consensus sequence to all mutated individuals.
In the case of the X4 consensus sequence, the degree of the consensus sequence was higher at smaller fitness thresholds, but the X4 consensus sequence could only reach the maximal degree at a fitness deviation of 1.0.

In the final analysis of the networks of all four-point mutants, we determined the node degree of the sequences of minimal fitness. In general, we found a higher degree for the least fit R5 sequences and than for the least X4 sequences (compare Figure 4.29. At the maximal nt Hamming distance of four and a fitness deviation of 0.4, the least fit R5 sequences were connected to 245 of 256 sequences of the network. In contrast, the 64 X4 mutants with zero fitness were still isolated at this level of fitness deviation. A transition that was comparable to the increase in the node degree observed for the R5 network at a fitness deviation of 0.4 could be observed at a fitness deviation threshold of ≥ 0.75 for the X4 network. Using this threshold, the node degree increased to 174. Since our network definition restricted sequence mutants of zero fitness to interconnect, the degree of the nodes of minimal fitness in the X4 networks was lower.
The comparison of the R5 and the X4 fitness landscape of four-point mutants revealed that the R5 fitness landscape contains a few sequences of high fitness, that result in a steep fitness decay upon the introduction of mutations, levelling off into a non-zero fitness plateau built by a set of least fit (but $\neq 0.0$) sequences. In contrast, the X4 consensus sequence fitness peak is surrounded by a number of mutated variants with small fitness deviations and the fitness decrease close to the peak is less steep. On the other hand, the X4 fitness landscape contains fitness holes built by 64 sequences of zero fitness.

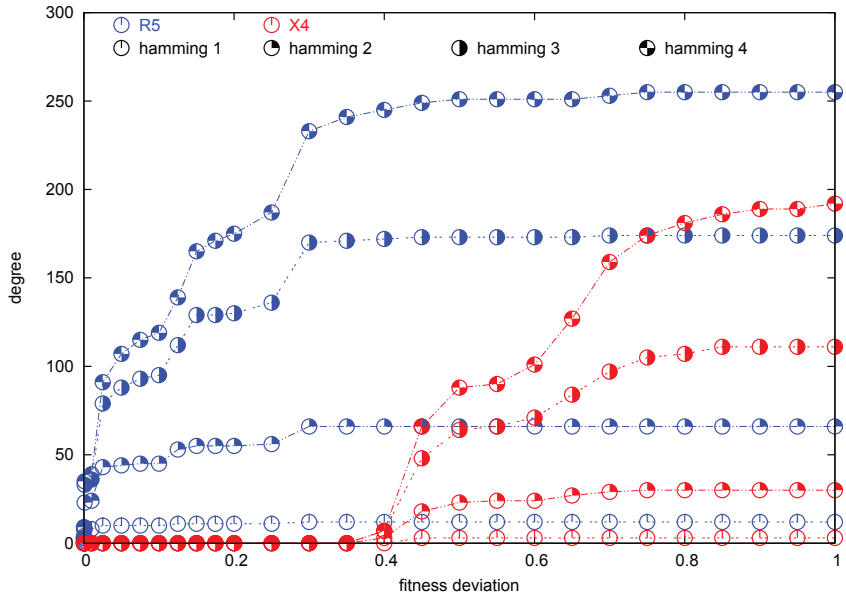


Figure 4.29.: **Degree of least fit sequences in R5 and X4 network of four-point mutants** At an nt Hamming distance of four and a fitness deviation of 0.4, the R5 sequences of minimal fitness achieved a node degree of 245. The least fit X4 mutants were still isolated at this level of fitness deviation and were connected to the central X4 network at a fitness threshold ≥ 0.75.

Fitness landscape of six-point mutants After a detailed analysis of the fitness landscapes built by the four-point mutants, we examined the fitness landscapes described by the populations of six-point mutants. Corresponding analyses of the R5 and X4 mutants confirmed the previous observations of the fitness landscapes of the four-point mutants and did not reveal additional findings. Thus, we only discuss two exemplary analyses.
According to Table 4.3, the networks contained 4,096 vertices and at most 8,386,560 edges. Figure 4.30 illustrates the number of edges that were realised in the networks, depending on the selected nt Hamming distance and the fitness deviation threshold. The maximal number of edges was created only in the R5 network (of an nt Hamming distance of six), with the majority of all possible edges (7,455,424) already being realised at a fitness deviation ≥ 0.3. Our findings confirmed the observation of a majority of homogeneous fitness values on the one hand, and the distinct separation of a few highly fit R5 sequences on the other hand, which were completely connected to all sequences only at fitness deviations ≥ 0.9.
In contrast to the R5 network, the X4 network contained approximately two thirds of all possible edges (5,384,534) at a fitness deviation level of 0.3 and at an nt Hamming distance of six. Furthermore, 523,776 edges ($\sim 6\%$ of the possible edges) were not realised in any X4 network, which was again a result of the restriction to connect two vertices that both carry sequences of zero replicative fitness.

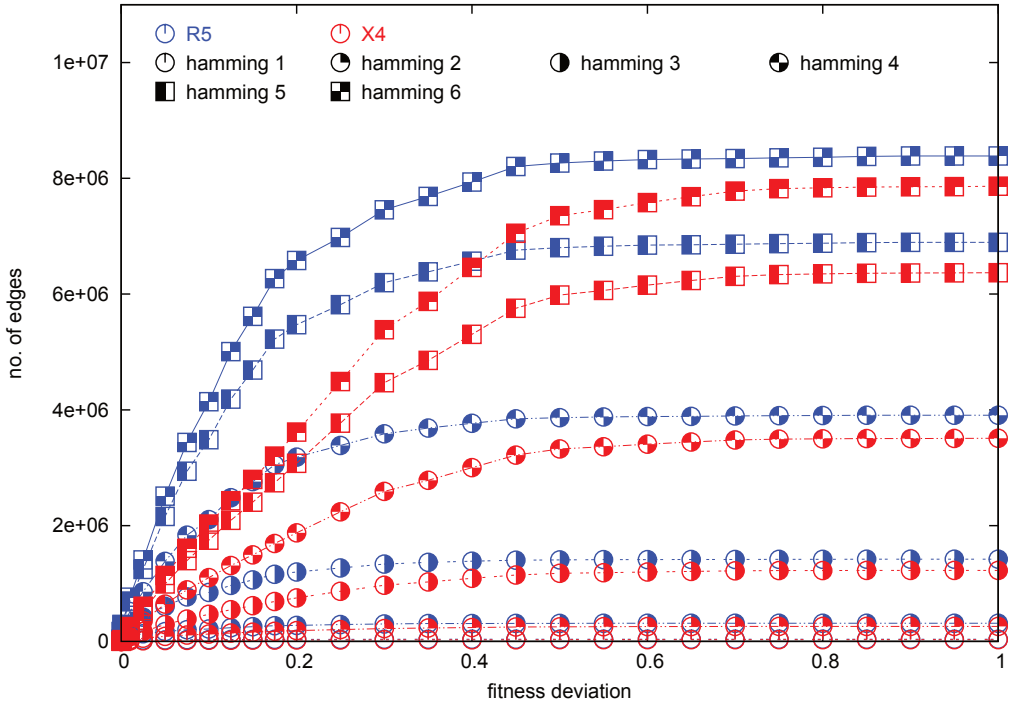


Figure 4.30.: **Number of edges in R5 and X4 network of six-point-mutants**
The graphic depicts the number of realised edges in the R5 and the X4 networks of six-point-mutants, depending on the selected nt Hamming distance (different lines) and the fitness deviation threshold (x-axis).
Most of the edges of the R5 network are already created at a fitness deviation ≥ 0.3; the maximum number of edges was realised at a fitness deviation ≥ 0.9 at an nt Hamming distance of six. In contrast, the X4 network realised at most two-thirds of all possible edges.

Figure 4.31 illustrates the maximal node degree that was observed in the networks of six-point mutant. In the case of the R5 network, a fitness deviation of 0.15 resulted in a maximal node degree of 3,679 ($\sim$ 90% of all possible edges), while a comparable transition in the X4 network was observed at a fitness threshold of 0.25 (3,670). A fitness deviation of 0.15 restricted the number of edges of the X4 network to 2,517 ($\sim$ 61.5%). The maximal node degree of 4,095 was reached almost in parallel for a fitness deviation of 0.45 in the R5 network and a fitness deviation of 0.55 in the X4 network.

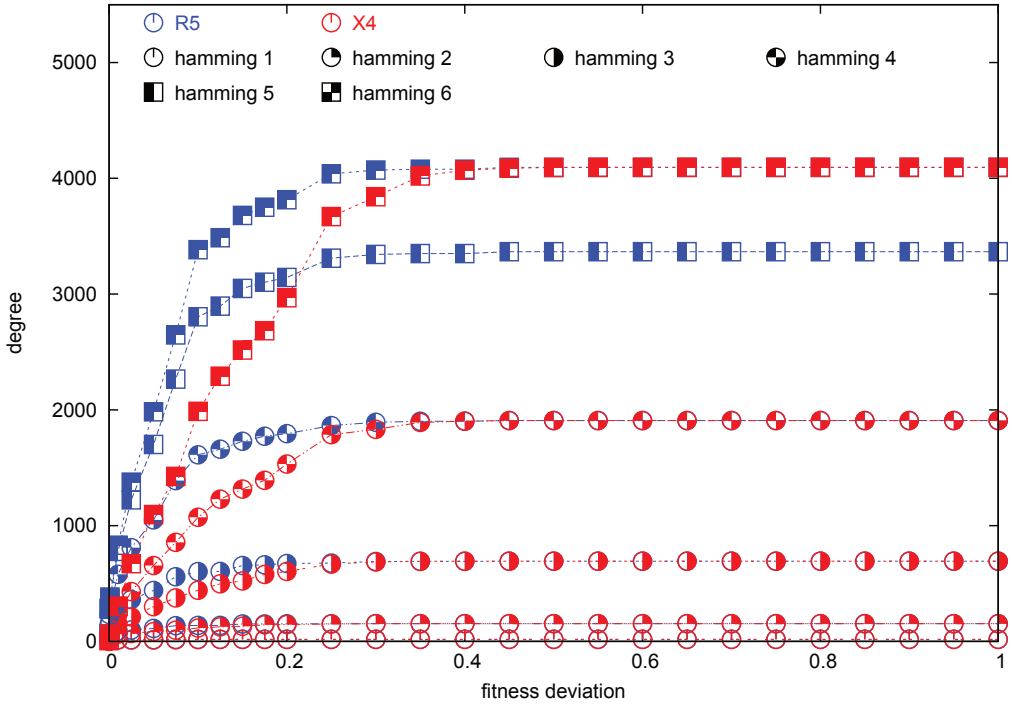


Figure 4.31.: **Maximal node degree in R5 and X4 network of six-point-mutants**
The graphic depicts the maximal node degree in the R5 and X4 networks of six-point mutants. At lower fitness deviation thresholds ≤ 0.4, the maximal node degree of the R5 network was higher compared to the X4 network. At a Hamming distance of six, the exact maximal degree was realised for a fitness deviation of 0.45 in the R5 and 0.55 in the X4 network.

In summary, the analyses of the six-point mutants did confirm the analyses of the four-point mutants, but did not contribute new findings. The observed network transitions are more regular upon changing fitness thresholds, illustrated by more smooth curves.
An inspection of the weakly conserved sequence positions revealed that the four most weakly conserved sequence positions of the X4 data set already comprised three different first codon positions. The six-point mutants included four and the eight-point mutants all five weakly conserved first codon positions of the X4 MSA. In case of the R5 MSA, the four-point mutants contained one first codon position, the six-point mutants two and the eight-point mutants all three weakly conserved first codon positions. Due to this observation, we neither expected remarkably different results for the eight- nor for the ten-point mutant networks and thus, we did not further deepen the analyses of this kind of mutant networks.

Definition of evolutionary networks

In the first network approach, we observed the influence of the nt Hamming distance and of a range of fitness deviation thresholds on the structure of the R5 and X4 mutant networks. In this section, we used an alternative definition for the construction of the networks of four-, six-, eight-, and ten-point mutants. We intended to analyse beneficial paths of evolution, and thus we applied directed networks. The networks were constructed as follows: for every individual of the population, we created one vertex, containing information about the nt sequence and the replicative fitness of the individual (based on Equation 4.12). Two vertices were connected by a directed edge, if the Hamming distance of the nt sequences was one (since the mutant sequences in each population were unique, no two sequences had a Hamming distance of zero). Furthermore, the edges were defined to start at the node with the lower fitness and point towards the node with the higher fitness. If both nodes had the same fitness, two parallel edges pointing into opposite directions were created to connect the nodes.
In summary, the evolutionary networks were defined in the following way:

- create a vertex v_i for each sequence i
- add a directed edge pointing from vertex v_i to vertex v_j if
  - $H(v_i, v_j) == 1$ and
  - $eFit_{v_i} \leq eFit_{v_j}$ and
  - $eFit_{v_i} \neq 0.0$

This approach again restricted the construction of links between two vertices representing sequences with zero replicative fitness. Links pointing from a sequence of non-zero fitness to a sequence of zero fitness are allowed, since sequences with replicative fitness $\neq 0.0$ could replicate and generate offspring with zero fitness, but not vice versa.
Using this definition, we created finite directed cyclic graphs of the R5 and the X4 mutant population. Cycles can only contain nodes that represent sequences with identical replicative fitness (i.e. different nt sequences that were translated into the same aa sequence). Self-loops did not contribute additional information to this analysis, thus, they were restricted to keep the graphs more simple. In contrast to the previous approach, this definition created only one network for each R5 and X4 mutant population.
The resulting graphs enabled us to follow paths of beneficial or neutral evolution, represented by an increasing or steady fitness along any path through the network. Analyses of the network structure enabled us to determine difference between the underlying fitness landscapes of the R5 and X4 mutant population and to find deviations in the length of the evolutionary pathways. Table 4.2 summarises the analyses of the R5 and X4 networks for the four-, six-, eight-, and ten-point mutant populations.

Evolutionary networks of four-point mutants We started with an analysis of the evolutionary networks of the R5 and X4 four-point mutants. The directed networks were both very sparse. In the case of R5, 1,664 edges were realised, representing 2.549% of all possible edges. The average node degree was 13, which was also the degree of each of the 256 nodes. Both the average in- and out-degree of the R5 network were 6.5.
The R5 network comprised two nodes of the maximal fitness of 1.0, representing the

consensus master sequence and an alternative nt mutant that consistently was translated into the aa consensus sequence. The in-degree of both R5 nodes of maximal fitness was one, representing the possible mutational pathway from one node to the other, exchanging only one nucleotide. The out degree was 12 for both nodes, resembling the four mutated nt positions and the three alternative nucleotides at each position.
As we expected, the R5 population did not contain any sequence of zero fitness. The least fitness value was 0.285 and represented four different nt sequences that all translated into the same aa sequence. The evolutionary network of the X4 mutants contained 1,248 edges, which represents a fraction of 1.911% of the theoretic maximum. The average node degree was 9.75, with an average in-degree of 4.875 and an average out-degree of 4.875. A maximal degree of 12 could be observed for 192 nodes, and the remaining 64 nodes had an (in-)degree of only three. The X4 network comprised only one node of the maximal fitness of 1.0, which in consequence had an in-degree of zero and an out-degree of 12. No silent nt mutation was possible for the X4 consensus master sequence.
We found 64 sequences of zero fitness among the X4 mutants, all containing at least one stop codon in the translated aa sequence. Only three directed paths led into each of these nodes and stopped there. The three pathways represented the three alternative nt at the respective sequence position.
This first analysis showed that the R5 network contained more edges than the X4 network, an observation that was already made for the previously described networks. An inspection of the underlying populations revealed that the differences between the R5 and the X4 network resulted from the 64 sequences of zero replicative fitness in the X4 population.
We next analysed the shortest paths in the networks. In the R5 network, the average length of all 35,928 reachable shortest paths (of length 1.0 to 6.0) was 2.676. In contrast, the number and average shortest path length in the X4 network was slightly decreased. We found 24,144 reachable shortest paths of length 1.0 to 5.0, with an average length of 2.583. Though the R5 network in general had a higher node degree than the X4 network, the shortest paths in the R5 network were longer. Analyses of single shortest paths of the R5 network showed that difficult mutational pathways were evolutionary reachable via detour and thus increased the shortest path length. In contrast, difficult evolutionary paths in the X4 network showed a tendency to become disconnected at some vertex. Thus, the shortest paths lengths in the X4 network were shorter (since long shortest paths more often were not reachable).
We next focussed the shortest paths analyses to those paths that connected a sequence of minimal fitness to a sequence of maximal fitness (min-max path), following the shortest mutational pathway to evolve from the least to the most fit sequence (or vice versa). The R5 network contained eight possible min-max paths, each of a length of 4.0, connecting any of the four nodes of minimal fitness to the two nodes of maximal fitness. In the X4 network, with one sequence of maximal fitness and 64 sequences of minimal fitness, we found 52 possible shortest paths of length 1.0 to 4.0. The network restriction to nt Hamming distances of one blocked 12 of the possible min-max pathways. The average shortest path length of 4.0 for the min-max paths was higher in the R5 network, than the average shortest path length of 3.135 for the 52 min-max paths in the X4 network.
Though the general shortest path length of the R5 and X4 network only varied by 3.6% (R5: 2.676 vs. X4: 2.583), the length of the shortest min-max paths connecting sequences of minimal to sequences of maximal fitness varied by $\sim$ 27.6% (R5: 4.0 vs. X4: 3.135)

between the R5 and X4 population.
This observation indicated a faster evolution of sequences of minimal fitness into sequences of maximal fitness in the X4 population, and gave further evidence to our hypothesis that the X4 population is evolutionary closer to sequences with minimal fitness than the R5 population.
This idea gained further support by an analysis of the maximal and minimal node betweenness, a value that represents the rate of all shortest paths that pass through a node. The measure describes the importance of the central nodes of a network as well as the network extension (i.e. for a network of sequences the extension in sequence space).
The maximal betweenness was $2.5 \cdot 10^{-3}$ for the R5 network and $2.1 \cdot 10^{-3}$ for the X4 network. These numbers indicated a slightly reduced importance of the central nodes of the X4 network as well as a larger expansion in sequence space. This observation was confirmed by the minimal betweenness of $2.9 \cdot 10^{-4}$ for the R5 and 0.0 for the X4 network. Thus, the R5 network was more central and dense and contained more evolutionary paths that traversed through the central nodes, while the X4 network was further extended in sequence space.
An analysis of the network closeness, which is an alternative measure of node centrality, further stressed this point. The maximal closeness of 0.332 in the R5 network was higher than the maximal closeness of 0.315 in the X4 network, and also the minimal closeness of the R5 network was higher than the minimal closeness of the X4 network $4.0 \cdot 10^{-3}$ vs. 0.0). In summary we found that the R5 network of all four-point mutants was more dense and concentrated in a smaller region in sequence space. The betweenness and closeness of the R5 network was higher, indicating more important central nodes. Evolutionary pathways between the R5 mutants in general were short (avg. length of 2.676), but paths that connected sequences of maximal fitness to sequences of minimal fitness were as long as 4.0.
In contrast, the network of the X4 mutant population was more sparse and extended in a larger region of sequence space. The X4 network enabled less evolutionary paths connecting the sequences of the population, and the possible pathways in general were slightly shorter (average length of 2.583), presumably due an interruption or blocking of longer evolutionary paths. Pathways from sequences of minimal to sequences of maximal fitness traversed on average 3.135 edges. A comparison of the average shortest path length and the shortest min-max path length indicated a reduced shortest min-max path length in the X4 network. Thus, a direct comparison of the fitness landscapes of the R5 and X4 four-point mutants supported our hypothesis that the pathways between sequences of maximal fitness and sequences of minimal fitness are evolutionary shorter in the X4 network than in the R5 network.

Table 4.2.: **Network measures of R5 and X4 mutant network**
The table gives a summary of the network analyses. The number of the sequences of minimal and maximal fitness is given in parenthesis (obs.).

network	measure	R5	(obs.)	X4	(obs.)
4-p. mut.	no. of edges	1,664		1,248	
6-p. mut.	no. of edges	46,112		29,184	
8-p. mut.	no. of edges	934,264		495,360	
4-p. mut.	avg. degree	13.00		9.75	
6-p. mut.	avg. degree	22.52		14.25	
8-p. mut.	avg. degree	28.51		16.54	
4-p. mut.	min. fit	$2.8 \cdot 10^{-1}$	(4)	0.00	(64)
6-p. mut.	min. fit	$1.3 \cdot 10^{-1}$	(48)	0.00	(1,024)
8-p. mut.	min. fit	$2.56 \cdot 10^{-2}$	(32)	0.00	(30,976)
4-p. mut.	avg. shortest path	2.68		2.58	
6-p. mut.	avg. shortest path	4.03		3.74	
8-p. mut.	avg. shortest path	5.26		5.12	
4-p. mut.	avg. shortest min - max path	4.00		3.14	
6-p. mut.	avg. shortest min - max path	5.75		4.64	
8-p. mut.	avg. shortest min - max path	7.75		6.04	
4-p. mut.	not reachable min - max paths	0		12	
6-p. mut.	not reachable min - max paths	0		192	
8-p. mut.	not reachable min - max paths	0		9,504	
4-p. mut.	max. betweenness	$2.49 \cdot 10^{-3}$		$2.10 \cdot 10^{-3}$	
6-p. mut.	max. betweenness	$1.83 \cdot 10^{-4}$		$9.43 \cdot 10^{-5}$	
8-p. mut.	max. betweenness	$1.93 \cdot 10^{-5}$		$9.36 \cdot 10^{-6}$	
4-p. mut.	min. betweenness	$2.92 \cdot 10^{-3}$		0.00	
6-p. mut.	min. betweenness	$5.93 \cdot 10^{-5}$		0.00	
8-p. mut.	min. betweenness	$2.29 \cdot 10^{-6}$		0.00	
4-p. mut.	max. closeness	$3.32 \cdot 10^{-1}$		$3.15 \cdot 10^{-1}$	
6-p. mut.	max. closeness	$2.22 \cdot 10^{-1}$		$2.10 \cdot 10^{-1}$	
8-p. mut.	max. closeness	$1.67 \cdot 10^{-1}$		$1.58 \cdot 10^{-1}$	
4-p. mut.	min. closeness	$3.92 \cdot 10^{-3}$		0.00	
6-p. mut.	min. closeness	$4.09 \cdot 10^{-3}$		0.00	
8-p. mut.	min. closeness	$1.23 \cdot 10^{-4}$		0.00	

Evolutionary networks of six-point mutants Analogue to the networks of four-point mutants, we analysed the R5 and X4 network of all six-point mutants. A summary of the network measures is given in Table 4.2.
The networks consisted of $4^6 = 4,096$ nodes. While 46,112 edges were realised in the R5 network, the X4 network only contained 29,184 edges. The higher density of the R5 network was reflected in a higher average node degree of 22.52, compared to 14.25 in the X4 network.
The average shortest path length was 4.03 (path length 1.0 to 8.0) in the R5 network, and 3.74 in the X4 network (paths length 1.0 to 7.0). Thus, the shortest paths in the X4

network are 7.75% shorter than in the R5 network, though the R5 network contains more edges and a higher average node degree.
An analysis of the sequences of minimal and maximal fitness showed that the R5 population contained eight individuals of maximal fitness (1.0), and 48 individuals of a minimal fitness of 0.13. The X4 population contained only one sequence of maximal fitness (1.0), but 1,024 individuals of a minimal replicative fitness of 0.0. Thus, the highly fit R5 sequences can integrate a number of different neutral nt mutations without a fitness loss. While the X4 population still contains only one sequences of maximal fitness, it accumulated further sequences of zero fitness, compared to the X4 population of four-point mutants.
All 384 possible shortest pathways between the eight sequences of maximal fitness and the 48 sequences of minimal fitness were reachable in the R5 network, while 192 of the possible 1,024 shortest pathways between the sequence of maximal and a sequence of minimal fitness in the X4 network could not be realised. The resulting shortest min-max path length varied between 6.0 and 7.0 in the R5 network and between 1.0 and 6.0 in the X4 network. On average, the shortest mutational pathway between a sequence of maximal and a sequence of minimal fitness was 5.75 in the R5 and 4.64 in the X4 network. Though the general shortest paths length only varied by 7.75% between the R5 and X4 network, the shortest min-max paths showed in a length difference of 23.92% and were again significantly shorter in the X4 network than in the R5 network.
The comparison of the centrality measures indicated a more condensed R5 network with a higher importance of the central nodes (maximal betweenness of $1.83 \cdot 10^{-4}$ and maximal closeness of 0.22), in contrast to a less central X4 network (maximal betweenness of $9.43 \cdot 10^{-5}$ and maximal closeness of 0.21.
Thus, the results of the R5 and X4 network of the six-point mutants again confirmed that the R5 network is more central, while the decreased centrality measures of the X4 network indicate a less concentrated around its centre. Furthermore, the shortest paths analyses of all shortest paths in comparison to the min-max shortest paths further confirmed our hypothesis of a closer proximity of sequences of maximal fitness to sequences of minimal fitness in the X4 sequence population.

Evolutionary networks of eight- and ten-point mutants Analyses of the evolutionary networks of the eight-point mutants revealed that the observed differences that we found for the networks of four- and six points mutants of the R5 and the X4 population were also present in the population of all eight-point mutants. The R5 population contained eight sequences of maximal fitness and the X4 population two individuals. We observed a minimal fitness of 0.0256 for 32 R5 mutants, while the minimal fitness in the X4 population was again zero, representing 30,976 mutants with stop codons.
Our analyses indicated a more dense R5 network, containing almost twice as many edges as the X4 network (934,264 edges versus 495,360 edges). Despite this difference, the general shortest path length of 5.26 edges in the R5 network and of 5.12 in the X4 network were comparable. In contrast, the length of the shortest evolutionary pathways between a sequence of minimal fitness and a sequence of maximal fitness differed remarkably from these numbers. An average shortest min-max pathway traversed 7.75 edges in the R5 network, but only 6.04 edges in the X4 network. Thus, the evolutionary distance of a sequence of maximal fitness to sequence of minimal fitness was in general 28.31% closer in the X4 network than in the R5 network. In addition, 9,504 (15,34%) of the 61,952 possible

evolutionary pathways between a sequence of minimal and a sequence of maximal fitness in the X4 network were not reachable, while each of the 256 corresponding pathways in the R5 network were reachable.
A final comparison of the network centrality measures of both networks showed that both the maximal betweenness (R5: $1.93 \cdot 10^{-5}$, X4: $9.36 \cdot 10^{-6}$) and the maximal closeness (R5: 0.17, X4: 0.16) of the R5 network were higher, which was also true for the corresponding minimal values.
Thus, the R5 network of all eight-point mutants was more central and concentrated on a specific region of the sequence space, and many paths traversed the central nodes of the network. In contrast, the central nodes of the X4 network were less important for the network structure. This finding again indicated a larger extension of the X4 network.
The detailed results are given in the summary in Table 4.2.
First analyses of the R5 and X4 population of the ten-point mutants showed a similar tendency. The inclusion of the ninth and tenth weakly conserved nucleotide position did not add a new first codon position to the mutated sequences, neither in the R5 nor the X4 data set, therefore we did not expect contradictory observations upon further analysis of the R5 and X4 network of ten-point mutants. On basis of the results for the previous network approaches and the different mutant networks, we decided to discontinue the analysis of the ten-point mutants and to omit the time-consuming analyses of shortest paths and centrality measures.
In summary, this part of the network analyses showed that the R5 networks concentrated in a condensed region of the sequence space. Central nodes of the network are an important component of many shortest evolutionary paths. The R5 networks contained no sequences of zero fitness, and pathways between nodes of maximal fitness and nodes of minimal fitness were longer than the average shortest path length. In comparison, the centrality values of the X4 networks showed that the networks were less concentrated. The central nodes were less important for the network structure, since they were less frequent integrated into the shortest network paths. Though the shortest min-max paths were also longer than the average shortest network pathways, the length difference was less pronounced. A comparison of the shortest min-max pathways in the R5 and X4 networks showed that the evolutionary distance of the sequences of minimal fitness and the sequences of maximal fitness was shorter in the X4 than in the R5 network. We assume that the X4 sequences yield an increased tendency to mutate towards low fit sequences that are no longer able to replicate efficiently.

Definition of neutral networks

In the two previous network approaches, we observed the structure of the R5 and X4 mutant networks using all available mutant sequences and enabling edges under different conditions (undirected edges at variable nt Hamming distances and different fitness deviation thresholds as well as evolutionary networks of directed edges from low to high fitness at a fixed nt Hamming distance of one). In the last part of the network analysis, we combined the fixed Hamming distance with a fitness constraint. Upon the network construction, we limited the sequences to a fraction of the R5 and the X4 sequences that surpassed a defined fitness threshold $eFit_{min}$. From the remaining sequences, we created an R5 respectively an X4 network based on two conditions. First, the nt Hamming distance between two sequences was one, and second, the representing vertices were connected via a directed edge, starting at the node with lower fitness and pointing towards the node with higher fitness. If both sequences had the same replicative fitness, two directed edges with an opposite orientation were created.

- create a vertex v_i if $eFit_{v_j} \geq eFit_{min}$
- add a directed edge pointing from vertex v_i to vertex v_j if
  - $H(v_i, v_j) == 1$ and
  - $eFit_{v_i} \leq eFit_{v_j}$

The fitness $eFit$ was determined based on Equation 4.12.
The idea behind the definition of a fitness restriction for the inclusion of the nodes respectively sequences into the networks was to reduce the networks to sequences of a high replicative fitness, since we presume those sequences to be biologically relevant for the course of the infection and the evolution of the sequences. Sequences with low or even zero replicative fitness are supposed to build a less relevant minority. The resulting networks should only contain sequences that are evolutionary neutral with respect to the replicative fitness of the underlying sequences.
Based on these definitions, we repeatedly analysed the structural differences between the R5 and X4 mutant networks. Depending on the selected fitness threshold $eFit_{min}$, we found a significant deviation in the size of the corresponding R5 and X4 networks, as presented in Table 4.3. The reason for the different network sizes are the deviations in the underlying fitness distributions discussed and presented in Figure 4.23.
The observed differences in size were largest in the networks of eight-point mutants using fitness constraints of 0.5 and 0.6. The definition of a fitness threshold led to networks that varied up to a factor of ten in size, making it less sensible to compare the resulting networks and to interpret the results. Clearly, the numbers of edges were smaller and also the paths lengths were remarkably shorter in the small networks.

Table 4.3.: **Graph size using minimal fitness constraint**
The table gives a overview over the number of nodes and the number of edges of the networks of four-, six-, eight-, and ten-point mutants, depending on the selected fitness threshold $eFit_{min}$.

		nodes		edges	
network	$eFit_{min}$	R5	X4	R5	X4
4-p. mut.	1.0	2	1	2	0
4-p. mut.	0.9	4	3	8	3
4-p. mut.	0.8	4	11	8	14
4-p. mut.	0.7	8	33	20	79
4-p. mut.	0.6	16	91	44	385
4-p. mut.	0.5	74	104	352	458
4-p. mut.	0.0	256	256	1,664	1,248
6-p. mut.	1.0	8	1	32	0
6-p. mut.	0.9	16	8	80	14
6-p. mut.	0.8	16	25	80	53
6-p. mut.	0.7	32	75	176	217
6-p. mut.	0.6	64	253	368	1,216
6-p. mut.	0.5	312	425	2,392	2,179
6-p. mut.	0.0	4,096	4,096	46,112	29,184
8-p. mut.	1.0	8	2	32	2
8-p. mut.	0.9	16	32	80	146
8-p. mut.	0.8	16	94	80	458
8-p. mut.	0.7	32	311	176	1,721
8-p. mut.	0.6	64	1,076	368	7,939
8-p. mut.	0.5	328	2,331	2,488	18,185
8-p. mut.	0.0	65,536	65,536	934,264	495,360

We decided to modify the fitness constraint and to use a fixed percentage of the most fit sequences of each population to balance the network sizes. As we discussed upon the introduction of the fitness function 4.12, the explicit fitness values are only valid for an intra-R5 or intra-X4 ranking of the replicative fitness. Due to a missing *in vitro* validation, we had no method to perform a direct comparison of the replicative fitness of the R5 and X4 fitness values.
Therefore, this relaxation of the network constraint is sensible. We tested fractions of sequences in a range x from 1% to 25%. In consequence, the neutral networks were defined in the following way:

- create a vertex v_i for the x percent of the most fit sequences i
- add a directed edge pointing from vertex v_i to vertex v_j if
  - $H(v_i, v_j) == 1$ and
  - $eFit_{v_i} \leq eFit_{v_j}$

Using this definition, we created networks of a fixed size, containing only the x percent most fit sequences of the R5 and X4 mutant population, up to a proportion of $x = 25\%$. The resulting networks are finite directed cyclic graphs. Cycles can only contain nodes representing sequences of identical replicative fitness (self-loops were restricted).
The neutral networks enabled us to analyse the differences between the underlying fitness landscapes of the R5 and the X4 sequences of the highest replicative fitness, presumably

representing the dominant sequences in an HIV infection.
We excluded the population of the four-point mutants from this analysis, since it contained only 256 sequences and resulted in very small networks. First analyses of these small networks showed that it was less informing to analyse and compare the network measures of these networks.
The restriction to use the 1 - 25% most fit sequences created R5 and X4 networks of equal size, but with differing least fitness values. These differences resulted from the deviations between the R5 and X4 fitness distribution we already described in Figure 4.23. Therefore, the least fit sequences that were included into the X4 network showed a higher fitness value than the least fit sequences that were included into the R5 network.
Using the 1% most fit sequences of the R5 population of all 6-point mutants for example comprised the sequence with fitness $eFit_{R5} = 0.69$ as least fit node. In the corresponding X4 network, the node with minimal fitness had a fitness of $eFit_{X4} = 0.75$. Correspondingly, the 1% most fit sequences of the 8-point mutants resulted in least fitness values of $eFit_{R5} = 0.40$ in the R5 network and of $eFit_{X4} = 0.66$ in the X4 network.
As stated above, we had no *in vitro* method to calibrate this pure numerical fitness values in terms of replicative capacity. Therefore, the numerical fitness value are valid for a ranking of the replicative capacity within the R5 or within the X4 population, but not across the populations. The numerical deviation of two fitness values, e.g. $eFit_{R5} = 0.40$ and $eFit_{X4} = 0.66$, could only be an artefact of our computational method and *in vitro* result in the same replicative fitness.
The minimal numerical fitness values of the six- and eight-point mutant sequences that were included into the R5 and X4 networks are compared in Figure 4.32, depending on the used proportion x of the most fit sequences. An overview over the fitness values and the additional network measures presented in the following lines are given in the Appendix in Table A.5 for the six-point mutants and in Table A.6 for the eight-point mutants.

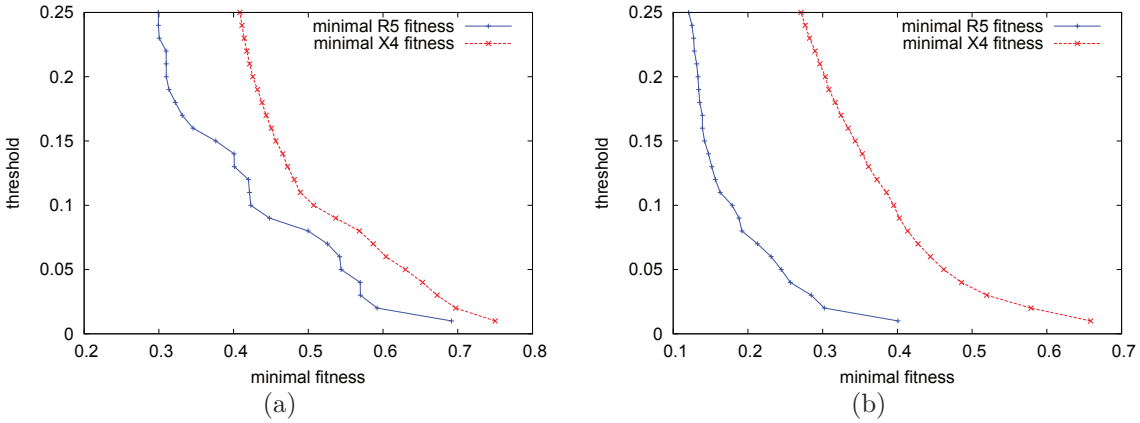


Figure 4.32.: **Minimal node fitness of R5 and X4 network**
The illustrations compare the minimal node fitness included into the R5 (blue) and X4 (red) network, depending on the selected threshold x. Figure (a) shows the minimal fitness for the six-point mutant networks, and Figure (b) for the eight-point mutant networks.

The analyses of the network measures of the neutral R5 and X4 networks of varying size confirmed our previous findings. A comparison of the maximal node betweenness of the R5 and the X4 networks of the most fit six- and eight-point mutants revealed that the maximal node betweenness of the R5 networks was always higher than the maximal node betweenness observed in the corresponding X4 networks (see Figure 4.33). Further comparisons of the average node betweenness as well as the minimal node betweenness witnessed this finding. Comparable to the evolutionary networks analysed before, the minimal node betweenness in the X4 networks was zero, but in general differed from zero for the R5 networks (data not shown).
The higher node betweenness of the nodes of the R5 network confirmed that the R5 networks are more centralised and that the central nodes are more often traversed by the shortest paths than the central vertices of the X4 networks. This result consolidates the idea that the R5 population is more condensed and centralised in sequence space, in contrast to the X4 sequences that are farther extended in sequence space.

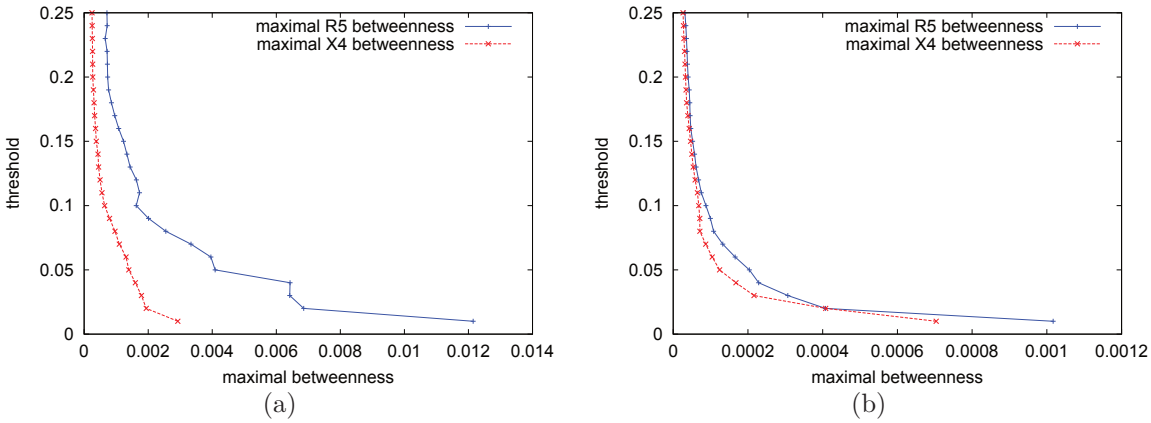


Figure 4.33.: **Maximal node betweenness of R5 and X4 network**
The illustrations compare the maximal node betweenness of the R5 (blue) and X4 (red) network, depending on the selected threshold x. Figure (a) shows the maximal betweenness of the six-point mutant networks, and Figure (b) of the eight-point mutant networks.

We found the same result upon the computation of the node closeness as an alternative centrality measure. The maximal node closeness for any threshold x in the range of 1 - 25% was higher for the R5 network than for the X4 network (compare Figure 4.34), and this observation was confirmed by additional analyses of the average and the minimal node closeness of the respective networks (data not shown). In agreement to the findings from the previous network analyses (see Table 4.2), the minimal node closeness in any of the X4 networks was zero, while the minimal closeness in the respective R5 networks in general differed from zero. The analyses of the node centrality measures coincided with the findings for the previous R5 and X4 network definitions.

A subsequent analysis of the maximal and the average node degree showed that the node degree in the R5 was higher (data not shown).

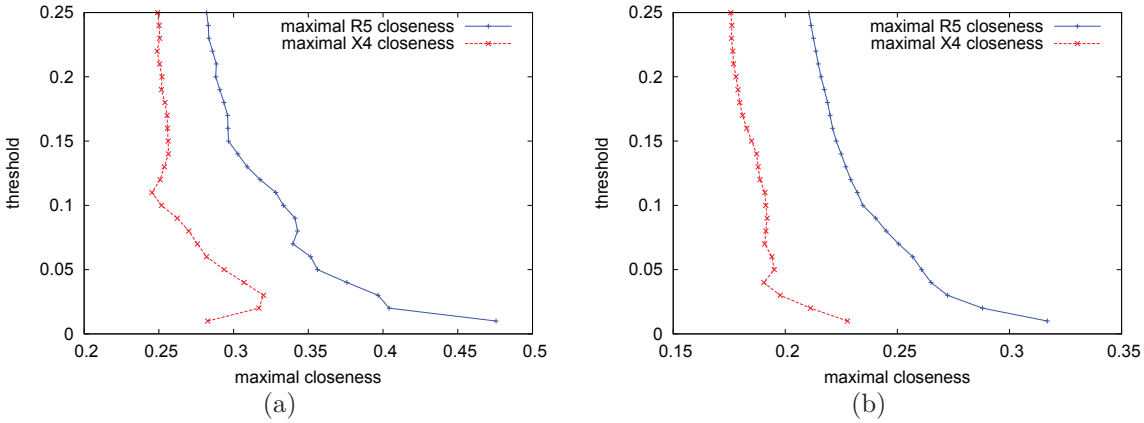


Figure 4.34.: **Maximal node closeness of R5 and X4 network**
The illustrations compare the maximal node closeness of the R5 (blue) and X4 (red) network, depending on the selected threshold x. Figure (a) shows the maximal closeness of the six-point mutant networks, and Figure (b) of the eight-point mutant networks.

Further computations showed that the shortest paths in the X4 networks were in general longer than in the corresponding R5 networks (compare Figure 4.35. This finding differed from the previous shortest path analysis 4.2, in which we found comparable average shortest path lengths in the R5 and the X4 networks. The deviation in the path lengths between the neutral networks and the evolutionary networks presumably resulted from the exclusion of the sequences of low fitness. In the previous evolutionary network definition, the low fit X4 sequences often led to an interruption of network paths in the more sparse and extended X4 network. Starting at a sequence of maximal fitness, a fraction of the least fit sequences was not reachable in each X4 network (compare number of not reachable min-max paths in Table 4.2). Using the present neutral network definition, the shortest path length of the X4 network increased, since less fit mutants were excluded and thus less frequently led to an interruption of the paths.

In summary, the analyses of the neutral networks of sequences of similar replicative fitness agreed with the findings for the previous network types. Thus, our results are independent of the details of the network definition. We conclude that the highly conserved R5-tropic viral sequences are evolutionary close and condensed in sequence space, while the more variable sequences of the X4-tropic viral populations extend farther to more distant regions of the V3 loop sequence space.

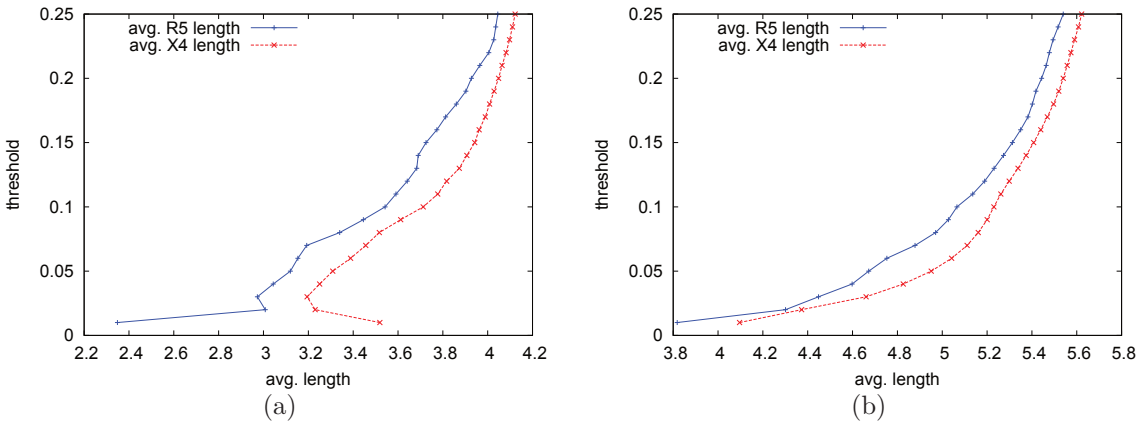


Figure 4.35.: **Shortest path length of R5 and X4 network**
The illustrations compare the average shortest path length of the R5 (blue) and X4 (red) network, depending on the selected threshold x. Figure (a) shows the average shortest path length of the six-point mutant networks, and Figure (b) of the eight-point mutant networks.

4.5. Discussion

In the second part of the work, we used the Los Alamos database [100] to collect a large V3 loop data set. We separated the sequences into two independent R5- and X4-tropic data sets, calculated multiple sequence alignments (MSA) of the R5- and X4-tropic sequences, and analysed the differences of the MSAs and the resulting consensus sequences. We further used the MSA to derive two fitness functions to describe the replicative fitness of R5- and X4-tropic V3 loop sequences. Finally, we studied evolutionary pathways defined by the fitness landscapes using methods from network theory to compare the underlying R5 and X4 fitness landscape.

We found deviations in the sequence conservation of the nucleotide (nt) and amino acid (aa) MSAs of the R5 and X4 data set. In general, the X4-tropic viral sequences are less conserved than the R5-tropic sequences. In consequence, random mutations of the more variable X4 consensus sequence altered the replicative fitness less than the introduction of random mutations into the highly conserved R5 consensus sequence.

During our analyses, we further revealed a putative fitness deficit of the X4-tropic strains. Upon analyses of the conservation of the MSAs with respect to the codon position, we observed a preference of the most weakly conserved X4 nt to occur in the first codon position. Among the ten most weakly conserved X4 consensus sequence positions, we found five times a first codon position. Though mutations in the ten least conserved X4 consensus sequence positions in general had a small impact on the X4 fitness, alterations in the five weakly conserved first codon positions frequently introduced stop codons into the mutant sequences of the X4 population. Based on the definition of our fitness functions, these stop mutations inhibit the replication of the respective X4 mutants.

In contrast to the five first codon positions among the ten most weakly conserved X4 sequence positions, we observed three first codon positions among the ten most weakly conserved R5 sequence positions. We further showed that no point mutation in the ten most weakly conserved R5 nt positions of our MSA (including the three first codon positions) could introduce a stop codon into any individual of the R5 mutant population. We formulate the hypothesis that a random mutation in general has a more negative impact on the replicative fitness of the R5 sequences, due the higher sequence conservation of the R5 consensus sequence, but a complete loss of the replicative fitness upon the introduction of a stop codon might occur in the less conserved X4 sequences at a higher probability.

Subsequent network analyses of the R5 and X4 mutant populations showed that the central nodes of the R5 network were part of many shortest paths and that the R5 sequences concentrated in a small region in sequence space, while the sparse and less centralised X4 network extended into a farther regions of the sequence space.

Furthermore, we found that the average shortest path length of the R5 and X4 networks was comparable, but the shortest evolutionary pathways between a sequence of maximal fitness and a sequence of minimal fitness were approximately one fourth shorter in the X4 network than in the R5 sequence network. These observations again indicated that the more variable sequences of the X4 population are evolutionary closer to sequences with low replicative fitness than the conserved sequences of the R5 population.

Based on our observations and on additional knowledge from literature, we hypothesise the following explanation for the co-receptor switch:

The fast replication of the HIV sequences and the high error rate of the reverse transcriptase result in the accumulation of replication errors. Since the R5 and X4 sequences are in close proximity in the sequence space, the accumulation of mutations frequently transforms R5 sequences into X4 and vice versa, and creates both R5- and X4-tropic HIV sequences.
In early stages of the disease, the immune system is strong. Due to a preference to detect X4-tropic viruses, X4 sequences are the major target of the immune system and the remaining R5-tropic sequences dominate the infection. In later stages of the disease, the decreasing immune pressure results in a diminishing immune selection, and the continuously high error rate of the reverse transcriptase creates a high sequence variability, which enables a dominance of the more variable X4-tropic sequences.
Thus, our hypothesis of the co-receptor switch merges the idea of the immune control hypothesis with our observation of a less conserved X4 population.

4.6. Outlook

Upon the comparison of local R5 and X4 mutant populations, our work revealed differences in the most weakly conserved sequence positions. In further studies, it is essential to extend our local studies onto larger populations of V3 loop sequences. To increase the observed sequence space, the next analyses should compare populations that carry mutations at the identical sequence positions, for example we could combine the weakly conserved R5 and X4 nucleotide positions and mutate both consensus sequences at the same sequence positions.
Furthermore, the replicative fitness of the V3 loop sequences should be confirmed experimentally in *in vitro* experiments. Above all, it would be very interesting to directly compare the replicative capacity of the R5 and the X4 consensus sequence. Though our data analyses indicated a higher replicative fitness of the highly conserved R5 consensus sequence, this observation is presumably an artefact of our computational method, that rewarded the higher sequence conservation with a higher replicative fitness.
In our work, we derived a fitness function that is composed of an amino acid based main fitness contribution and a second term that evaluates epistatic interactions between pairs of coupled amino acid mutations. In further studies, it would be challenging to include higher dimensions of amino acid interactions into the fitness function.
Our analyses were based on a cross section of data bank sequences that we assumed to represent a steady state of the V3 loop population. For further studies, it would be interesting to further increase the data set. One could include viral populations from deep sequencing approaches, and in addition gain fitness information based on the number of sequence duplicates found within the patient-specific viral populations.
Last but not least, our approach presumed static fitness landscapes. It would be interesting to transform the static fitness landscapes into dynamic fitness landscapes, and to model for example changes in the immune pressure, the cell availability, and in the administered therapy. The increasing availability of high-throughput sequencing techniques could enable an approach to formulate a dynamic fitness landscape based on the longitudinal deep sequencing information of a single patient, for example at different time points during the infection, supplemented by the additional clinical information as presented in the first part of this work.

5. Simulation of evolution of HIV-1 V3 loop

5.1. Introduction

In the last part of this work, we developed a model to simulate the evolution of HIV V3 loop populations in an artificial *in silico* setting.
The observation of a patients complete viral population is time consuming, costly and depends on the compliance of the patient. Our model enables us to create a multitude of different viral populations, starting from varying founder sequences or founder populations, and to observe their evolution in detail. Using this simulation tool, we are not limited by the restrictions of a clinical study, neither by the compliance of the patient and the physician, by the sensitivity of the sequencing method, by the observation time, or by the costs.
The only restriction we face in an *in silico* approach is the availability of sufficient computation time and hardware to perform the simulations, to store the data, and to analyse the simulated populations, but these are restrictions we also face upon the analysis of real sequencing data, for example from a high-throughput sequencing approach.

5.2. Methods

In the following sections, we describe the established evolutionary model of Moran [108] on which our simulation of the V3 loop evolution is based. We first introduce the basic concept of the model, then we describe some model extensions and introduce the adaptations we used in our simulation.

5.2.1. Moran model

The basic idea of the Moran model was introduced by Moran [108] in 1958. In general, the model describes the evolution of a population of N individuals over time.

Basic Moran model

In the standard Moran model, two representations, a and A, build a population of N individuals. In each time step, one random individual x is replaced by another random individual which is selected from the whole population (including x). Thus, the allele frequencies of A and a can change by 1 in each time step.
Starting from k individuals A and $N-k$ individuals a, with the birth process b and death process d, the frequency of A changes following these equations:

$$\begin{aligned} k \to k+1 : b_k &= (N-k) \cdot \frac{k}{N} \\ k \to k-1 : d_k &= k\frac{N-k}{N} \end{aligned} \tag{5.1}$$

Note that the birth and death rate, b_i and d_i, are identical.

Adaptation of the Moran model for simulation

Our simulations are based on the Moran model. In contrast to the classical model, in which one individual is replaced in every time step, our simulation tool replaces ten individuals in every time step. This modification was introduced to decrease the computation time of the simulation runs.
In test simulations, the modification did not influence the evolutionary course of the simulated sequences, but changes of the default population size of the model could alter the impact of this adaptation.

5.2.2. Simulation

We used the Perl programming language to develop the simulation tool. Each run of the simulation is divided into three phases:

- model initialisation
- simulation turns
- sampling and computation of mean population fitness

A graphical description of the simplified model workflow is given in Figure 5.1.
During initialisation (1), the user defines the evolutionary parameters of the model. These

are the duration of the simulation, the population size, the rate of sequence sampling, the mutation rate, the strength of the epistasis, (the parameter β in Equation 4.12), the rate of the R5 and the X4 fitness contribution (the parameter α in Equation 4.12), their variation during the simulation, and the method to select the replicating sequences.
After initialisation, the simulation starts to run for the predefined number of turns (2). Each turn starts with a replication phase, during which new offspring are created and random individuals die and are removed from the population (except the newly created offspring). In this phase, the mean fitness of the population is calculated. After an user-specified number of turns, a random sequence sample is selected and saved to a sequence file.
At the end of the simulation (3), a final random sample is saved. In addition, the complete final population is saved for further analyses.

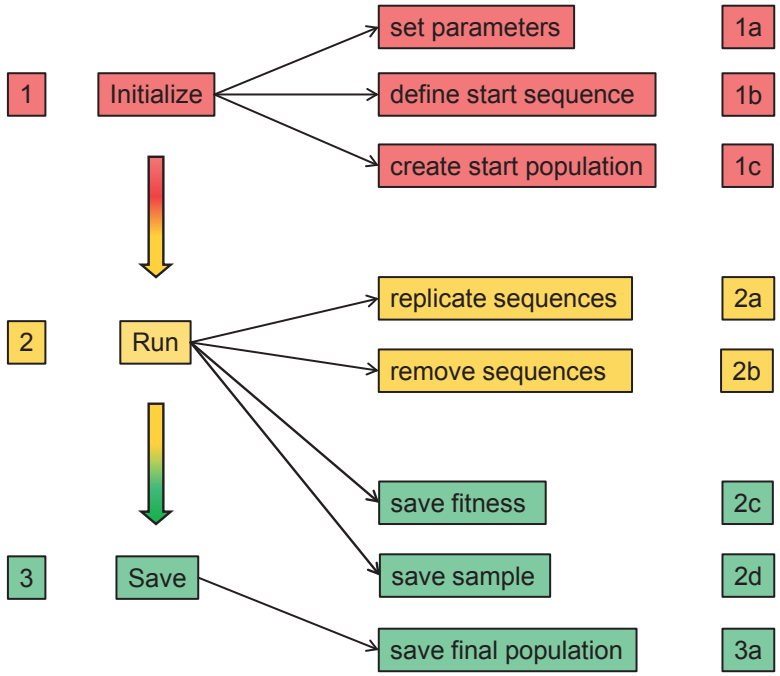


Figure 5.1.: **Graphical description of the workflow of the simulation**
The illustration describes the sequence of steps of a simulation run.

Due to the importance of the simulation parameters for the course of the evolution, the simulation is described in detail in the following sections.

Initialisation of the simulation

In the initialisation phase, the user defines the simulation parameters. The founder population is defined by the size of the population and by the selected sequence of the 105

nucleotides (nt) that build the V3 loop. The user can decide whether to use a constant population size or to allow the population to vary in size. In general, we use a population of a constant size of 3,000 sequences for our simulations. This is based on the knowledge that the virus population in the blood of a successful treated patient is almost constant over time, with an estimate of 3,000 cells infected. A population of variable size will increase or decrease, depending on the replicative fitness of the sequences of the population. The fitter the population, the more sequences are by chance selected for replication.
In addition to the size of the population, the user also decides whether the start population is homogeneous or heterogeneous. Choosing a uniform start population enables the user to decide about the exact nt start sequence. The simulation will then start with the defined number of identical, predefined nt sequences.
In the default setting, our simulation generates a homogeneous initial population of identical V3 loop sequences, based on the knowledge that an *in vivo* HIV infection typically is seeded by only one viral founder strain [52]. The homogeneous population consists of identical copies of a random ten-point mutant of a modified V3 loop consensus sequence. We previously analysed that the original R5 and X4 consensus sequences only differed in the nt positions 64 (R5: `G`, X4: `A`) and 73 (R5: `G`, X4: `A`). To minimise the risk to introduce a co-receptor bias by the founder population of the simulation, the R5 and X4 consensus sequence was modified into a neutral sequence with respect to the co-receptor usage. Thus, both position 64 and 73 the nt was altered manually into the alternative nt `T`, mutating the aa from `A` (R5) and `T` (X4) into `S`, respectively from `D` (R5) and `K` (X4) into `Y`.
The resulting *co-receptor neutral* sequence is then mutated in ten random nt positions to create a start sequence for the simulations. For additional variability, a back-switch of the randomly selected nt position into the consensus nt symbol is not restricted. Therefore, the resulting start sequence has a nt Hamming distance in the range of two to twelve with respect to the R5 and X4 nt consensus sequence.
It is also possible to start from a heterogeneous population of nt sequences of length 105, with the nt sequences randomly created from the letters of the DNA alphabet (G, C, A, T). The simulation ensures a founder population of nt sequences without stop codons. Random sequences with stop codons are created again until they are replaced by a sequence without any stop codon.
The duration of the simulation is measured in replication cycles (termed *turns*). The default duration of the simulation was set to 30,000 turns and can be modified by the user upon the simulation start. Using the default parameters, a simulation of 30,000 turns ensured the convergence of the population towards the consensus sequence.
Next, the user can define the mutation rate. We determined a default mutation rate of 0.016 to result in an average of one amino acid mutation per replication. The average number of mutations is based on an analysis of 1,000 simulation runs. In parallel, we determined the mutation rate of 0.016 analytically A.3. This mutation rate is set as default mutation rate for the simulation. Note that higher mutation rates increase and lower rates decrease the speed of evolution, leaving the course of the evolution in general unaffected.
The user can further decide between roulette and tournament selection to select a sequence for replication (see next Section), and between an additive or a multiplicative fitness function to calculate the replicative fitness of the sequences. The roulette selection method based on a multiplicative fitness function is defined as the default setting.

Finally, the user can define the sampling rate and the sampling time. The sampling rate determines the number of sequences that are saved and the sampling time determines after which number of turns a sample is saved to the result file.
An overview over the simulation parameters is given in the following enumeration. The default values are given in parenthesis.

- number of turns (30,000)
- initial population size (3,000)
- homogeneous or heterogeneous starting population (homogeneous)
- start sequence (ten-point mutant of consensus sequence)
- constant or variable population size (constant)
- mutation rate ($\mu = 0.016$)
- roulette or tournament selection (roulette)
- fitness function: additive/multiplicative (multiplicative (see Equation 4.12)
- epistatic strength ($\beta = 0.8$)
- R5 fitness contribution ($\alpha = 0.5$)
- X4 fitness contribution $1 - \alpha$
- sampling time: every t turns ($t = 100$)
- fraction of sampling: r ($r = 0.05$)

In addition, the user can decide to change the influence of the R5 or the X4 fitness function during the simulation. Therefore, an optional parameter x can be defined which determines the number of turns after which the parameter α is altered. In general, $x = \frac{numberofturns}{10}$. Using this setting, α is increased every 10% of the turns by 0.1. Also this value can be modified upon simulation initialisation.
Apart from this expansion, the user can decide to use only the main or only the epistatic fitness term by an alteration of the parameter β, as described in Equation 4.12.
Optionally, the user can initialise the random number generator of the simulation with a random seed to get reproducible simulation results.

Simulation turns

After the initialisation, the simulation of the sequence evolution is started. During each turn of the simulation, ten sequences of the population are selected for replication and ten sequences are removed from the population.

Replication During the replication step, parental sequences are selected to create offspring based on their fitness. The underlying selection algorithm is either roulette or tournament selection. In general, all sequences have a chance to get selected by both methods, but sequences with a larger replicative fitness are selected for replication at a higher frequency. Upon roulette selection, a sequence is selected proportional to its fitness. Sequences with a higher replicative fitness get selected with a higher probability, and sequences with a fitness of zero (i.e. sequences that carry at least one stop codon) are selected with a probability of zero , thus they are excluded from the selection process.
If tournament selection is used to select a parental sequence, the selection is a two-step process. First, two sequences are randomly picked from the population (i.e. not depending

on their fitness). In the second step, the two individuals compete and only the sequence with the higher fitness is replicated.
In contrast to the classical Moran model, an average of ten parental sequences is selected to replicate per turn.

Mutation The simulation performs the sequence replication with a chance for a mutation (i.e. error-prone replication). Applying the defined mutation rate μ, each of the 105 nt positions is checked for a possible mutation upon replication. The resulting nt sequence of the offspring is translated into an aa sequence and the fitness of the newly created offspring is calculated based on the selected fitness function.

Death After replication, an number of sequences equal to the number of newly created offspring is randomly selected and removed from the population. Since the classical Moran model assumes a parallel birth and death process, those offspring that were produced in the same turn are excluded from the death selection process.

5.2.3. Save simulation results

During the simulation run, the results are saved to a result file at equidistant time points. The distance of the time points can be defined by the user. The saved results comprise the nt and aa sequences of the individuals of the population, their individual replicative fitness as well as the mean replicative fitness of the population.
For large populations, the user can decide to save only a fraction of all sequences, i.e. a limited sequence sample, instead of the complete population. The sampling rate can be selected upon the initialisation of the simulation.
Independent of the defined sampling rate, the complete final population is saved at the end of the simulation.

5.3. Results

In the following section, we analyse a number of simulation properties and evaluate the capability to reproduce the course of the V3 loop evolution. We start with an analysis of the influence of the simulation parameters. Subsequently, we compare the effect of the different definitions of the underlying fitness functions and finally, we present selected simulation results using the default simulation parameters given in Section 5.2.2.

5.3.1. Parameter analysis

During the design of the simulation tool, we tested a variety of different simulation parameters and a range of parameter values. The main purpose was to get a first impression of the features of our model simulation.
Some of the parameter settings had no biologically meaningful equivalent and were excluded from further considerations (e.g. simulations of the viral evolution merely based on epistatic interactions, constant and very small viral populations, a random selection of replicating parental sequences, simulations with heterogeneous random starting populations, or with unrealistic viral start sequences as for example mere poly-A sequences).
The parameter analyses that were relevant to mimic a biologically sensible V3 loop sequence evolution are described in the following lines.

Population size and mutation rate

We performed a parameter scan for both roulette and tournament selection to determine biologically meaningful parameters for the simulations. Therefore, homogeneous test populations were simulated for 50,000 turns and tested in ten simulations. We analysed the influence of the mutation rate μ in the range of 0.001 to 0.032 and of the population size in the range of 50 to 3,000 individuals.
The parameters of a simulation were defined to be appropriate, if the simulated sequences evolved towards the most fit consensus sequence, represented by a fitness value of 1.0 (based on Equation 4.12), at some time point during the simulation (termed *timed to convergence*). This definition relies on the observation that the consensus sequence in general spreads fast throughout the population shortly after its first occurrence.
Using the roulette selection method, the time to convergence was closely related with the population size (compare Figure 5.2). The larger the population, the earlier we observed a sequence that evolved towards the consensus sequence. While the largest mutation rate of 0.032 only converged for simulated populations of more than 500 sequences, the smallest mutation rate of 0.001 evolved towards the consensus sequence only in populations larger than 1,000, (using a simulation duration of 50,000 turns). Additional simulations of a duration of up to 500,000 turns showed only slightly improved result.
A detailed inspection of the populations during the simulations revealed that the sequences do not persist long enough in small populations to enable their replication, since the probability that they become removed from the population is higher. To avoid sequence with increased fitness to be swept out before they create offspring, we either had to increase the population size or the life span of the sequences in our simulations.
In a further analysis of the simulated populations, we found that small mutation rates (e.g. 0.001) increased the mean population fitness, while large mutation rates (e.g. 0.032)

decreased the mean fitness (data not shown). The simulated data indicated that large mutation rates tend to overshoot the mark. Once the simulated sequences reached the consensus sequence, simulation runs with large mutation rates continued to alter multiple nucleotides within one round of replication and thus failed to stabilise the population at the fitness optimum. We observed that large populations could cope with this problem. Though the sequences spread wide in sequences space, a proportion of the simulated sequences in a large population conserved the maximal fit consensus sequence. In the case of small mutation rates, the majority of the sequences exactly hit the most fit consensus sequence, since only rare mutations were introduced into the offspring upon replication.

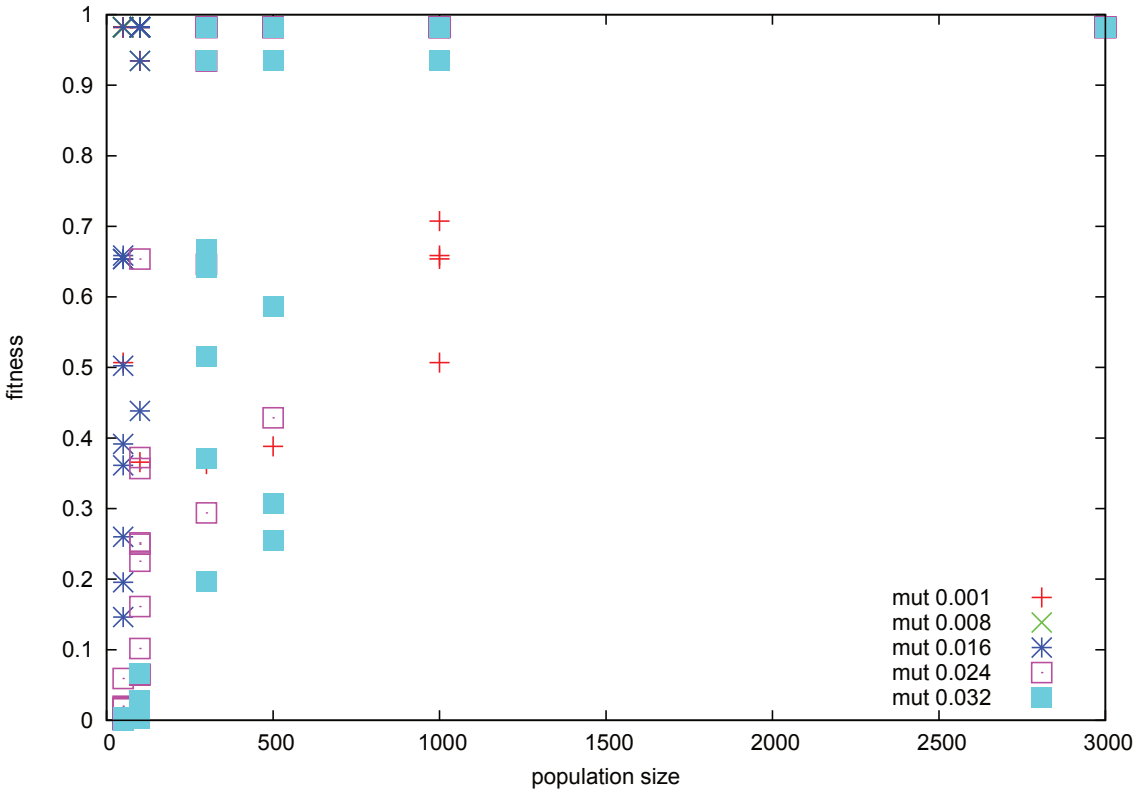


Figure 5.2.: **Parameter scan for simulation using roulette selection**
The figure illustrates the association of the population size, the mutation rate μ, and the maximal fitness of the simulated population upon roulette selection. A fitness of 1.0 is indicative of an evolution towards the consensus sequence.
A mutation rate of 0.032 converged for population sizes larger than 500, and a mutation rate of 0.001 converged for population sizes larger than 1,000.

We performed the same parameter test for the tournament selection method (compare Figure 5.3). The simulations showed that the method was very sensitive to the selected mutation rate. Mutation rates $\mu > 0.008$ (i.e. on average one aa mutation in every second turn) restricted the population to evolve towards the consensus sequence during simulations of 50,000 turns. A five- to tenfold duration of the simulations (i.e. up to 500,000 turns) increased the success, but still some simulations failed to evolve to the consensus sequence.

We found that this was mainly a consequence of the first step of the tournament selection method. As log as there are many sequences of low fitness and only a few sequence of high fitness in the population, the random selection picks sequences of high fitness at a low probability. Therefore, the fitness-dependent second step can only decide between two less fit sequences. Only if there are enough high fit sequences in the population, the highly fit sequences are selected steadily for replication. In consequence, the time to convergence is longer upon the usage of tournament selection.

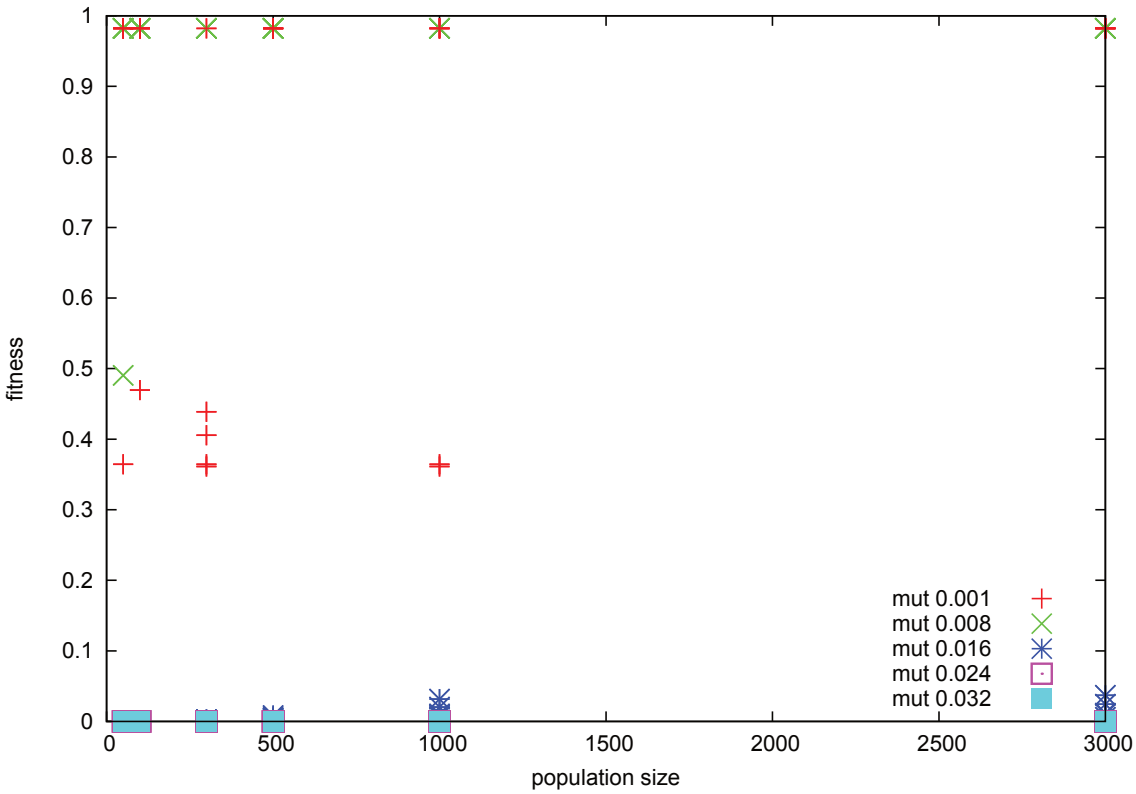


Figure 5.3.: **Parameter scan for simulation using tournament selection**
The figure illustrates the association of the population size, the mutation rate μ, and the maximal fitness of the simulated population upon tournament selection. A fitness of 1.0 is indicative of an evolution towards the consensus sequence.
The use of the tournament selection method in our simulations was very sensitive to the mutation rate. Mutation rates larger than 0.008 often failed to evolve towards the consensus sequence during simulations of 50,000 turns.

Due the increased computation time, the simulations based on tournament selection were less useful to observe the evolution of many populations in parallel. Thus, we decided to use roulette selection. We used populations $\geq 1,000$ sequences, since populations of that size were quite robust to any of the tested mutation rates, as analysed in Figure 5.2. Higher mutation rates resulted in shorter convergence times, but introduced more sequence diversity, while simulations with lower mutation rates required a longer simulation duration, but resulted in populations of less sequence diversity. Thus, we selected a fast default mutation rate of 0.016 for roulette selection (on average one aa mutation per replication).

Sampling

During a simulation run, a user-defined fraction of sequences is written to a result file. Given enough memory space, the user can decide to save all intermediate sequences of each generation. The subsequent analyses of the full sequence data of the default simulation with 3,000 sequences and 30,000 turns would then comprise 90,000,000 sequences. In test simulations, this amount of data required more computational time for the determination of the pairwise nt Hamming distances and the calculation of the diversity and the divergence than the simulations themselves.
For practical reasons, we found a sampling process in predefined timely intervals (e.g. every 100 turns) more suitable. Furthermore, limited sequence samples with intervals of unknown populations in between mimic the conditions of our longitudinal patient study, during which sparse samples were taken in variable time intervals.
To analyse the effect of sampling and to find an optimal sampling rate, we performed a parameter scan for the sampling rate. Figures 5.4 and 5.5 illustrate the influence of the sampling rate on the population diversity and divergence over time, as defined in Section 3.2.1. For a population of 3,000 sequences, we tested sampling rates of 100%, 10%, 5%, and 2% , i.e. all 3,000 sequences and 300, 150, and 60 randomly drawn sequences. The sampling rate of 100% represents the real sequence diversity and divergence of the simulated population.

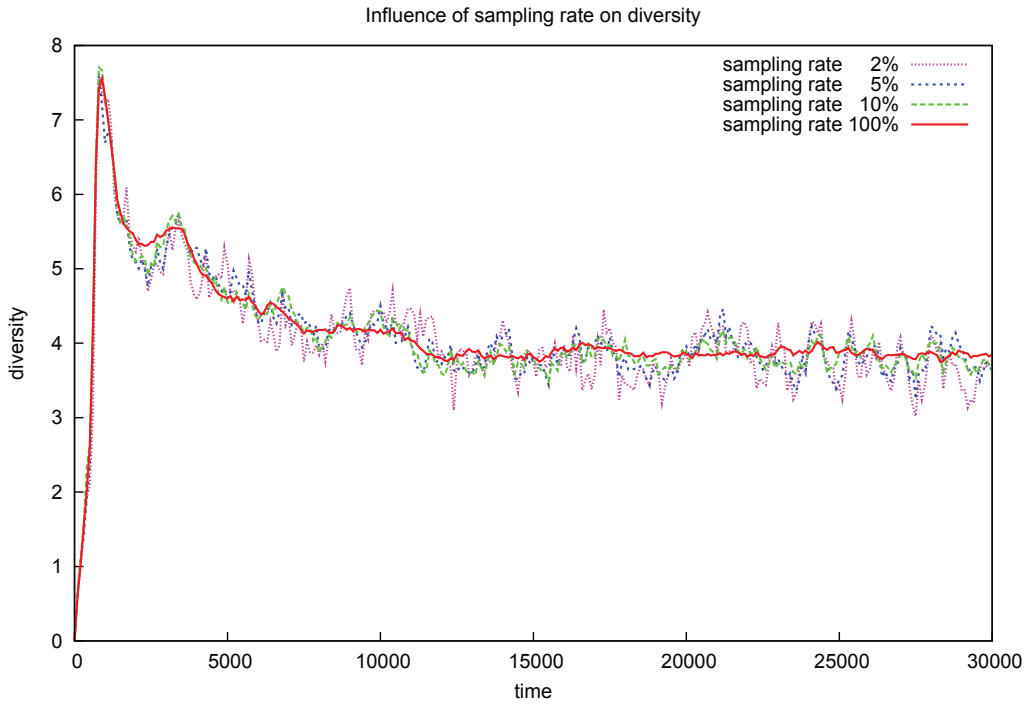


Figure 5.4.: **Parameter scan for sampling rate**
The figure illustrates the influence of the sampling rate on the evolution of the population diversity over time. The red curve (100% of all sequences saved) equals the real sequence diversity of the simulated population and served as a standard to test sampling rates of 10%, 5%, and 2% for a population of 3,000 sequences.

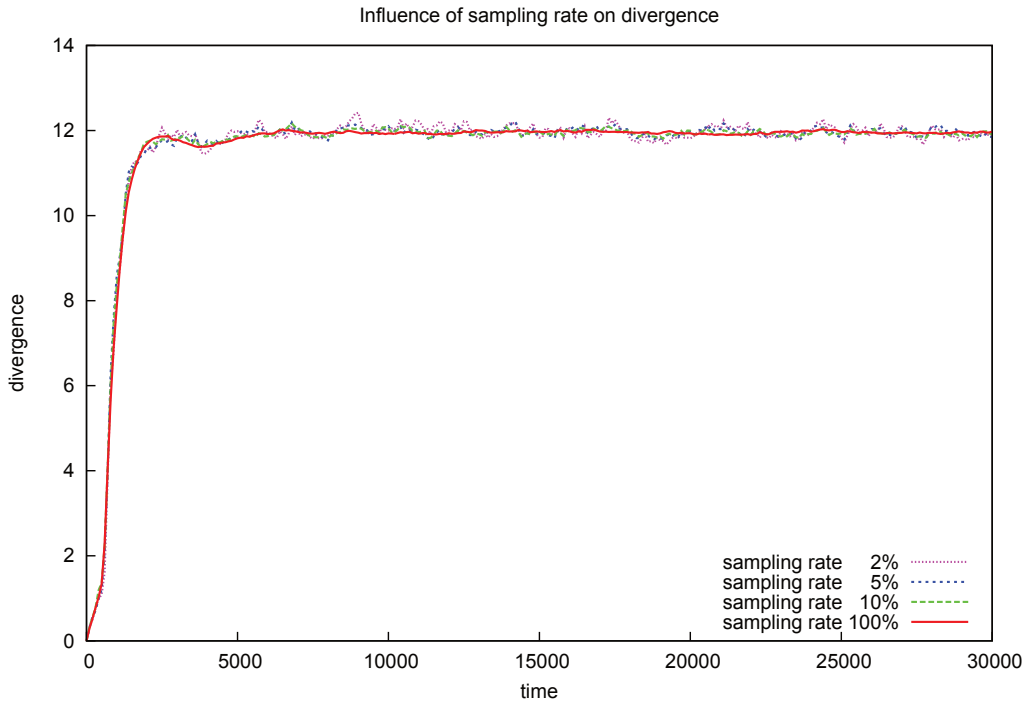


Figure 5.5.: **Parameter scan for sampling rate**
The figure illustrates the influence of the sampling rate on the evolution of the population divergence over time. The red curve (100% of all sequences saved) equals the real sequence divergence of the simulated population and served as a standard to test sampling rates of 10%, 5%, and 2% for a population of 3,000 sequences.

As expected, the comparison showed an increasing deviation of the real measures from the sample estimates with decreasing sampling rate. The required memory space for the analyses scaled with the same rate, while the computation time scaled quadratically with the sampling rate, due to the calculation of the Hamming distance of each sequence pair upon the computation of the diversity. Based on these observations and with respect to the consumption of memory space and computation time, we used a default sampling rate of five percent, i.e. a sampling parameter of $\frac{1}{20}$ for the simulations.

Additive versus multiplicative fitness

In our model, two fitness contributions influence the fitness value of a sequence. The fitness is composed of the main fitness term, based on the aa sequence, and of the epistatic fitness term, describing the effect of pairs of interacting mutations. Resolving the mathematical description of the fitness terms in more detail, both the main and the epistatic fitness contribution are composed of multiple individual fitness contributions. The exact definitions were described in Sections 4.4.3 and 4.4.5.
We considered different possibilities to join the single fitness contributions: the summation of all single fitness contributions, the multiplication of all single fitness contributions (as described in Equation 4.10 and analysed in Section 4.4.3), and a combination of the summation and the multiplication (i.e. the summation of the single position specific contributions, and the multiplication of the resulting main and epistatic term).

We tested the different methods and analysed the effect on the resulting replicative fitness as well as on the course and the outcome of the simulations. The test simulations confirmed our theoretical considerations presented in Section 4.4.3 (results not shown). Simulations based on the multiplicative fitness function sifted the less fit sequences out fast and soundly, due to an increased chance to select sequences of high fitness for replication. In the case of the additive fitness function, the simulations run approximately the tenfold number of turns to evolve towards the consensus sequence and the resulting final populations were more heterogeneous and showed a decreased mean population fitness, compared to simulations that were based on the multiplicative fitness function. The rational behind this observation is that the multiplicative fitness function has a distinct and steep fitness peak, representing the optimal consensus sequence, while the additive fitness function creates a fitness landscape which is more flat and has a less pronounced consensus sequence peak. Piganeau *et al.* [117] described that additive selective effects are not effective when the number of selected sites is larger than the effective population size, an assumption they state to be realistic for current molecular data.
Based on these results and the previous considerations (compare Section 4.4.3), we decided to use a the multiplicative formulation of the fitness function for the default simulation model, as presented in Equation 4.12. The resulting values were normalised to the interval $[0.0, 1.0]$ as described upon the introduction of the fitness function.

5.3.2. Results of the default model

We used the default parameter settings described in Section 5.2.2 to perform multiple simulations of the evolution of the V3 loop, starting from homogeneous populations of random ten-point mutants of the R5 and X4 nt consensus sequences presented in Section 4.3.2.

Evolution of the diversity and the divergence

Based on the default simulations, we analysed the course of the diversity and the divergence (defined by Equation 3.2.1). Figure 5.6 illustrates the course of the diversity, using the default simulation parameters (see Section 5.2.2) and the fitness function presented in Equation 4.12. The simulations started from homogeneous populations with zero diversity, and the diversity increased during the first 2,000 turns, representing approximately 20,000 replications. Around this time, one of the newly created offspring mutated into the most fit consensus sequence. In the subsequent simulation turns, the consensus sequence became dominant and spread throughout the population. As a consequence, the diversity of the population decreased and finally stabilised at an average diversity level depending on the mutation rate.
A comparison of the simulation results with the results of Shankarappa *et al.* [138] showed that the course of diversity created by our simulations nicely coincided with the three-staged pattern of viral evolution published in 1999.

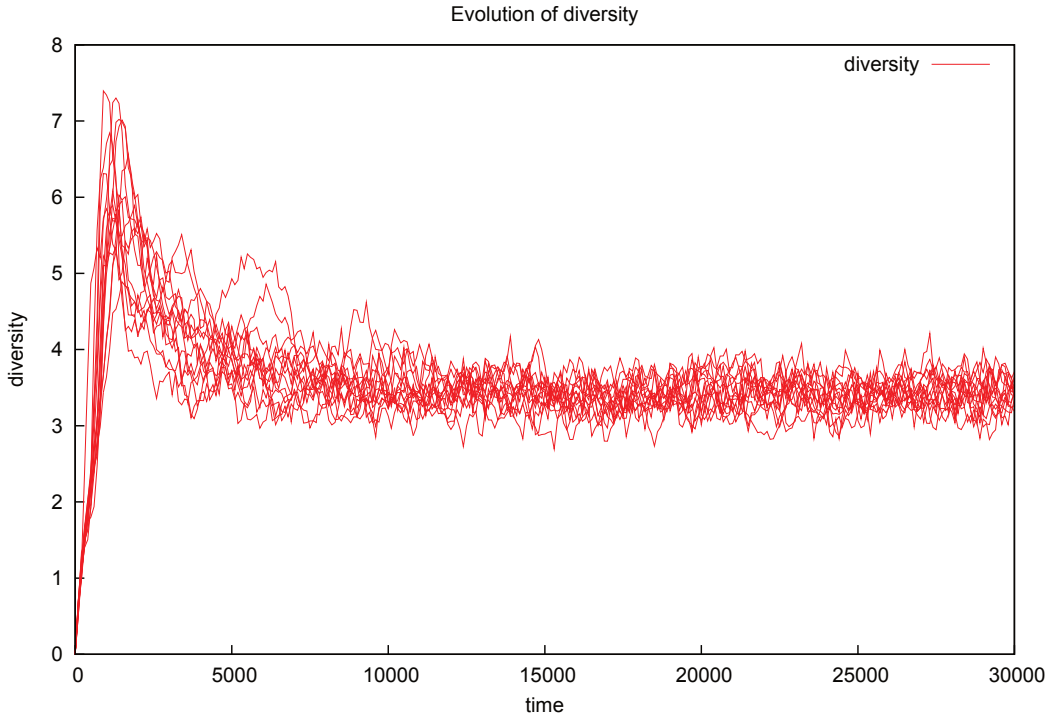


Figure 5.6.: **Evolution of diversity in the default model**
The figure illustrates the course of the diversity of 15 example runs, using the default simulation parameters.
The diversity showed a three-staged pattern: an increase during the first 2,000 turns, a decreasing diversity while the consensus sequence became dominant, and a final stabilisation at an average diversity level.

The corresponding analysis of the course of the divergence is shown in Figure 5.7. After an initial increase of the divergence during the first 2,000 turns, the divergence of the population stabilised at a specific nt Hamming distance, representing the genetic distance between the initial sequence and the consensus sequence. The extend of the final divergence of the population depends on the exact number of mutations of the founder strain.
Using the default weighting parameter $\alpha = 0.5$, we observed a closer genetic distance of the simulated sequences to the R5 consensus sequence in the majority of all simulations. A temporarily increase or decrease of the divergence during the course of evolution was observed in single simulation runs. Detailed analyses showed that this was a result of a transient dominance of the X4 consensus sequence. The closer genetic distance of the simulated sequences to the R5 consensus sequence could so far not be completely resolved. The reduced epistatic interactions in intermediate sequences seem to be one cause. We further presume that the R5 dominance is at least partially a consequence of the finite population size and the limited simulation time.
A comparison of the simulated course of divergence with the evolutionary course of an HIV-1 infection described by Shankarappa *et al.* [138] showed similar patterns. After an initial increase of the divergence during the acute phase of the infection, the genetic distance stabilised at some time point during the disease and stayed at that level until the end of the simulations.

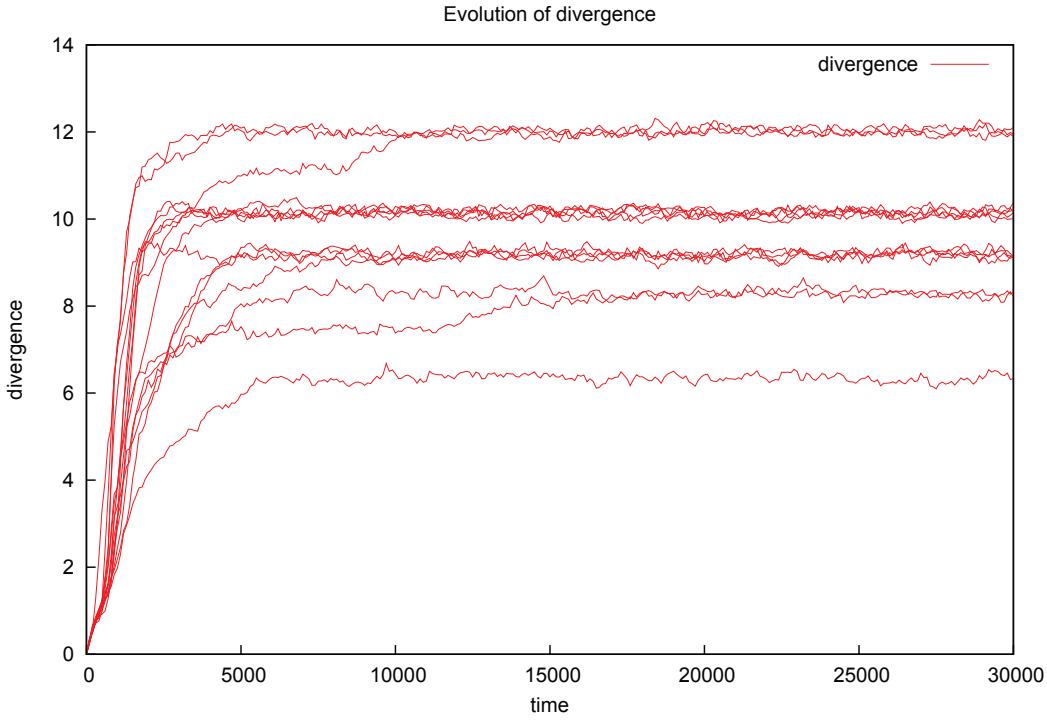


Figure 5.7.: **Evolution of divergence in the default model**
The figure illustrates the course of divergence of 15 example runs, using the default simulation parameters.
The divergence increased during the first 2,000 turns, until one of the offspring evolved towards the most fit consensus sequence. While the consensus sequence became dominant, the divergence of the population stabilised.

Thus, regarding the viral diversity and divergence, our simulation tool is capable to mimic the evolutionary course of an HIV-1 infection as described in early biological studies of Shankarappa *et al.* [138]

Hamming distances of the final population

Subsequent to the analyses of the viral diversity and divergence, we analysed the genetic distance between the simulated populations and the consensus sequences. Therefore, we calculated the aa Hamming distance of each sequence of the final population to both the R5 and the X4 aa consensus sequence. All default simulations resulted in comparable final populations.
Figure 5.8 shows the distribution of the resulting aa Hamming distances. As previously described, the R5-tropic sequences dominated the population at the end of the simulation runs, despite an equal weight for both the R5 and X4 fitness contribution ($\alpha = 0.5$). The definition of our fitness function (defined in Equation 4.12) favoured the R5 sequence due to less epistatic interactions in intermediate sequences. Using $\alpha = 0.5$, a population in favour of the X4 consensus sequence was only observed transiently in intermediate time steps during the simulation.
In the default simulations, about $\frac{1}{6}$ of the 3,000 sequences of the final population of each simulation exactly matched the R5 consensus sequence, i.e. the aa Hamming distance to

the R5 consensus sequence was zero for about 500 sequences per run. The majority of the sequences ($\sim 1,000$ sequences) accumulated at a Hamming distance of one. We found this to be a result of the continuous replication and mutation process. Once the consensus sequence is present in the population, it is most likely selected for replication due to the maximal fitness value. Upon the following replication, the default mutation rate of 0.016 introduces on average one aa mutation into the newly created offspring. This phenomenon is further discussed at the end of the section.

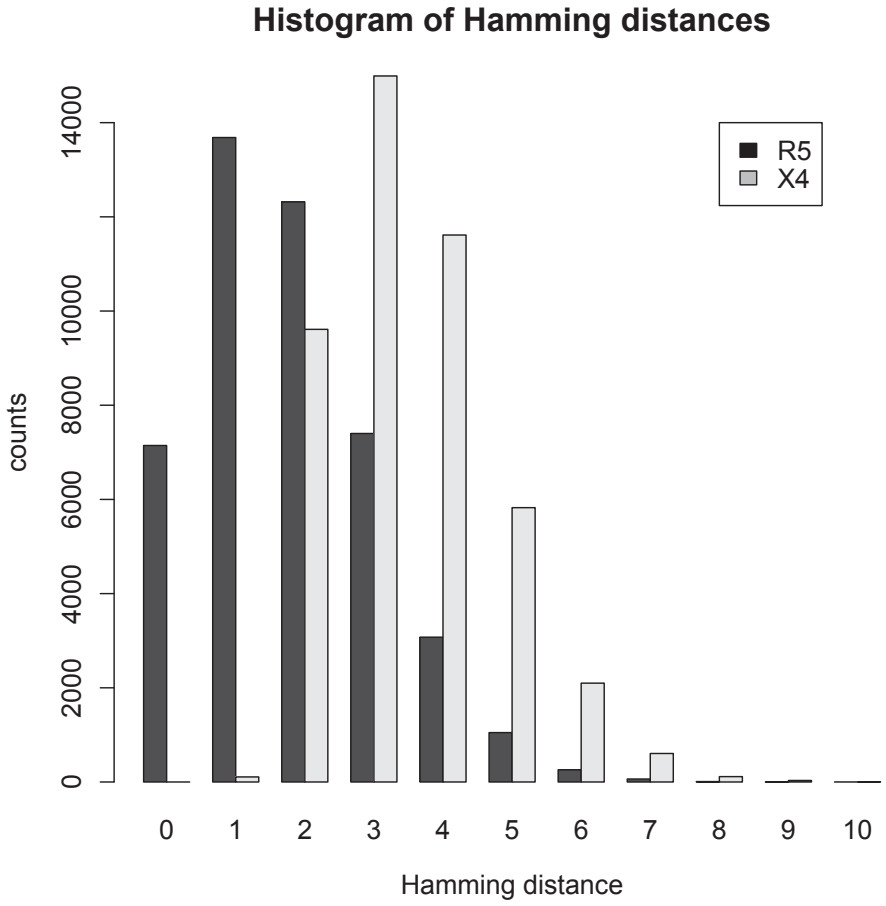
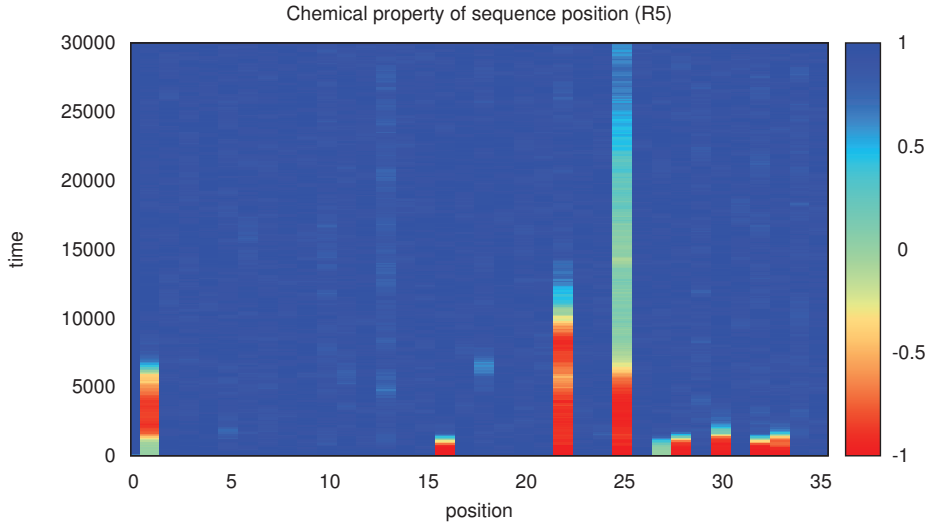


Figure 5.8.: **Hamming distances of the final population of the default model**
The figure illustrates the distribution of the aa Hamming distances of the sequences of the final population to the R5 and the X4 consensus sequence.
Both the Kolmogorow-Smirnov [92, 143] and the χ^2 [114] test showed that the distributions of aa the Hamming distances were significantly different ($p - value =< 0.0001$ for both tests).

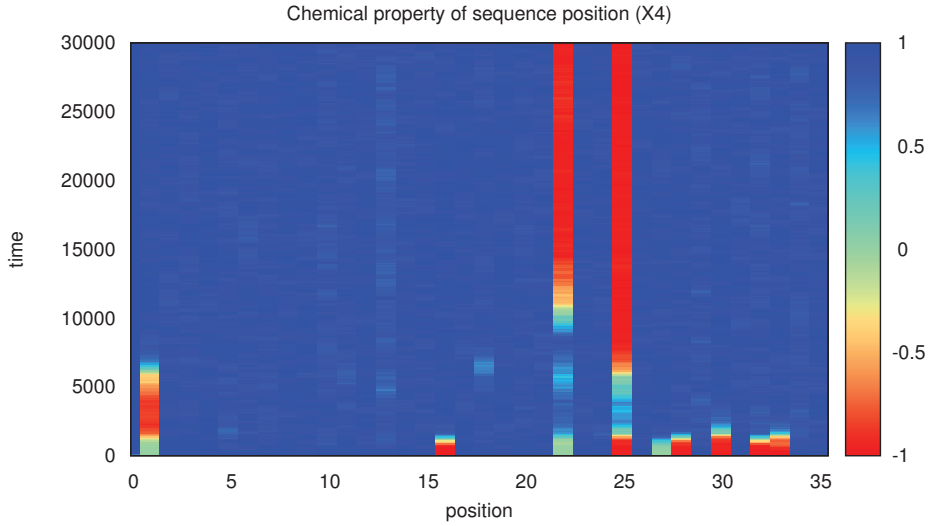
In additional simulation runs, we found that an alteration of the weighting parameter α in favour of the X4 fitness function (i.e. $\alpha < 0.5$) directed the viral evolution either towards mixed R5- and X4-tropic or mere X4-tropic final populations, depending on the exact parameter value (data not shown).

Evolution of position specific amino acids over time

Following the analyses of the course of the diversity and divergence and of the Hamming distances of the final populations, we next asked whether the course of evolution is chemically sensible. Therefore, we classified the amino acids into four chemical groups: basic (K, R, H), acidic (E, D), polar (Y, T, Q, G, S, C, N), and non-polar (A, V, M, L, I, P, W, F) (compare Table A.7).
During the simulations, the sampled sequences were translated into a chemical score with respect to the target consensus sequence, regarding the following scheme: amino acids matching the consensus sequence amino acid were scored as 1.0, differing amino acids belonging to the same chemical group as the consensus amino acid were scored as 0.0 and differing amino acids of a differing chemical group were scored as -1.0. We scored all sequences of each sample and calculated the position-specific sample average.
Figure 5.9 depicts the position-specific average of the samples compared to the respective R5 (a) or X4 (b) consensus sequence. The illustrations show that evolution of the mutated positions follows a meaningful chemical course. In general, mutated positions first met the correct chemical group and finally evolved towards the correct amino acid. This effect is especially pronounced in figure (a) for the R5 sequence position one, 22, and 25.
A comparison of the chemical evolution with respect to the R5 and the X4 consensus sequence confirmed the finding, that the X4 consensus sequence dominated the population only transiently. In the illustrated example in Figure 5.9, this can be seen for the first 5,000 to 10,000 turns in the positions 22 and 25 of the simulation. During this time, the correct chemical group with respect to the X4 sequence dominated both positions, represented by the blue colour in the X4 (b) plot. In consequence, the chemical group in the positions 22 and 25 differed from the R5 consensus sequence during that time, represented by the red colour in the R5 plot (a). After approximately 12,000 turns, the R5 consensus aa dominated, leading to a switch of the chemical property, represented by the positive score (blue colour) in the R5 plot (a) and the negative score (red colour) in the X4 plot (b).
Figure 5.9 also nicely visualises the fluctuating character of the position-specific aa of the population. Once the correct aa in a sequence positions occurred, it dominated that position, but some individuals of the population continuously acquired new mutations during later rounds of replication (illustrated by the change between dark and light blue in position 25 beginning from turn 25,000).
From the analyses of the chemical course of the evolution we found that our model was able to simulate chemically meaningful evolutionary pathways.



(a)



(b)

Figure 5.9.: **Evolution of chemical properties**
The figure illustrates the evolution of the chemical properties during the simulation. Amino acids are translated position-dependent into a chemical score, with respect to the consensus sequence. Matching amino acids are scored as 1.0, differing amino acids belonging to the same chemical group than the consensus amino acid are scored as 0.0 and differing amino acids of a differing chemical group are scored as -1.0.
The plots depict the position-specific average of the sampled population. (a) illustrates the course of the chemical properties with respect to the R5 consensus sequence, and (b) with respect to the X4 consensus sequence.

5.4. Discussion

In the third part of the work we developed a model to simulate the sequence evolution of the HIV-1 V3 loop. The simulation tool is based on the findings from the previous parts of the project, i.e. the sequence analyses, the observed evolutionary course of the evolution, and the fitness function to describe the replicative fitness of the simulated sequences.
The model is flexible and enables the user to examine the influence of changing simulation parameters on the timely course of the evolution, for example the impact of the mutation rate, of the simulation time, of the population size, or the effect of different homogeneous or heterogeneous founder populations.
We determined a set of parameters that serve as a default model to simulate the evolution of the V3 loop over time. Our analyses showed that the default model is able to evolve a population of sequences towards the observed R5 and/or X4 consensus sequence. The evolutionary speed of the simulation depends on the composition of the founder population, the mutation rate, and the size of the population.
We further found that the model is capable to reproduce the evolutionary course of the viral diversity and divergence over time, as it was described by Shankarappa *et al.* [138]. Further analyses confirmed that the simulated evolutionary pathways follow sensible chemical paths, first accumulating amino acids with correct chemical properties and finally converging into the correct position specific consensus amino acid.
Upon the inspection of the final steady state populations of the default model, we detected that the majority of the sequences resided at an amino acid Hamming distance of one to the consensus sequence. The exact distance is defined by the underlying mutation rate (default 0.016). We found that the final Hamming distance is a consequence of the continuous mutation process. The sequences are driven by the selection process towards the fitness optimum, which is the consensus sequence. Once the fitness peak is reached, the sequences slide down on the fitness landscape to positions of lower fitness. From the lower level of the fitness landscape, they again step uphill towards the optimum, creating a continuous exchange of the sequences at the top of the fitness peak and the sequences at the observed Hamming distance.
Based on our observations we state that our *in silico* simulation model is a useful tool to study V3 loop evolution. The model is able to cope with the restrictions of *in vivo* or *in vitro* studies, e.g. the decreasing patient compliance during the years of a longitudinal study, the high costs of the deep sequencing approaches to enable analyses of complete viral populations, or the sensitivity of the sequencing methods.

5.5. Outlook

In further studies, we would like to focus on the simulation of the co-receptor switch and to evaluate the neutral networks generated by our simulation tool. We want to analyse the network changes that happen during the simulated evolution. By a variation of the simulation parameters during a running simulation, analyses of dynamic fitness landscapes could be possible. This property has already been implemented into the simulation. Last but not least, a combination of the simulation tool with a co-receptor prediction method is also highly recommended to enable co-receptor predictions during the simulation.
We would further like to combine the simulation data with longitudinal studies of *in vivo* HIV evolution. First, it would be very interesting to study HIV evolution in humanised mice. Humanised mice are mice in which the mouse-specific immune system (in parts) is replaced by a human immune system or at least a couple of human immune cells. These studies would enable us to decide whether the simulated evolutionary pathways are comparable with the viral evolution in a model organism. We could further imagine to compare the intra-host evolution of a longitudinal study of HIV-infected patients with the simulated evolution.
To do so, the simulation should be supplemented with a cell dynamics model, mimicking the availability of CCR5 and CXCR4 positive T cells as well as the level of viral load. This extension of the model was used in early version of the simulation tool and we were already able to reproduce the dynamics of susceptible and infected (SI) cells.
A further extension could model the adaptation of the immune system to the viral sequences via the integration of a death or clearance process based on adapted immune cells. The presented model is further prepared to utilise ageing information of virions. Some very first test were made to distinguish between productively and latently infected cells and to introduce compartmentalisation into the model (i.e. to compare the viral populations in different body compartments). Last but not least, the model could be extended to mimic HAART treatment, for example by the introduction of a fitness benefit for drug specific amino acid mutations during the course of simulation.

6. Discussion

6.1. Discussion

The present PhD project is a composed of three intertwined parts. While the first part is based directly on longitudinal data of 36 HIV-infected patients, the second part utilised a cross-section of biological sequence data from 2,768 patients comprised in the Los Alamos HIV data base. In the third part, the findings were combined to develop an *in silico* tool to describe intra-host HIV evolution.

Though the general clinical and evolutionary course of an HIV-1 infection was known from the early days of HIV research, so far it was not analysed whether the course of infection of HAART treated patients is comparable to the course of untreated patients or patients with antiretroviral monotherapy. We found some first indications that the course of the infection of treated and untreated patients is highly related, but the timely course of the infection varies, depending on the success of the administered therapy.

In vitro co-receptor studies further revealed that intermediate blinks of X4-tropic viral strains can be suppressed by successful HAART therapy. Thus, we hypothesise that both the course of infection and the co-receptor switch are not one- but bi-directional. Our first findings should be confirmed in further longitudinal studies.

Based on approximately 80,000 Los Alamos V3 loop sequences we next analysed the genetic differences between the R5- and the X4-tropic strains, since the conditions that lead to a co-receptor switch in about 50% of all patients are still obscure. Therefore we derived two independent fitness functions to describe the replicative fitness of R5- and X4-tropic V3 loop sequences. Based on the fitness functions, we used methods from graph theory to analyse the underlying fitness landscapes.

With our studies, we were able to confirm the differences between the R5 and the X4 sequence space of the V3 loop, that were described in prior publications, e.g. by Bozek *et al.* [15]. Analysing the sequence conservation of our data set, we could show that the R5-tropic sequences are more conserved than the X4-tropic sequences, a fact that was also described earlier.

In addition, we found that the weakest conserved sequence positions deviate between the R5 and the X4 sequences, leading to far-reaching consequences. Detailed studies of the weakest conserved sequence positions and the respective nucleotide codons showed that the most weakly conserved X4 sequence codons are evolutionary close to stop codons. In several of the weakest conserved codons, only one nucleotide mutation suffices to introduce a stop codon into the respective X4-tropic V3 loop sequence. In contrast, none of the weakest conserved nucleotides of the R5 consensus sequence enables a mutation that introduced a stop codon into the R5 sequence.

Following these observations we hypothesise that the less conserved X4 sequences have a fitness disadvantage in a setting with high immune pressure. The better immune recognition of X4-tropic sequences yields an additional disadvantage. Based on these ideas,

the R5-tropic sequences dominate early in the infection due to a mutational robustness and a worse immune recognition, while a weak immune pressure is a prerequisite to enable the occurrence and replication of X4-tropic sequences. Thus, we suppose that the co-receptor switch can be explained by the immune control hypothesis, in combination with the observation of a weakly conserved X4 population.

We comprised our observations in the third part of this work by the development of an *in silico* tool to simulate the sequence evolution of the HIV V3 loop. The heart of the simulation is the fitness function derived in the second part of the work. The method is at a early state and is still missing some important factors that influence *in vivo* viral evolution (e.g. strength of immune pressure, availability of R5 and X4 target cells, drug-virus interactions). Therefore we were not able to reproduce the course of the viral evolution of our study patients. So far, the simulations enabled us to analyse some evolutionary processes of simulated viral populations in fast time and at low costs. With our *in silico* model we were able to mimic the course of the evolution of HIV populations with respect to the viral diversity and the divergence. Further analyses showed that the simulated viral evolution followed a chemically sensible course.

In summary, the present project gave us some clues regarding the mechanisms of the co-receptor switch and was successful to gain a number of new and interesting insights into the aspects of viral intra-host HIV-1 evolution and into the genetic differences between the R5-tropic and the X4-tropic viral strains.

Bibliography

[1] H. Abdi and L. J. Williams. Principal component analysis. *CompStats*, 2(4):433–459, Jul 2010.

[2] J. Aldrich. Correlations Genuine and Spurious in Pearson and Yule. *Statistical Science*, 10(4):364–376, 1995.

[3] K. Allers, G. Hutter, J. Hofmann, C. Loddenkemper, K. Rieger, E. Thiel, and T. Schneider. Evidence for the cure of HIV infection by CCR5Δ32/Δ32 stem cell transplantation. *Blood*, 117(10):2791–2799, Mar 2011.

[4] D. G. Altman and J. M. Bland. Absence of evidence is not evidence of absence. *BMJ*, 311(7003):485, Aug 1995.

[5] P. Artimo, M. Jonnalagedda, K. Arnold, D. Baratin, G. Csardi, E. de Castro, S. Duvaud, V. Flegel, A. Fortier, E. Gasteiger, A. Grosdidier, C. Hernandez, V. Ioannidis, D. Kuznetsov, R. Liechti, S. Moretti, K. Mostaguir, N. Redaschi, G. Rossier, I. Xenarios, and H. Stockinger. ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Res.*, 40(Web Server issue):597–603, Jul 2012.

[6] I. Bahar, A. R. Atilgan, and B. Erman. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold Des*, 2(3):173–181, 1997.

[7] C. Balasubramanian, G. Chillemi, I. Abbate, M. R. Capobianchi, G. Rozera, and A. Desideri. Importance of V3 Loop Flexibility and Net Charge in the Context of Co-Receptor Recognition. A Molecular Dynamics Study on HIV gp120. *J. Biomol. Struct. Dyn.*, 29(5):1–13, Apr 2012.

[8] G. Barr. *Perl 'List::Util' Package*, 1997–2007. Licensed under the LGPL version 2.

[9] F. Barre-Sinoussi, J. C. Chermann, F. Rey, M. T. Nugeyre, S. Chamaret, J. Gruest, C. Dauguet, C. Axler-Blin, F. Vezinet-Brun, C. Rouzioux, W. Rozenbaum, and L. Montagnier. Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science*, 220(4599):868–871, May 1983.

[10] J. E. Bennett, A. Racine-Poon, J. C. Wakefield, W. R. eds. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London, 1996. ISBN 0 412 05551 1.

[11] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Res.*, 28(1):235–242, Jan 2000. `http://www.pdb.org`.

[12] F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, 112(3):535–542, May 1977.

[13] S. Beyer, R. Ortalo, and J. Leto. *Perl 'Math::MatrixReal' Package*, 1996-2002. Licensed under the LGPL version 2.

[14] D. Binninger-Schinzel, D. Muller, T. Wolf, B. Krause, B. Meye, G. Winskowsky, S. Raupp, S. Norley, R. Brodt, and A. Werner. Characterization of a chemokine receptor CCR5-negative T cell line and its use in determining human immunodeficiency virus type 1 phenotype. *J. Med. Virol.*, 80:192–200, Feb 2008.

[15] K. Bozek, A. Thielen, S. Sierra, R. Kaiser, and T. Lengauer. V3 loop sequence space analysis suggests different evolutionary patterns of CCR5- and CXCR4-tropic HIV. *PLoS ONE*, 4(10):e7387, 2009.

[16] R. Bron, P. J. Klasse, D. Wilkinson, P. R. Clapham, A. Pelchen-Matthews, C. Power, T. N. Wells, J. Kim, S. C. Peiper, J. A. Hoxie, and M. Marsh. Promiscuous use of CC and CXC chemokine receptors in cell-to-cell fusion mediated by a human immunodeficiency virus type 2 envelope protein. *J. Virol.*, 71(11):8405–8415, Nov 1997.

[17] I. N. Bronstein. *Taschenbuch der Mathematik*. Verlag Harri Deutsch. ISBN 3-80-855670-6.

[18] B. Brooks and M. Karplus. Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proc. Natl. Acad. Sci. U.S.A.*, 80(21):6571–6575, Nov 1983.

[19] D. M. Cardo, D. H. Culver, C. A. Ciesielski, P. U. Srivastava, R. Marcus, D. Abiteboul, J. Heptonstall, G. Ippolito, F. Lot, P. S. McKibben, and D. M. Bell. A case-control study of HIV seroconversion in health care workers after percutaneous exposure. Centers for Disease Control and Prevention Needlestick Surveillance Group. *N. Engl. J. Med.*, 337(21):1485–1490, Nov 1997.

[20] L. L. Cavalli-Sforza and A. W. Edwards. Phylogenetic analysis. Models and estimation procedures. *Am. J. Hum. Genet.*, 19(3 Pt 1):233–257, May 1967.

[21] Centers for Disease Control and Prevention, Atlanta, USA. Current Trends Human T-Lymphotropic Virus Type III / Lymphadenopathy-Associated Virus: Agent Summary Statement. *Morbidity and Mortality Weekly Report*, 35(34):540–542, 547–549, Aug 1986.

[22] B. Chandramouli, G. Chillemi, I. Abbate, M. R. Capobianchi, G. Rozera, and A. Desideri. Importance of V3 loop flexibility and net charge in the context of co-receptor recognition. A molecular dynamics study on HIV gp120. *J. Biomol. Struct. Dyn.*, 29(5):879–891, 2012.

[23] B. Charlesworth. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat. Rev. Genet.*, 10(3):195–205, Mar 2009.

[24] C. Cheng-Mayer, M. Quiroga, J. W. Tung, D. Dina, and J. A. Levy. Viral determinants of human immunodeficiency virus type 1 T-cell or macrophage tropism, cytopathogenicity, and CD4 antigen modulation. *J. Virol.*, 64(9):4390–4398, Sep 1990.

[25] R. Chenna, H. Sugawara, T. Koike, R. Lopez, T. J. Gibson, D. G. Higgins, and J. D. Thompson. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.*, 31(13):3497–3500, Jul 2003.

[26] T. Christiansen, B. D. Foy, L. Wall, and J. Orwant. *Programming Perl*. O'Reilly Media, 4th edition, 2012. ISBN 978-0596004927.

[27] M. Clerici, N. I. Stocks, R. A. Zajac, R. N. Boswell, D. R. Lucey, C. S. Via, and G. M. Shearer. Detection of three distinct patterns of T helper cell dysfunction in asymptomatic, human immunodeficiency virus-seropositive patients. Independence of CD4+ cell numbers and clinical staging. *J. Clin. Invest.*, 84(6):1892–1899, Dec 1989.

[28] J. Coffin, A. Haase, J. A. Levy, L. Montagnier, S. Oroszlan, N. Teich, H. Temin, K. Toyoshima, H. Varmus, and P. Vogt. What to call the AIDS virus? *Nature*, 321(6065):10, 1986.

[29] R. I. Connor, K. E. Sheridan, D. Ceradini, S. Choe, and N. R. Landau. Change in coreceptor use correlates with disease progression in HIV-1–infected individuals. *J. Exp. Med.*, 185(4):621–628, Feb 1997.

[30] H. Cramer. *Mathematical Methods of Statistics*. Princeton University Press, Princeton, 1999. ISBN 0-69-100547-8.

[31] G. E. Crooks, G. Hon, J. M. Chandonia, and S. E. Brenner. WebLogo: a sequence logo generator. *Genome Res.*, 14(6):1188–1190, Jun 2004.

[32] V. Dahirel, K. Shekhar, F. Pereyra, T. Miura, M. Artyomov, S. Talsania, T. M. Allen, M. Altfeld, M. Carrington, D. J. Irvine, B. D. Walker, and A. K. Chakraborty. Coordinate linkage of HIV evolution reveals regions of immunological vulnerability. *Proc. Natl. Acad. Sci. U.S.A.*, 108(28):11530–11535, Jul 2011.

[33] C. R. Darwin. *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. John Murray, London, first edition, 1859. John van Wyhe, ed. 2002-. The Complete Work of Charles Darwin Online.

[34] M. P. Davenport, J. J. Zaunders, M. D. Hazenberg, H. Schuitemaker, and R. P. van Rij. Cell turnover and cell tropism in HIV-1 infection. *Trends Microbiol.*, 10(6):275–278, Jun 2002.

[35] P. Delobel, K. Sandres-Saune, M. Cazabat, C. Pasquier, B. Marchou, P. Massip, and J. Izopet. R5 to X4 switch of the predominant HIV-1 population in cellular reservoirs during effective highly active antiretroviral therapy. *J. Acquir. Immune Defic. Syndr.*, 38(4):382–392, Apr 2005.

[36] H. Deng, R. Liu, W. Ellmeier, S. Choe, D. Unutmaz, M. Burkhart, P. Di Marzio, S. Marmon, R. E. Sutton, C. M. Hill, C. B. Davis, S. C. Peiper, T. J. Schall, D. R. Littman, and N. R. Landau. Identification of a major co-receptor for primary isolates of HIV-1. *Nature*, 381(6584):661–666, Jun 1996.

[37] R. W. Doms and D. Trono. The plasma membrane as a combat zone in the HIV battlefield. *Genes Dev.*, 14(21):2677–2688, Nov 2000.

[38] B. J. Doranz, J. Rucker, Y. Yi, R. J. Smyth, M. Samson, S. C. Peiper, M. Parmentier, R. G. Collman, and R. W. Doms. A dual-tropic primary HIV-1 isolate that uses fusin and the beta-chemokine receptors CKR-5, CKR-3, and CKR-2b as fusion cofactors. *Cell*, 85(7):1149–1158, Jun 1996.

[39] A. J. Drummond, G. K. Nicholls, A. G. Rodrigo, and W. Solomon. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*, 161(3):1307–1320, Jul 2002.

[40] A. J. Drummond and A. Rambaut. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.*, 7:214, 2007.

[41] A. J. Drummond and A. Rambaut. *Tracer, version 1.4*, 2007. [Online; accessed 19-December-2011].

[42] A. J. Drummond, A. Rambaut, B. Shapiro, and O. G. Pybus. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.*, 22(5):1185–1192, May 2005.

[43] A. J. Drummond, A. Rambaut, and M. Suchard. *BEASTWiki*. [Online; accessed 19-December-2011].

[44] M. P. D'Souza and V. A. Harden. Chemokines and HIV-1 second receptors. Confluence of two fields generates optimism in AIDS research. *Nat. Med.*, 2(12):1293–1300, Dec 1996.

[45] M. Eigen. Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften*, 58(10):465–523, Oct 1971.

[46] M. Eigen. The origin of genetic information: viruses as models. *Gene*, 135(1-2):37–47, Dec 1993.

[47] M. Eigen, J. McCaskill, and P. Schuster. Molecular Quasi-Species. *J. Phys. Chem.*, 92:6881–6891, Jun 1988.

[48] L. Euler. Solutio problematis ad geometriam situs pertinentis. *Commentarii Academiae Scientiarum Imperialis Petropolitanae*, 8:128–140, 1736. Opera Omnia (1) 7 (1911-56), 1-10.

[49] G. Fatkenheuer and et al. Subgroup analyses of maraviroc in previously treated R5 HIV-1 infection. *N. Engl. J. Med.*, 359(14):1442–1455, Oct 2008.

[50] A. S. Fauci, G. Pantaleo, S. Stanley, and D. Weissman. Immunopathogenic mechanisms of HIV infection. *Ann. Intern. Med.*, 124(7):654–663, Apr 1996.

[51] A. S. Fauci, S. M. Schnittman, G. Poli, S. Koenig, and G. Pantaleo. NIH conference. Immunopathogenic mechanisms in human immunodeficiency virus (HIV) infection. *Ann. Intern. Med.*, 114(8):678–693, Apr 1991.

[52] W. Fischer, V. V. Ganusov, E. E. Giorgi, P. T. Hraber, B. F. Keele, T. Leitner, C. S. Han, C. D. Gleasner, L. Green, C. C. Lo, A. Nag, T. C. Wallstrom, S. Wang, A. J. McMichael, B. F. Haynes, B. H. Hahn, A. S. Perelson, P. Borrow, G. M. Shaw, T. Bhattacharya, and B. T. Korber. Transmission of single HIV-1 genomes and dynamics of early immune escape revealed by ultra-deep sequencing. *PLoS ONE*, 5(8):e12303, 2010.

[53] R. A. Fisher. *The Design of Experiments*. Macmillan Pub Co, 9 edition, 1971. ISBN 0-028-44690-9.

[54] R. A. Fouchier, M. Groenink, N. A. Kootstra, M. Tersmette, H. G. Huisman, F. Miedema, and H. Schuitemaker. Phenotype-associated sequence variation in the third variable domain of the human immunodeficiency virus type 1 gp120 molecule. *J. Virol.*, 66(5):3183–3187, May 1992.

[55] A. D. Frankel and J. A. Young. HIV-1: fifteen proteins and an RNA. *Annu. Rev. Biochem.*, 67:1–25, 1998.

[56] G. Fuellen, R. Resnick, S. E. Brenner, C. Dagdigian, S. Chervitz, E. Birney, J. Gilbert, and E. o. Stupka. *Perl 'Bioperl' Package*, 1996-2009.

[57] S. Gavrilets. Evolution and speciation on holey adaptive landscapes. *Trends Ecol. Evol. (Amst.)*, 12(8):307–312, Aug 1997.

[58] L. Y. Geer, A. Marchler-Bauer, R. C. Geer, L. Han, J. He, S. He, C. Liu, W. Shi, and S. H. Bryant. The NCBI BioSystems database. *Nucleic Acids Res.*, 38(Database issue):D492–496, Jan 2010.

[59] G. B. Gloor, L. C. Martin, L. M. Wahl, and S. D. Dunn. Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. *Biochemistry*, 44(19):7156–7165, May 2005.

[60] N. Go, T. Noguti, and T. Nishikawa. Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proc. Natl. Acad. Sci. U.S.A.*, 80(12):3696–3700, Jun 1983.

[61] S. N. Goodman. Toward evidence-based medical statistics. 2: The Bayes factor. *Ann. Intern. Med.*, 130(12):1005–1013, Jun 1999.

[62] P. R. Gorry and P. Ancuta. Coreceptors and HIV-1 pathogenesis. *Curr HIV/AIDS Rep*, 8(1):45–53, Mar 2011.

[63] R. Grant and . co workers. Preexposure chemoprophylaxis forHIV prevention in men who have sex with men. *N. Engl. J. Med.*, 363(27):2587–99, Dec 2010.

[64] C. Graziosi, G. Pantaleo, J. F. Demarest, O. J. Cohen, M. Vaccarezza, L. Butini, M. Montroni, and A. S. Fauci. HIV-1 infection in the lymphoid organs. *AIDS*, 7 Suppl 2:S53–58, Nov 1993.

[65] X. Gu, Y. X. Fu, and W. H. Li. Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. *Mol. Biol. Evol.*, 12(4):546–557, Jul 1995.

[66] L. A. Guay, P. Musoke, T. Fleming, D. Bagenda, M. Allen, C. Nakabiito, J. Sherman, P. Bakaki, C. Ducar, M. Deseyve, L. Emel, M. Mirochnick, M. G. Fowler, L. Mofenson, P. Miotti, K. Dransfield, D. Bray, F. Mmiro, and J. B. Jackson. Intrapartum and neonatal single-dose nevirapine compared with zidovudine for prevention of mother-to-child transmission of HIV-1 in Kampala, Uganda: HIVNET 012 randomised trial. *Lancet*, 354(9181):795–802, Sep 1999.

[67] R. M. Gulick and et al. Maraviroc for previously treated patients with R5 HIV-1 infection. *N. Engl. J. Med.*, 359(14):1429–1441, Oct 2008.

[68] A. T. Haase. Population biology of HIV-1 infection: viral and CD4+ T cell demographics and dynamics in lymphatic tissues. *Annu. Rev. Immunol.*, 17:625–656, 1999.

[69] A. A. Hagberg, D. A. Schult, and P. J. Swart. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15, Pasadena, CA USA, Aug. 2008.

[70] B. H. Hahn, G. M. Shaw, S. K. Arya, M. Popovic, R. C. Gallo, and F. Wong-Staal. Molecular cloning and characterization of the HTLV-III virus associated with AIDS. *Nature*, 312(5990):166–169, 1984.

[71] R. W. Hamming. Error Detecting and Error Correcting Codes. *The Bell System Technical Journal*, 29(2):147–160, Apr 1950.

[72] R. W. Hamming. *Coding and Information Theory*. Prentice Hall, second edition, 2002. ISBN 0131390724.

[73] A. Hildebrandt, A. K. Dehof, A. Rurainski, A. Bertsch, M. Schumann, N. C. Toussaint, A. Moll, D. Stockel, S. Nickels, S. C. Mueller, H. P. Lenhof, and O. Kohlbacher. BALL–biochemical algorithms library 1.3. *BMC Bioinformatics*, 11:531, 2010.

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

[75] A. Hodgkinson and A. Eyre-Walker. Variation in the mutation rate across mammalian genomes. *Nat. Rev. Genet.*, 12(11):756–766, Nov 2011.

[76] F. Hoffgaard. *Biomolecular Correlation in Physical and Sequence Space*. Technische Universität Darmstadt, 2011. PhD thesis.

[77] F. Hoffgaard, P. Weil, and K. Hamacher. BioPhysConnectoR: Connecting sequence information and biophysical models. *BMC Bioinformatics*, 11:199, 2010.

[78] C. C. Huang, S. N. Lam, P. Acharya, M. Tang, S. H. Xiang, S. S. Hussan, R. L. Stanfield, J. Robinson, J. Sodroski, I. A. Wilson, R. Wyatt, C. A. Bewley, and P. D. Kwong. Structures of the CCR5 N terminus and of a tyrosine-sulfated antibody with HIV-1 gp120 and CD4. *Science*, 317(5846):1930–1934, Sep 2007.

[79] C. C. Huang, M. Tang, M. Y. Zhang, S. Majeed, E. Montabana, R. L. Stanfield, D. S. Dimitrov, B. Korber, J. Sodroski, I. A. Wilson, R. Wyatt, and P. D. Kwong. Structure of a V3-containing HIV-1 gp120 core. *Science*, 310(5750):1025–1028, Nov 2005.

[80] G. Hutter, D. Nowak, M. Mossner, S. Ganepola, A. Mussig, K. Allers, T. Schneider, J. Hofmann, C. Kucherer, O. Blau, I. W. Blau, W. K. Hofmann, and E. Thiel. Long-term control of HIV by CCR5 Delta32/Delta32 stem-cell transplantation. *N. Engl. J. Med.*, 360(7):692–698, Feb 2009.

[81] M. A. Jensen, M. Coetzer, A. B. van 't Wout, L. Morris, and J. I. Mullins. A reliable phenotype predictor for human immunodeficiency virus type 1 subtype C based on envelope V3 sequences. *J. Virol.*, 80(10):4698–4704, May 2006.

[82] M. A. Jensen, F. S. Li, A. B. van't Wout, D. C. Nickle, D. Shriner, H. X. He, S. McLaughlin, R. Shankarappa, J. B. Margolick, and J. I. Mullins. Improved coreceptor usage prediction and genotypic monitoring of R5-to-X4 transition by motif analysis of human immunodeficiency virus type 1 env V3 loop sequences. *J. Virol.*, 77(24):13376–13388, Dec 2003.

[83] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabasi. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, Oct 2000.

[84] Joint United Nations Programme on HIV/AIDS (UNAIDS). *UNAIDS Report on the global AIDS epidemic*. WHO Library Cataloguing-in-Publication Data, 20 Avenue Appia, CH-1211 Geneva 27, Switzerland, 2012. ISBN 978-92-9173-996-7.

[85] C. Kamp, T. Wolf, I. G. Bravo, B. Kraus, B. Krause, B. Neumann, G. Winskowsky, A. Thielen, A. Werner, and B. S. Schnierle. Decreased HIV diversity after allogeneic stem cell transplantation of an HIV-1 infected patient: a case report. *Virol. J.*, 7:55, 2010.

[86] J. Karn and C. M. Stoltzfus. Transcriptional and posttranscriptional regulation of HIV-1 gene expression. *Cold Spring Harb Perspect Med*, 2(2):a006916, Feb 2012.

[87] R. E. Kass and A. E. Raftery. Bayes Factors. *J Am Stat Asoc.*, 90(430):773–795, Jun 1995.

[88] S. A. Kauffman. Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol.*, 22(3):437–467, Mar 1969.

[89] M. S. Killian and J. A. Levy. HIV/AIDS: 30 years of progress and future challenges. *Eur. J. Immunol.*, 41(12):3401–3411, Dec 2011.

[90] J. Kingman. The coalescent. *Stochastic Processes and their Applications*, 13(3):235 – 248, 1982.

[91] J. F. Kingman. Origins of the coalescent. 1974-1982. *Genetics*, 156(4):1461–1463, Dec 2000.

[92] A. Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. *G. Ist. Ital. Attuari*, 4:83, 1933.

[93] B. T. Korber, R. M. Farber, D. H. Wolpert, and A. S. Lapedes. Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. *Proc. Natl. Acad. Sci. U.S.A.*, 90(15):7176–7180, Aug 1993.

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

[95] S. Kumar, K. Tamura, and M. Nei. MEGA: Molecular Evolutionary Genetics Analysis software for microcomputers. *Comput. Appl. Biosci.*, 10(2):189–191, Apr 1994.

[96] C. Lanave, G. Preparata, C. Saccone, and G. Serio. A new method for calculating evolutionary substitution rates. *J. Mol. Evol.*, 20(1):86–93, 1984.

[97] T. Lengauer, O. Sander, S. Sierra, A. Thielen, and R. Kaiser. Bioinformatics prediction of HIV coreceptor usage. *Nat. Biotechnol.*, 25:1407–1410, Dec 2007.

[98] J. Levy. *HIV and the Pathogenesis of AIDS*. ASM Press, Washington, DC, third edition, 2007. ISBN 1-55581-393-3.

[99] J. A. Levy. HIV pathogenesis: 25 years of progress and persistent challenges. *AIDS*, 23(2):147–160, Jan 2009.

[100] Los Alamos National Laboratory, Los Alamos, NM 87545. *Los Alamos database*. [Online; accessed 15-May-2012].

[101] A. J. Low, W. Dong, D. Chan, T. Sing, R. Swanstrom, M. Jensen, S. Pillai, B. Good, and P. R. Harrigan. Current V3 genotyping algorithms are inadequate for predicting X4 co-receptor usage in clinical isolates. *AIDS*, 21(14):17–24, Sep 2007.

[102] L. M. Mansky and H. M. Temin. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *J. Virol.*, 69(8):5087–5094, Aug 1995.

[103] L. C. Martin, G. B. Gloor, S. D. Dunn, and L. M. Wahl. Using information theory to search for co-evolving residues in proteins. *Bioinformatics*, 21(22):4116–4124, Nov 2005.

[104] J. W. Mellors, A. Munoz, J. V. Giorgi, J. B. Margolick, C. J. Tassoni, P. Gupta, L. A. Kingsley, J. A. Todd, A. J. Saah, R. Detels, J. P. Phair, and C. R. Rinaldo. Plasma viral load and CD4+ lymphocytes as prognostic markers of HIV-1 infection. *Ann. Intern. Med.*, 126(12):946–954, Jun 1997.

[105] P. Miller. *Perl 'Statistics::Basic' Package*, 2012. Licensed under the LGPL version 2.

[106] A. Moll, A. Hildebrandt, H. P. Lenhof, and O. Kohlbacher. BALLView: an object-oriented molecular visualization and modeling framework. *J. Comput. Aided Mol. Des.*, 19(11):791–800, Nov 2005.

[107] A. Moll, A. Hildebrandt, H. P. Lenhof, and O. Kohlbacher. BALLView: a tool for research and education in molecular modeling. *Bioinformatics*, 22(3):365–366, Feb 2006.

[108] P. Moran. Random processes in genetics. *Mathematical Proceedings of the Cambridge Philosophical Society*, 54:60–71, Jan 1958.

[109] H. J. Muller. Further studies on the nature and causes of gene mutations. *Proceedings of the 6th International Congress of Genetics*, pages 213–255, 1932.

[110] M. Newton and A. Raftery. Approximate Bayesian inference by the weighted likelihood bootstrap (with Discussion). *J R Stat Soc Series B*, 56:3–48, 1994.

[111] J. Neyman and E. S. Pearson. On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of The Royal Society A: Mathematical, Physical and Engineering Sciences*, 231:289–337, 1933.

[112] O. A. Okwundu, C. I.and Uthman and O. C. A. Antiretroviral pre-exposure prophylaxis (PrEP) for preventing HIV in high-risk individuals. *J Evid Based Med*, 5(3):186, Aug 2012.

[113] C. Pastore, A. Ramos, and D. E. Mosier. Intrinsic obstacles to human immunodeficiency virus type 1 coreceptor switching. *J. Virol.*, 78(14):7565–7574, Jul 2004.

[114] K. Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Phil. Mag.*, 50(302):157–175, 1900.

[115] K. Pearson. On lines and planes of closest fit to systems of points in space. *Phil. Mag.*, 2:559–572, 1901.

[116] D. Persaud, H. Gay, C. Ziemniak, Y. H. Chen, M. Piatak, T.-W. Chun, M. Strain, D. Richman, and K. Luzuriaga. Functional HIV cure after very early ART of an infected infant. *Program and abstracts of the 20th Conference on Retroviruses and Opportunistic Infections*, Mar 2013. Abstract 48LB.

[117] G. Piganeau, R. Westrelin, B. Tourancheau, and C. Gautier. Multiplicative versus additive selection in relation to genome evolution: a simulation study. *Genet. Res.*, 78(2):171–175, Oct 2001.

[118] E. Poveda, E. Seclen, M. d. e. l. M. Gonzalez, F. Garcia, N. Chueca, A. Aguilera, J. J. Rodriguez, J. Gonzalez-Lahoz, and V. Soriano. Design and validation of new genotypic tools for easy and reliable estimation of HIV tropism before using CCR5 antagonists. *J. Antimicrob. Chemother.*, 63(5):1006–1010, May 2009.

[119] M. C. Prosperi, L. Bracciale, M. Fabbiani, S. Di Giambenedetto, F. Razzolini, G. Meini, M. Colafigli, A. Marzocchetti, R. Cauda, M. Zazzi, and A. De Luca. Comparative determination of HIV-1 co-receptor tropism by Enhanced Sensitivity Trofile, gp120 V3-loop RNA and DNA genotyping. *Retrovirology*, 7:56, 2010.

[120] Python Software Foundation. *The Python Language Reference*, 2010. Version 2.7.

[121] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0, [Online; accessed 19-December-2011].

[122] A. Rambaut. *FigTree*, 2009. version 1.2.3.

[123] R. R. Regoes and S. Bonhoeffer. The HIV coreceptor switch: a population dynamical perspective. *Trends Microbiol.*, 13(6):269–277, Jun 2005.

[124] R. M. Ribeiro, M. D. Hazenberg, A. S. Perelson, and M. P. Davenport. NaÃŕve and memory cell turnover as drivers of CCR5-to-CXCR4 tropism switch in human immunodeficiency virus type 1: implications for therapy. *J. Virol.*, 80(2):802–809, Jan 2006.

[125] N. E. Riddick, E. A. Hermann, L. M. Loftin, S. T. Elliott, W. C. Wey, B. Cervasi, J. Taaffe, J. C. Engram, B. Li, J. G. Else, Y. Li, B. H. Hahn, C. A. Derdeyn, D. L. Sodora, C. Apetrei, M. Paiardini, G. Silvestri, and R. G. Collman. A novel CCR5 mutation common in sooty mangabeys reveals SIVsmm infection of CCR5-null natural hosts and efficient alternative coreceptor use in vivo. *PLoS Pathog.*, 6(8):e1001064, 2010.

[126] J. D. Roberts, K. Bebenek, and T. A. Kunkel. The accuracy of reverse transcriptase from HIV-1. *Science*, 242(4882):1171–1173, Nov 1988.

[127] A. G. Rodrigo. Dynamics of syncytium-inducing and non-syncytium-inducing type 1 human immunodeficiency viruses during primary infection. *AIDS Res. Hum. Retroviruses*, 13(17):1447–1451, Nov 1997.

[128] Rohrer. Aids-Therapie aus dem Computer. *MaxPlanck- Forschung*, 3:22, Mar 2005. `http://www.mpg.de/979037/MPF_2005_3`.

[129] K. H. Rosen. *Discrete Mathematics and its Applications*. McGraw-Hill, New York, fifth edition, 2003. ISBN 0-07-242434-6.

[130] T. S. Some probabilistic and statisical problems on the analysis of DNA sequences. *Lect Math Life Sci*, 17:57–86, 1986.

[131] A. Saez-Cirion, C. Bacchus, L. Hocqueloux, V. Avettand-Fenoel, I. Girault, C. Lecuroux, V. Potard, P. Versmisse, A. Melard, T. Prazuck, B. Descours, J. Guergnon, J.-P. Viard, F. Boufassa, O. Lambotte, C. Goujard, L. Meyer, D. Costagliola, A. Venet, G. Pancino, B. Autran, C. Rouzioux, and the ANRS VISCONTI Study Group. Post-Treatment HIV-1 Controllers with a Long-Term Virological Remission after the Interruption of Early Initiated Antiretroviral Therapy ANRS VISCONTI Study. *PLoS Pathog.*, 9(3):e1003211, 2013.

[132] S. A. Sawyer, J. Parsch, Z. Zhang, and D. L. Hartl. Prevalence of positive selection among nearly neutral amino acid replacements in Drosophila. *Proc. Natl. Acad. Sci. U.S.A.*, 104(16):6504–6510, Apr 2007.

[133] E. Schneider, S. Whitmore, K. M. Glynn, K. Dominguez, A. Mitsch, and M. T. McKenna. Revised surveillance case definitions for HIV infection among adults, adolescents, and children aged ¡18 months and for HIV infection and AIDS among children aged 18 months to ¡13 years–United States, 2008. *MMWR Recomm Rep*, 57(RR-10):1–12, Dec 2008.

[134] T. D. Schneider and R. M. Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, 18(20):6097–6100, Oct 1990.

[135] H. Schuitemaker, M. Koot, N. A. Kootstra, M. W. Dercksen, R. E. de Goede, R. P. van Steenwijk, J. M. Lange, J. K. Schattenkerk, F. Miedema, and M. Tersmette. Biological phenotype of human immunodeficiency virus type 1 clones at different stages of infection: progression of disease is associated with a shift from monocytotropic to T-cell-tropic virus population. *J. Virol.*, 66(3):1354–1360, Mar 1992.

[136] P. Schuster and J. Swetina. Stationary mutant distributions and evolutionary optimization. *Bull. Math. Biol.*, 50(6):635–660, 1988.

[137] E. Seclen, V. Soriano, M. M. Gonzalez, S. Gomez, A. Thielen, and E. Poveda. High concordance between the position-specific scoring matrix and geno2pheno algorithms for genotypic interpretation of HIV-1 tropism: V3 length as the major cause of disagreement. *J. Clin. Microbiol.*, 49(9):3380–3382, Sep 2011.

[138] R. Shankarappa, J. B. Margolick, S. J. Gange, A. G. Rodrigo, D. Upchurch, H. Farzadegan, P. Gupta, C. R. Rinaldo, G. H. Learn, X. He, X. L. Huang, and J. I. Mullins. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *J. Virol.*, 73:10489–10502, Dec 1999.

[139] C. E. Shannon. The mathematical theory of communication. *Bell Syst. Tech. J.*, 27(3):379–423, Jul 1948.

[140] M. Sharon, N. Kessler, R. Levy, S. Zolla-Pazner, M. Gorlach, and J. Anglister. Alternative conformations of HIV-1 V3 loops mimic beta hairpins in chemokines,

suggesting a mechanism for coreceptor selectivity. *Structure*, 11(2):225–236, Feb 2003.

[141] T. Shioda, J. A. Levy, and C. Cheng-Mayer. Small amino acid changes in the V3 hypervariable region of gp120 can affect the T-cell-line and macrophage tropism of human immunodeficiency virus type 1. *Proc. Natl. Acad. Sci. U.S.A.*, 89(20):9434–9438, Oct 1992.

[142] R. F. Siliciano and W. C. Greene. HIV latency. *Cold Spring Harb Perspect Med*, 1(1):a007096, Sep 2011.

[143] N. V. Smirnov. Tables for estimating the goodness of fit of empirical distributions. *Ann. Math. Stat.*, 19:279, 1948.

[144] A. Stamatakis. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690, Nov 2006.

[145] A. Stamatakis, P. Hoover, and J. Rougemont. A Fast Bootstrapping Algorithm for the RAxML Web-Servers. *Systematic Biology*, 57(5):758–771, 2008.

[146] T. Strachan and A. P. Read. *Human Molecular Genetics*. Wiley-Liss, Bios Scientific Publishers, New York, 2. edition, 1999. ISBN 1859962025.

[147] K. Tamura, J. Dudley, M. Nei, and S. Kumar. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.*, 24:1596–1599, Aug 2007.

[148] M. C. Thigpen and et al. Antiretroviral preexposure prophylaxis for heterosexual HIV transmission in Botswana. *N. Engl. J. Med.*, 367(5):423–434, Aug 2012.

[149] J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22(22):4673–4680, Nov 1994.

[150] A. N. Tikhonov. Solution of incorrectly formulated problems and the regularization method. *Soviet Math.*, 4:1035–1038, 1963.

[151] L. P. Vandekerckhove and et al. European guidelines on the clinical management of HIV-1 tropism testing. *Lancet Infect Dis*, 11(5):394–407, May 2011.

[152] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.

[153] W. F. Vranken, M. Budesinsky, F. Fant, K. Boulez, and F. A. Borremans. The complete Consensus V3 loop peptide of the envelope protein gp120 of HIV-1 shows pronounced helical character in solution. *FEBS Lett.*, 374(1):117–121, Oct 1995.

[154] P. J. Waddell and M. A. Steel. General time-reversible distances with unequal rates across sites: mixing gamma and inverse Gaussian distributions with invariant sites. *Mol. Phylogenet. Evol.*, 8(3):398–414, Dec 1997.

[155] R. S. Waples. Genetic methods for estimating the effective size of cetacean populations. *Report of the international whaling commision*, 13:279–300, 1991.

[156] J. D. Watson. *Molecular Biology of the Gene*. Cold Spring Harbor Laboratory Press, 6 edition, 2007. ISBN 080539592X.

[157] J. Weber. The pathogenesis of HIV-1 infection. *Br. Med. Bull.*, 58:61–72, 2001.

[158] P. Weil. *Koevolution in molekularen Komplexen*. Technische Universität Darmstadt, 2012. PhD thesis.

[159] A. Werner, D. Binninger-Schinzel, D. Muller, B. Krause, B. Meye, G. Winskowsky, T. Wolf, R. Brodt, and B. Schnierle. Evolution of HIV Geno- and Phenotypes during Anti-Retroviral Therapy. *European Journal of Medical Research*, 12:12–13, Aug 2007.

[160] WHO. *WHO Case Definitions of HIV for Surveillance and Revised Clinical Staging and Immunological Classification of HIV-Related Disease in Adults and Children*. WHO Library Cataloguing-in-Publication Data, Geneva, Switzerland, 2007. ISBN 978 92 4 159562 9.

[161] C. O. Wilke, C. Ronnewinkel, and T. Martinetz. Dynamic fitness landscapes in molecular evolution. *Phys. Rep.*, 349(5):395–446, Aug 2001.

[162] T. Wolf, V. Rickerts, S. Staszewski, S. Kriener, B. Wassmann, G. Bug, M. Bickel, P. Gute, H. R. Brodt, and H. Martin. First case of successful allogeneic stem cell transplantation in an HIV-patient who acquired severe aplastic anemia. *Haematologica*, 92:e56–58, Apr 2007.

[163] S. Wright. Evolution in Mendelian Populations. *Genetics*, 16(2):97–159, Mar 1931.

[164] S. Wright. The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proc. of the VI International Congress of Genetrics:*, 1:356–366, 1932.

[165] L. Xiao, S. M. Owen, I. Goldman, A. A. Lal, J. J. deJong, J. Goudsmit, and R. B. Lal. CCR5 coreceptor usage of non-syncytium-inducing primary HIV-1 is independent of phylogenetically distinct global HIV-1 isolates: delineation of consensus motif in the V3 domain that predicts CCR-5 usage. *J. Virol.*, 240(1):83–92, Jan 1998.

[166] Z. Yu, M. J. Dobro, C. L. Woodward, A. Levandovsky, C. M. Danielson, V. Sandrin, J. Shi, C. Aiken, R. Zandi, T. J. Hope, and G. J. Jensen. Unclosed hiv-1 capsids suggest a curled sheet model of assembly. *Journal of Molecular Biology*, 425(1):112 – 123, 2013.

[167] L. Zhang, T. He, Y. Huang, Z. Chen, Y. Guo, S. Wu, K. J. Kunstman, R. C. Brown, J. P. Phair, A. U. Neumann, D. D. Ho, and S. M. Wolinsky. Chemokine coreceptor usage by diverse primary isolates of human immunodeficiency virus type 1. *J. Virol.*, 72(11):9307–9312, Nov 1998.

A. Appendix

A.1. Supplementary results

A.1.1. Pearson correlation coefficient

Table A.1.: **Pearson correlation of diversity and CD4$^+$ cell count** The table lists the Pearson correlation values r and the respective p-values for the association of the diversity and the CD4$^+$ cell count. 'ID' is the unique patient identifier and the column 'obs.' gives the number of observations.

ID	obs.	r	p-value	ID	obs.	r	p-value
4	8	-0.168	0.691	98	7	0.183	0.694
5	6	-0.363	0.479	100	4	-0.967	0.033
7	4	-0.419	0.581	107	8	0.025	0.952
10	5	-	-	109	3	0.023	0.985
13	6	0.855	0.030	127	7	0.845	0.017
24	4	-0.501	0.499	132	4	0.109	0.891
26	3	0.978	0.133	143	3	0.322	0.791
32	6	0.135	0.799	166	5	-0.145	0.816
40	5	-0.022	0.973	178	3	0.816	0.393
41	6	0.081	0.879	180	12	0.356	0.256
43	3	-	-	190	5	0.665	0.220
49	5	0.851	0.067	194	7	0.307	0.504
51	5	0.299	0.624	196	4	-0.377	0.623
62	5	-0.082	0.895	197	5	0.517	0.373
68	4	-0.748	0.252	212	5	0.342	0.573
72	7	0.252	0.585	222	3	-0.530	0.644
85	3	-0.975	0.141	265	11	0.031	0.927
97	3	0.863	0.337	268	7	0.015	0.975

Table A.2.: **Pearson correlation of divergence and CD4$^+$ cell count** The table lists the Pearson correlation values r and the respective p-values for the association of the divergence and the CD4$^+$ cell count. 'ID' is the unique patient identifier and the column 'obs.' gives the number of observations.

ID	obs.	r	p-value	ID	obs.	r	p-value
4	8	-0.725	0.042	98	7	-0.010	0.983
5	6	0.519	0.291	100	4	-0.893	0.107
7	4	0.545	0.455	107	8	0.343	0.405
10	5	-	-	109	3	0.185	0.881
13	6	0.564	0.244	127	7	0.837	0.019
24	4	0.111	0.889	132	4	-0.741	0.259
26	3	-0.983	0.119	143	3	0.727	0.482
32	6	0.465	0.353	166	5	0.194	0.755
40	5	0.099	0.874	178	3	0.925	0.247
41	6	0.058	0.913	180	12	0.443	0.149
43	3	-	-	190	5	-0.273	0.657
49	5	0.840	0.075	194	7	0.160	0.732
51	5	0.602	0.283	196	4	-0.836	0.164
62	5	-0.358	0.554	197	5	0.705	0.184
68	4	0.939	0.061	212	5	0.577	0.309
72	7	0.286	0.535	222	3	-0.496	0.670
85	3	0.459	0.696	265	11	-0.015	0.966
97	3	0.629	0.567	268	7	0.470	0.287

Table A.3.: **Pearson correlation of viral load and diversity** The table lists the Pearson correlation values r and the respective p-values for the association of the viral load and the diversity. 'ID' is the unique patient identifier and the column 'obs.' gives the number of observations.

ID	obs.	r	p-value	ID	obs.	r	p-value
4	7	-0.203	0.663	98	7	-0.445	0.317
5	5	-0.586	0.299	100	4	0.748	0.252
7	3	1.000	0.015	107	8	0.377	0.357
10	5	-	-	109	3	-0.390	0.745
13	6	-0.742	0.092	127	7	-0.133	0.776
24	4	-0.068	0.932	132	6	0.261	0.617
26	3	-0.233	0.850	143	3	0.744	0.466
32	6	-0.158	0.765	166	5	0.357	0.555
40	3	0.861	0.340	178	4	-0.260	0.740
41	5	0.339	0.577	180	9	-0.089	0.820
43	3	-	-	190	5	0.305	0.618
49	5	-0.439	0.459	194	7	-0.026	0.956
51	5	-0.938	0.019	196	5	0.060	0.924
62	5	-0.533	0.355	197	6	0.262	0.616
68	3	0.993	0.074	212	5	0.259	0.674
72	6	-0.379	0.459	222	4	-	-
85	7	-	-	265	9	0.143	0.713
97	4	-0.162	0.838	268	6	-0.457	0.362

Table A.4.: **Pearson correlation of viral load and N_e]** The table lists the Pearson correlation values r and the respective p-values for the association of the viral load and the effective population size N_e. 'ID' is the unique patient identifier and the column 'obs.' gives the number of observations.

ID	obs.	r	p-value	ID	obs.	r	p-value
4	32	-0.178	0.330	98	21	-0.758	<0.001
5	30	-0.806	<0.001	100	11	0.799	0.003
7	23	-0.009	0.969	107	36	0.080	0.642
10	10	0.344	0.330	109	10	-0.333	0.347
13	15	-0.805	<0.001	127	23	0.145	0.509
24	29	0.148	0.445	132	19	0.861	<0.001
26	14	-0.044	0.882	143	21	0.973	<0.001
32	28	-0.241	0.216	166	28	0.722	<0.001
40	19	0.539	0.017	178	10	-0.151	0.677
41	23	0.425	0.043	180	39	0.495	0.001
43	11	0.775	0.005	190	11	0.715	0.013
49	28	-0.399	0.036	194	34	-0.150	0.398
51	21	-0.467	0.033	196	35	-0.587	<0.001
62	34	-0.158	0.373	197	20	-0.372	0.107
68	31	0.075	0.689	212	13	0.518	0.070
72	20	-0.336	0.148	222	5	-0.137	0.826
85	40	-0.605	<0.001	265	26	0.438	0.025
97	18	-0.315	0.202	268	11	-0.311	0.352

Table A.5.: **Graph measures of 6-point mutant networks.** The table lists the minimal node fitness, the maximal node betweenness, the maximal node closeness and the average shortest path length of the respective R5 and X4 six-point mutant networks, depending on the percentage of the most fit nodes (%) included into the six-point mutant networks.

	min. fit.		max. bet.		max. close.		avg. length	
%	R5	X4	R5	X4	R5	X4	R5	X4
1	0.692	0.750	$1.2 \cdot 10^2$	$2.9 \cdot 10^3$	0.476	0.283	2.349	3.519
2	0.592	0.697	$6.9 \cdot 10^3$	$1.9 \cdot 10^3$	0.404	0.317	3.008	3.231
3	0.570	0.672	$6.4 \cdot 10^3$	$1.8 \cdot 10^3$	0.397	0.320	2.973	3.194
4	0.569	0.653	$6.4 \cdot 10^3$	$1.6 \cdot 10^3$	0.376	0.307	3.043	3.250
5	0.544	0.630	$4.1 \cdot 10^3$	$1.4 \cdot 10^3$	0.356	0.294	3.120	3.309
6	0.542	0.604	$4.0 \cdot 10^3$	$1.3 \cdot 10^3$	0.352	0.282	3.153	3.389
7	0.525	0.587	$3.3 \cdot 10^3$	$1.1 \cdot 10^3$	0.340	0.276	3.193	3.456
8	0.500	0.568	$2.5 \cdot 10^3$	$9.6 \cdot 10^4$	0.343	0.270	3.340	3.517
9	0.448	0.537	$2.0 \cdot 10^3$	$7.9 \cdot 10^4$	0.341	0.262	3.446	3.612
10	0.423	0.507	$1.6 \cdot 10^3$	$6.4 \cdot 10^4$	0.333	0.252	3.542	3.712
11	0.421	0.489	$1.7 \cdot 10^3$	$5.6 \cdot 10^4$	0.328	0.245	3.591	3.777
12	0.420	0.481	$1.6 \cdot 10^3$	$5.0 \cdot 10^4$	0.318	0.251	3.641	3.817
13	0.401	0.472	$1.4 \cdot 10^3$	$4.5 \cdot 10^4$	0.309	0.254	3.684	3.873
14	0.401	0.466	$1.3 \cdot 10^3$	$4.3 \cdot 10^4$	0.303	0.256	3.690	3.907
15	0.376	0.457	$1.2 \cdot 10^3$	$3.8 \cdot 10^4$	0.300	0.256	3.725	3.942
16	0.346	0.451	$1.1 \cdot 10^3$	$3.6 \cdot 10^4$	0.296	0.256	3.773	3.962
17	0.331	0.444	$9.6 \cdot 10^4$	$3.3 \cdot 10^4$	0.296	0.256	3.812	3.990
18	0.322	0.438	$8.5 \cdot 10^4$	$3.1 \cdot 10^4$	0.294	0.254	3.860	4.009
19	0.314	0.432	$7.6 \cdot 10^4$	$2.9 \cdot 10^4$	0.291	0.252	3.903	4.029
20	0.310	0.426	$7.3 \cdot 10^4$	$2.7 \cdot 10^4$	0.288	0.252	3.927	4.049
21	0.310	0.422	$7.3 \cdot 10^4$	$2.7 \cdot 10^4$	0.288	0.251	3.965	4.064
22	0.310	0.418	$7.2 \cdot 10^4$	$2.6 \cdot 10^4$	0.286	0.249	4.005	4.082
23	0.301	0.414	$6.6 \cdot 10^4$	$2.6 \cdot 10^4$	0.283	0.251	4.027	4.098
24	0.299	0.411	$7.2 \cdot 10^4$	$2.5 \cdot 10^4$	0.283	0.250	4.036	4.111
25	0.299	0.408	$7.1 \cdot 10^4$	$2.5 \cdot 10^4$	0.282	0.249	4.045	4.122

Table A.6.: **Graph measures of 8-point mutant networks** The table lists the minimal node fitness, the maximal node betweenness, the maximal node closeness and the average shortest path length of the respective R5 and X4 six-point mutant networks, depending on the percentage of the most fit nodes (%) included into the eight-point mutant networks.

	min. fit.		max. bet.		max. close.		avg. length	
%	R5	X4	R5	X4	R5	X4	R5	X4
1	0.401	0.659	$1.0 \cdot 10^3$	$7.0 \cdot 10^4$	0.317	0.228	3.818	4.096
2	0.302	0.579	$4.1 \cdot 10^4$	$4.1 \cdot 10^4$	0.288	0.211	4.301	4.373
3	0.285	0.520	$3.1 \cdot 10^4$	$2.1 \cdot 10^4$	0.272	0.198	4.448	4.661
4	0.257	0.486	$2.3 \cdot 10^4$	$1.7 \cdot 10^4$	0.265	0.190	4.598	4.826
5	0.245	0.462	$2.0 \cdot 10^4$	$1.2 \cdot 10^4$	0.261	0.195	4.672	4.951
6	0.231	0.444	$1.7 \cdot 10^4$	$1.0 \cdot 10^4$	0.257	0.194	4.753	5.042
7	0.213	0.427	$1.3 \cdot 10^4$	$8.7 \cdot 10^5$	0.251	0.191	4.879	5.111
8	0.192	0.414	$1.1 \cdot 10^4$	$7.1 \cdot 10^5$	0.245	0.191	4.971	5.161
9	0.188	0.403	$9.9 \cdot 10^5$	$7.1 \cdot 10^5$	0.240	0.192	5.027	5.200
10	0.179	0.395	$8.8 \cdot 10^5$	$6.9 \cdot 10^5$	0.235	0.191	5.065	5.230
11	0.163	0.385	$7.5 \cdot 10^5$	$6.5 \cdot 10^5$	0.232	0.191	5.136	5.261
12	0.157	0.373	$6.8 \cdot 10^5$	$5.9 \cdot 10^5$	0.229	0.189	5.189	5.299
13	0.152	0.361	$6.1 \cdot 10^5$	$5.4 \cdot 10^5$	0.227	0.188	5.232	5.338
14	0.147	0.353	$5.7 \cdot 10^5$	$5.0 \cdot 10^5$	0.225	0.187	5.273	5.374
15	0.142	0.344	$5.2 \cdot 10^5$	$4.6 \cdot 10^5$	0.223	0.185	5.313	5.408
16	0.139	0.334	$4.7 \cdot 10^5$	$4.3 \cdot 10^5$	0.221	0.183	5.350	5.439
17	0.139	0.325	$4.5 \cdot 10^5$	$3.9 \cdot 10^5$	0.220	0.181	5.383	5.469
18	0.135	0.317	$4.4 \cdot 10^5$	$3.6 \cdot 10^5$	0.219	0.180	5.402	5.496
19	0.134	0.308	$4.3 \cdot 10^5$	$3.4 \cdot 10^5$	0.217	0.179	5.418	5.520
20	0.133	0.304	$4.0 \cdot 10^5$	$3.3 \cdot 10^5$	0.216	0.178	5.443	5.540
21	0.131	0.296	$3.8 \cdot 10^5$	$3.2 \cdot 10^5$	0.215	0.177	5.464	5.557
22	0.128	0.290	$3.7 \cdot 10^5$	$3.0 \cdot 10^5$	0.214	0.177	5.478	5.574
23	0.127	0.282	$3.5 \cdot 10^5$	$2.9 \cdot 10^5$	0.213	0.176	5.494	5.590
24	0.125	0.277	$3.4 \cdot 10^5$	$2.8 \cdot 10^5$	0.212	0.176	5.517	5.608
25	0.121	0.271	$3.1 \cdot 10^5$	$2.6 \cdot 10^5$	0.211	0.176	5.540	5.622

A.2. Amino acid code and chemical properties

Table A.7.: Amino acid code and chemical properties

A	ALA	Alanine	non-polar
C	CYS	Cysteine	polar
D	ASP	Aspartic Acid	acidic
E	GLU	Glutamic Acid	acidic
F	PHE	Phenylalanine	non-polar
G	GLY	Glycine	polar
H	HIS	Histidine	basic
I	ILE	Isoleucine	non-polar
K	LYS	Lysine	basic
L	LEU	Leucine	non-polar
M	MET	Methionine	non-polar
N	ASN	Asparagine	polar
P	PRO	Proline	non-polar
Q	GLN	Glutamine	polar
R	ARG	Arginine	basic
S	SER	Serine	polar
T	THR	Threonine	polar
V	VAL	Valine	non-polar
W	TRP	Tryptophan	non-polar
Y	TYR	Tyrosine	polar

Table A.8.: Nucleotide code

A	Adenine	T	Thymine
C	Cytosine	G	Guanine

A.3. Analytical determination of the mutation rate

The 64 codons are translated into 21 different symbols, 20 amino acids and 3 stop codons.
First codon position:
19 symbols are altered into a differing symbol upon any nt mutation in the first codon position ($19 \cdot \frac{3}{4}$), two symbols are altered due to a mutation into two of four possible nts ($2 \cdot \frac{2}{4}$).

$$19 \cdot \frac{3}{4} + 2 \cdot \frac{2}{4} = \frac{61}{4}$$

Second codon position:
19 symbols are altered into a differing symbol upon any nt mutation in the second codon position ($19 \cdot \frac{3}{4}$), two symbols are altered due to a mutation into two of four possible nts ($2 \cdot \frac{2}{4}$).

$$19 \cdot \frac{3}{4} + 2 \cdot \frac{2}{4} = \frac{61}{4}$$

Third codon position: two symbols are altered into a differing symbol upon any nt mutation in the third codon position ($2 \cdot \frac{3}{4}$), ten symbols are altered due to a mutation into two of four possible nts ($10 \cdot \frac{2}{4}$), one symbol is altered only upon one specific nt ($1 \cdot \frac{1}{4}$), and eight symbols are not altered by any nt mutation in the third codon position ($8 \cdot \frac{0}{4}$).

$$2 \cdot \frac{3}{4} + 10 \cdot \frac{2}{4} + 1 \cdot \frac{1}{4} + 8 \cdot \frac{0}{4} = \frac{27}{4}$$

In summary, these considerations result in:

$$\frac{61}{4} + \frac{61}{4} + \frac{27}{4} = \frac{149}{4}$$

Based on 21 symbols, the result is divided by 21, resulting in a value of 1.7738. This number describes, that a mutation in the first, second, and third codon position in parallel alters on average 1.7738 symbols.
In consequence, we need on average $3 \cdot \frac{1}{1.7738} = 1.69$ nucleotide mutations to produce an average of one amino acid mutation per replication. Adapted to a nt sequence length of 105, this consideration results in a mutation rate of $\frac{1.69}{105} = 0.016$.

A.4. Alphabetical list of frequently used abbreviations

aa	amino acid
ART	antiretroviral therapy
bp	basepair
CCR5	one of the two predominant HIV-1 co-receptors, often used in the early to latent phase of infection
CXCR4	one of the two predominant HIV-1 co-receptors, in about 50% of patients used in latent to late phase of infection
CC	cross correlation
CD4$^+$	T-cell bearing CD4 receptor
CDC	United States Centers for Disease Control and Prevention
DNA	deoxyribonucleic acid
FPR	false positive rate
HAART	highly active antiretroviral therapy
nt	nucleotide
MI	mutual information
MSA	multiple sequence alignment
MRCA	most recent common ancestor
mRNA	messenger RNA
PBMC	peripheral blood mononuclear cell
PCR	polymerase chain reaction
R5	short form of the CCR5 co-receptor
RNA	ribonucleic acid
SUMI	subset mutual information
X4	short form of the CXCR4 co-receptor
V3	third variable loop, a specific region of the HIV-1 envelope gene, which is crucial in co-receptor binding and cell entry
HIV (HIV-1)	human immunodeficiency virus (type 1)
AIDS	acquired immunodeficiency syndrome
env	envelope gene of HIV-1
gp	glycoprotein
RT	reverse transcriptase, a central HIV protein
WHO	World Health Organization

Last but not least ...

Vielen Dank an alle, die mich durch meine Promotionszeit begleitet haben!
Der erste Dank gebührt Christel, Barbara und Peter, die mir die Chance gaben, unter ihrer Anleitung an einem tollen Thema zu arbeiten und die meine Doktorarbeit stets mit Interesse und zahlreichen guten Ideen und Ratschlägen unterstützt haben.
Danke an Kay, der mich als externe Doktorandin in seine Gruppe aufgenommen hat, obwohl er meist alle Hände voll zu tun hat. Die Diskussionen in der AG Hamacher haben mir so manches mal eine neue Perspektive auf meine Arbeit eröffnet. Nicht zu vergessen der wöchentliche Brunch, für den ich sogar ab und an ein Mittagessen mit Kollegen am PEI hab' sausen lassen. Aber nicht nur die Zeit mit euch, sondern auch mit euren zahlreichen Gäste war immer eine Bereicherung, und allzu oft musste ich die Runde verlassen, wenn ich gerne noch geblieben wäre.
Auch meinen Kollegen hier am PEI möchte ich herzlich für die gemeinsame Zeit danken. Trotz vieler Umzüge war ich in der Nachbarschaft jedes meiner wechselnden Büros stets willkommen und wurde mit offenen Armen empfangen. Auch wenn uns thematisch meist wenig verband, haben wir beim Mittagessen oder auf dem Flur am Drucker viele interessante Gespräche geführt.
Danke an Barbaras Gruppe, und insbesondere an Sarah, die meine Kenntnisse über die Methoden im Labor und die mit den Analysen verbundenen Schwierigkeiten erweitert haben. Danke an Jan, der sich jederzeit bereit erklärt hat, mit mir mathematische Fragestellungen zu erörtern oder zur rechten Zeit gemeinsam eine Pause von der Arbeit einzulegen. Danke an Kay für die Unterstützung bei den statistischen Auswertungen. Auch Susanne möchte ich auf diesem Weg nochmal danke sagen für die gemeinsamen Monate im Büro.
Ein besonderer Dank geht an die liebste Saarländerin im fernen Hessen. Es ist sehr schade, dass wir uns erst so spät beim Joggen im Wald begegnet sind. Auf die gemeinsamen Runden habe ich mich immer den ganzen Tag gefreut. Und auch Barbara hat uns so maches Mal begleitet, wenn es ihre Zeit und das Wetter zugelassen haben.
Meinen Freunden zu hause danke ich, dass sie trotz der wenigen gemeinsamen Zeit den Kontakt nicht haben einschlafen lassen. Ich hoffe sehr, dass wir in Zukunft wieder mehr Zeit miteinander verbringen können und dass der abgesagte gemeinsame Urlaub im nächsten Jahr statt findet.
Auch meine Familie und insbesondere mein Mann mussten viele Stunden während der letzten Jahre ohne mich verbringen. Danke, dass ihr auch über viele Kilometer Entfernung stets für mich da wart, euch gemeinsam mit mir gefreut und mir in schwierigen Zeiten Kraft gegeben habt - auch dann, wenn ich selbst manchmal nicht für euch da sein konnte, wenn ich es gerne gewesen wäre.
Danke auch an all die, die bisher unerwähnt geblieben sind. An Alexander Thielen für die Durchführung der Korezeptor Vorhersagen, an die projektverantwortlichen für die finanzielle Unterstützung dieser wichtigen Forschung, an meine Doktorandenkollegen hier am PEI, die mir die drei Jahre zu einer angenehmen Zeit gemacht haben.

Curriculum Vitae

Personal

Name	Miriam Carbon-Mangels
Date of Birth	24.09.1977
Place of Birth	Zweibrücken

Career

1984 - 1988	Grundschule Sechsmorgen, Zweibrücken
1988 - 1997	Helmholtz-Gymnasium, Zweibrücken
1997 - 2000	training on the job, Tax consultant assistant, Dr. Meyer und Partner, Pirmasens
2000 - 2005	Tax consultant assistant, StB Kämmerer, Kaiserslautern
2005 - 2010	study of Bioinformatics at the University of Saarland, Saarbrücken, BA-Thesis *In Silico Models for the Prediction of Metabolism by Cytochrome P450 Enzymes*, MA-Thesis *TNF alpha Antibody Design*
2010 - 2013	PhD at Paul-Ehrlich-Institut, Langen, and TU Darmstadt, *Intra-host HIV-1 evolution and the co-receptor switch*

Presentations

2010	3. HIV-Resistenz-Workshop, Berlin, Oral Presentation, *Ausblick HIV-Projekt*
	5th PEI Retreat, Kröckelbach, Oral Presentation, *Tracing intra-host HIV evolution by phylogenetic analysis*
2011	AREVIR-Meeting, Bonn, Oral Presentation, *Comparative study of intra-host HIV evolution: Emergence of viral diversity*
	12th International Conference on Systems Biology (ICSB), Mannheim, Poster Presentation, *Tracing intra-host HIV evolution by phylogenetic analysis*
2012	6th PEI Retreat, Tagungsstätte Löwenstein, Poster Presentation, *Correlating clinical and evolutionary parameters in intra-patient HIV evolution*
2013	7th PEI Retreat, Tagungsstätte Löwenstein, Oral Presentation, *Development of a fitness function for the V3 loop of HIV-1*