Dissertation

Bioinformatical Approaches to Ranking of anti-HIV Combination Therapies and Planning of Treatment Schedules

Author
André Altmann

Dissertation

zur Erlangung des Grades
des Doktors der Naturwissenschaften (Dr. rer. nat.)
der Naturwissenschaftlich–Technischen Fakultäten
der Universität des Saarlandes

Saarbrücken
2010

Tag des Kolloquiums:	8.6.2010
Dekan:	Prof. Dr. Holger Hermanns
Vorsitzender des Prüfungsausschusses:	Prof. Dr. Matthias Hein
Berichterstatter:	Prof. Dr. Thomas Lengauer, Ph.D.
	Prof. Dr. Jörg Rahnenführer
	Prof. Dr. Hans-Peter Lenhof
Beisitzer:	Dr. Francisco Domingues

Abstract

The human immunodeficiency virus (HIV) pandemic is one of the most serious health challenges humanity is facing today. Combination therapy comprising multiple antiretroviral drugs resulted in a dramatic decline in HIV-related mortality in the developed countries. However, the emergence of drug resistant HIV variants during treatment remains a problem for permanent treatment success and seriously hampers the composition of new active regimens.

In this thesis we use statistical learning for developing novel methods that rank combination therapies according to their chance of achieving treatment success. These depend on information regarding the treatment composition, the viral genotype, features of viral evolution, and the patient's therapy history. Moreover, we investigate different definitions of response to antiretroviral therapy and their impact on prediction performance of our method. We address the problem of extending purely data-driven approaches to support novel drugs with little available data. In addition, we explore the prospect of prediction systems that are centered on the patient's treatment history instead of the viral genotype. We present a framework for rapidly simulating resistance development during combination therapy that will eventually allow application of combination therapies in the best order.

Finally, we analyze surface proteins of HIV regarding their susceptibility to neutralizing antibodies with the aim of supporting HIV vaccine development.

Kurzfassung

Die Humane Immundefizienz-Virus (HIV) Pandemie ist eine der schwerwiegendsten gesundheitlichen Herausforderungen weltweit. Kombinationstherapien bestehend aus mehreren Medikamenten führten in entwickelten Ländern zu einem drastischen Rückgang der HIV-bedingten Sterblichkeit. Die Entstehung von Arzneimittel-resistenten Varianten während der Behandlung stellt allerdings ein Problem für den anhaltenden Behandlungserfolg dar und erschwert die Zusammenstellung von neuen aktiven Kombinationen.

In dieser Arbeit verwenden wir statistisches Lernen zur Entwicklung neuer Methoden, welche Kombinationstherapien bezüglich ihres erwarteten Behandlungserfolgs sortieren. Dabei nutzen wir Informationen über die Medikamente, das virale Erbgut, die Virus Evolution und die Therapiegeschichte des Patienten. Außerdem untersuchen wir unterschiedliche Definitionen für Therapieerfolg und ihre Auswirkungen auf die Güte unserer Modelle. Wir gehen das Problem der Erweiterung von daten-getriebenen Modellen bezüglich neuer Wirkstoffen an, und untersuchen weiterhin die Therapiegeschichte des Patienten als Ersatz für das virale Genom bei der Vorhersage. Wir stellen das Rahmenwerk für die schnelle Simulation von Resistenzentwicklung vor, welches schließlich erlaubt, die bestmögliche Reihenfolge von Kombinationstherapien zu suchen.

Schließlich analysieren wir das HIV Oberflächenprotein im Hinblick auf seine Anfälligkeit für neutralisierende Antikörper mit dem Ziel die Impfstoff Entwicklung zu unterstützen.

Acknowledgements

This work was carried out in the *Department for Computational Biology and Applied Algorithmics* at the *Max Planck Institute for Informatics* in Saarbrücken. Foremost, I would like to thank my advisor Prof. Thomas Lengauer for his encouragement, support, and advice throughout the years, and for creating an inspiring environment where I could follow my own ideas. Thank you for giving me the opportunity to participate in this great project. I am also grateful to Prof. Jörg Rahnenführer and Prof. Hans-Peter Lenhof, who kindly agreed to referee this thesis.

This work would not have been possible without the great number of even greater collaborators. I am much obliged to Robert W. Shafer (Stanford University), W. Jeffrey Fessel (Kaiser-Permanente Medical Care Program), and Klaus Überla (Ruhr-Universität Bochum). I am thankful to Rolf Kaiser and all members of his group at the Institute of Virology (University of Cologne), specifically Eugen Schülter, Nadine Sichtig, Melanie Balduin, and Jens Verheyen. I am grateful to the members of the Arevir consortium, foremost Rolf Kaiser, Martin Däumer, and Hauke Walter for sharing their knowledge on HIV. Thanks to all members of EuResist (IST-2004-027173) coordinated by Maurizio Zazzi and Francesca Incardona for creating this unique experience: especially Mattia Prosperi and Michal Rosen-Zvi for being partners in crime in developing the prediction engine.

Thanks to all former and present members of our group for making the department a great and fun place to work (and have breaks). Especially my former office mates, Christoph Bock, Alexander Thielen, and Kasia Bozek; Bastian Beggel, Jasmina Bogojeska, Alejandro Pironti, Hiroto Saigo, Oliver Sander, Laura Tolosi, Hendrik Weisser for all the joint projects we worked on; Alejandro Pironti, Alexander Thielen, and Francisco Domingues for proof-reading parts of this thesis; Christoph Bock, Francisco Domingues, Oliver Sander, Tobias Sing, and Alexander Thielen for the numerous discussions. Special thanks go to Niko Beerenwinkel, Tobias Sing, and Martin Däumer for their invaluable advice during the beginning of this work, and to Achim Büch and Ruth Schneppen-Christmann for their support and innumerable favors.

Yassen Assenov, Rayna Dimitrova, Konstantin Halachev, Christoph Hartmann, Andreas Kämper, Lars Kunert, Jochen Maydt, Oliver Sander, Tobias Sing, and Laura Tolosi were the main criminals for making life in Saarbrücken a very fun time. Thank you!

Furthermore, I would like to thank the students that worked with me on different projects: Fabian Müller for his excellent work in both his Bachelor's and Master's Thesis (congratulations for the Günter-Hotz medal!), Juliane Perner for her great work on her Bachelor's Thesis and her invaluable help for computing clinical cutoffs, and Peter Ebert for his superb job in his Bachelor's Thesis. It was great working with all of you. I wish you all the best!

Thanks go also to Jorge Cham for his PhD comics that are a continuous source of fun because they are so true and straight to the point of every-day PhD-life[1][2].

Finally, I would like to thank my family and friends for their love and support, especially my parents, Bärbel, and Günter. Above all though, I am grateful to Evangelia; for sharing ävery day with me: life is so much easier knowing you are by my side.

[1] http://www.phdcomics.com/comics/archive.php?comicid=1159
[2] http://www.phdcomics.com/comics/archive.php?comicid=1164

Contents

1 Introduction

Todays methods of modern molecular biology generate massive amounts of data. Proteomics focuses on the composition of proteins (our molecular machines), (epi-)genomics studies the influence of changes in a person's (epi-)genetic information on e.g. susceptibility to diseases, where transcriptomics analyzes which genes are active under certain conditions. Research conducted with these methods in the context of diseases often aims at providing personalized therapy.

In essence, personalized therapy refers to the delivery of the right drugs in the right dosage to the patient. Decisions regarding the selection of drugs are in general based on available biomarkers, e.g. the patient's genome; personalization ensures a (cost-)effective therapy. The ways leading to the "personalization" are manifold, for example:

1. Understanding the molecular basis of the disorder. It is crucial to understand how genetic changes lead to the disease, because deeper understanding will allow a classification of defects with similar phenotype into subgroups depending on their molecular origin. This, in turn, will lead to the development of better drugs targeted at the causes of the disease leading to shorter treatment times and reduced costs.

2. Understanding of pathways involved in drug metabolization. Here, genetic changes influencing the effective concentration of the drug are of interest. For instance, mutations resulting in overdosing of the drug increase the risk of side-effects, while mutations that lower the drug concentration may result in treatment failure.

3. Pathogen tailored treatment. Obviously, knowing the disease-causing pathogen allows the selection of better matched drugs for its eradication (e.g. targeted vs. broad-spectrum antibiotic). Moreover, resistance of pathogens against antimicrobial agents often requires screening for active compounds prior to treatment start.

Typically, *"omics"* strategies are employed for identifying disease-related biomarkers. In comparative studies, two groups of individuals (patients vs. healthy controls or responders vs. non-responders) are studied with proteomics, (epi-)genomics, and transcriptomics. In these settings, data are generated at a large-scale and their analysis is far beyond the feasibility of manual inspection. Thus, bioinformatics methods are needed. These methods often rely on methodologies from statistics and statistical/machine learning to separate healthy from sick individuals based on the given data. Moreover, bioinformatics methods are needed to design tools that analyze routinely generated *omics* data, and thereby bring personalized therapy into clinical routine.

In this thesis we focus on personalized treatment of infections with the human immunodeficiency virus (HIV). HIV therapy is one of the most prominent examples of personalized therapy today and belongs to the third category of personalized medicine: finding antiretroviral agents that are effective against the predominant viral variant in the host.

Personalized HIV therapy

Briefly, HIV mainly infects cells of the human immune system and, when untreated, HIV infections typically lead to the acquired immunodeficiency syndrome (AIDS) followed by death due to opportunistic infections. By now, AIDS accounts for an estimated annual 2 million deaths and thus poses one of the major worldwide health challenges.

Patients in developed countries have access to about two dozen antiretroviral drugs. These compounds interrupt the replication cycle of the virus and thereby allow for a partial recovery of the patient's immune system. Complete eradication of all viral particles from the patient is not possible. Moreover, treatment success as defined by undetectable amounts of virus in the blood can only be maintained for a limited time. The cause of the limited treatment success is the dynamic viral population within the host: high turnover rates and short replication cycles paired with an error-prone replication process lead to the evolutionary selection of mutations that reduce the susceptibility of the virus to the applied antiretroviral drug. These resistance mutations eventually render the antiretroviral drug useless for treating the patient.

To date, testing for drug resistance prior to prescription of antiretroviral compounds is highly recommended. Resistance can be assessed either by slow and expensive laboratory based tests in cell cultures or via fast and cheap standardized methods that determine the genetic sequence of the viral drug targets. For the output of the latter method, a multitude of tools provides classification for single drugs in terms of drug resistance.

In order to delay the emergence of resistance mutations, modern anti-HIV regimens combine at least three drugs with a minimum of two different mechanisms of action. The introduction of this highly active antiretroviral therapy (HAART) resulted in a dramatic decline of HIV-related mortality in developed countries. Resistance testing provides the pivotal tool for selecting active compounds against the viral population in the patient. Methods for resistance testing, however, still provide only recommendations for single drugs and thereby ignore the interplay between multiple drugs for attacking HIV.

Outline

We begin in Chapter 2 with a brief overview of the discovery of HIV, followed by a summary of the current pandemic. Then, we continue with a description of the HIV particle, its replication cycle, and clinical aspects of the HIV infection. This is followed by a review of state-of-the-art HIV treatment including a detailed description of the viral drug targets. We continue with a brief summary on HIV resistance testing and available genotype interpretation methods. We conclude the chapter with an outlook on potential advancements towards richer decision support systems.

In Chapter 3 we present ways for improving and updating GENO2PHENO, a web service that predicts drug resistance from the viral genome.

In Chapter 4 we start with a motivation for interpretation systems that consider combination therapies and summarize previously developed features that encode viral evolution. Then we present geno2pheno-THEO, a freely available interpretation tool, that predicts clinical response to combination therapy using information on the treatment, the virus, and evolutionary features. We continue with a retrospective clinical validation of geno2pheno-

THEO, where we compare the prediction performance to three widely-used expert-based interpretation systems. Moreover, we study the robustness of geno2pheno-THEO and explore the benefit of regimen-specific models.

Chapter 5 begins with an introduction to the EuResist project, which is followed by a presentation of the three decision support systems for inferring short-term *in vivo* response to combination therapy developed within the project. We continue with a description of the EuResist prediction engine including combination approaches, performance assessment (including human experts as reference), and the web service. We proceed with an investigation on alternative definitions for response to combination treatment and study the performance of resulting prediction engines. Then we approach the update-problem, which constitutes the Achilles' heel of data-driven decision support systems for HIV combination therapy. We conclude the chapter with the presentation of a prediction system that uses the patient's treatment history instead of the viral genotype to infer response to treatment.

In Chapter 6 we motivate the requirement for planning sequences of HIV therapies. We present transducers, a family of finite state machines, that constitute the computational framework for our novel method to rapidly simulate viral evolution during combination therapy. Furthermore, we introduce five mutation models that are based on *in vitro* and *in vivo* resistance data. The framework is studied in two validation settings. In the first setting we challenge the framework to separate failing from succeeding treatments on the basis of mutations caused by the preceding regimen. Within the second setting we assess whether overall resistance development is correctly reflected by resistance development within drug classes. The chapter is concluded by a case study comparing the impact of typical first-line regimens on future drug options.

Chapter 7 makes the transition from personalized anti-HIV therapy to the probably only true solution for the HIV pandemic: an HIV vaccine. In the beginning of the chapter we summarize the basic concept of the human immune system comprising innate and adaptive immune response. This summary is followed by a short description of how vaccines establish immunity. Further, we list the challenges posed by HIV to vaccine design and review the most prominent vaccine candidates. The focus of the chapter is the prediction of neutralization by antibodies based on the genetic sequence of the viral genome. Analysis of the statistical learning models reveals positions that are essential for antibody binding. We conclude the chapter with potential improvements for our approach and further applications to the development of HIV vaccines.

Chapter 8 concludes the thesis and provides an outlook on further advancements on personalized HIV therapy.

2 AIDS, HIV, and Antiretroviral Therapy

"In the field of observation, chance favours the prepared mind." – Louis Pasteur

This chapter aims at providing the necessary background for this thesis on the human immunodeficiency virus (HIV), the acquired immunodeficiency syndrome (AIDS), and state-of-the-art antiretroviral treatment with the focus on available antiretroviral drugs and routine treatment guidance. However, many important aspects that are not of direct relevance for this thesis had to be omitted.

2.1 The AIDS pandemic

In 1981 a mysterious disease was discovered among the gay community in the United States of America. The patients were hospitalized due to diseases resulting from an impaired immune system. Owing to the media, the illness quickly became known as GRID (gay-related immunodeficiency). The cause of the disease was unknown and the race for its discovery commenced. The first effective tool to screen for a possible infection with the unknown agent was a positive test for an infection with the Hepatitis B virus (HBV).

Roughly one year after the first cases of this immunodeficiency in the USA were reported, the virus was identified and isolated by researchers at the Institute Pasteur in Paris (France). For their discovery of the infectious agent that is now known as the Human Immunodeficiency Virus (HIV), Luc Montagnier and Françoise Barré-Sinoussi were awarded the Nobel-Price for Physiology or Medicine in 2008 (Barré-Sinoussi et al., 1983). HIV infects cells of the human immune system and eventually leads to their death, thus explaining the typical symptoms of immunodeficiency in HIV infected patients. Robert C. Gallo, who also claimed to be the first discoverer of the infectious agent causing AIDS was not awarded the Nobel Price (Gallo et al., 1983). Indeed, it turned out that the probes investigated by Gallo's group were co-infected with a Human T cell Leukemia Virus (HTLV), and thus "the paper by the Montaginer/Chermann group is unequivocally the first reported true isolation of HIV from a patient with lymphadenopathy" (Gallo, 2002). Gallo, however, demonstrated that the identified agent is the cause of AIDS (Popovic et al., 1984) and developed the first diagnostic blood test that prevented new infections and probably saved thousands of lives. The omission of Gallo from the prestigious Nobel Price, despite his numerous major contributions in the early years, resulted in a massive response by the international scientific community (Abbadessa et al., 2009). The history of the discovery of HIV has been briefly summarized by the main contributers Montagnier (2002) and Gallo (2002) in featured articles.

The period following the identification of HIV (which in the early works was termed LAV or HTLV-III due to its believed relation to the group of HTL viruses) as the cause of AIDS was part of the scientific success story of modern medicine. Just to name a few milestones:

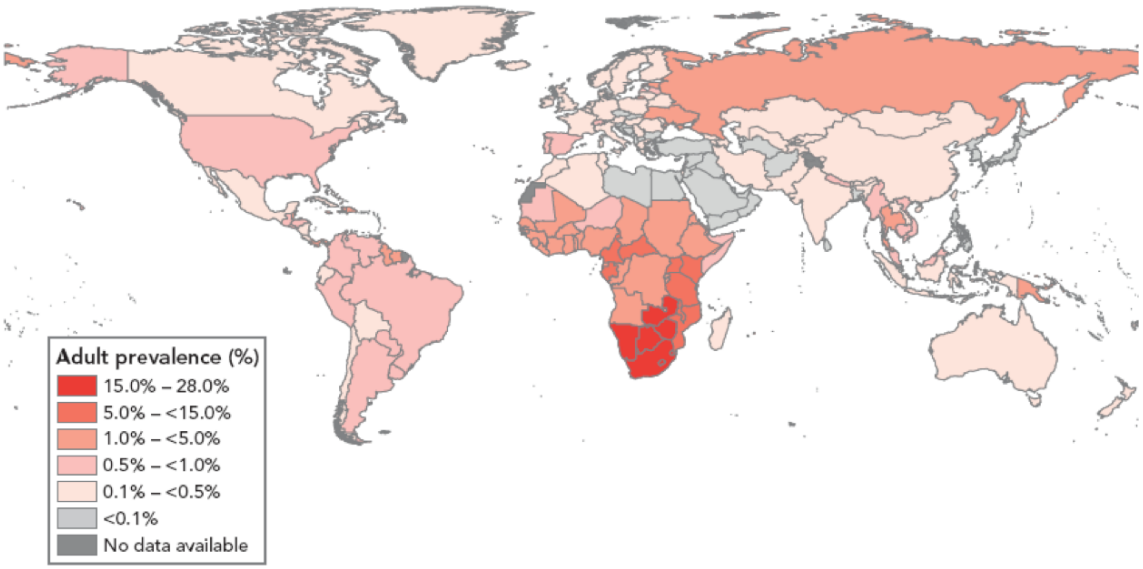


Figure 2.1: Global prevalence of HIV in December 2007. The figure was adapted from the WHO 2008 Report on the global AIDS epidemic available at http://www.who.int/hiv/data/en/. By the end of 2008 at total of 33.4 million [31.1-35.8] people were estimated to live with HIV.

the genome was sequenced and inter- as well as intra-patient variation in viral populations was detected (Shaw et al., 1984; Ratner et al., 1985; Wong-Staal et al., 1985; Hahn et al., 1986), the virus was found in the brain of AIDS patients (Shaw et al., 1985), the modes of HIV transmission were elucidated, all of HIV's genes and most proteins were defined. But most importantly, the blood supplies in most developed nations were rendered safe by screening for HIV. The capability of screening for HIV infections, however, provided a first hint on how severe the HIV pandemic should become, e.g. sera from hemophiliacs in Japan were tested HIV negative early 1984, by end of 1984 already 20% of those patients were tested HIV positive after they had been treated with HIV-contaminated blood products from the United States (Gallo, 2002).

The spread of AIDS was rapid, soon after first reported in 1981 in the USA, fourteen countries reported cases of AIDS in 1982, and 33 already in 1983. By the end of 2008, 33.4 million people worldwide were estimated to be living with HIV. Figure 2.1 displays the prevalence of HIV in different regions of the world. According to the WHO 2009 report on the global AIDS epidemic, 2.7 [2.4-3.0] million people were newly infected with HIV and about 2.0 million [1.7-2.4] million people died of AIDS in 2008. HIV is believed to be responsible for approximately 25 million deaths, thus rendering it the most serious biological killer to date.

The prevalence of HIV infected individuals varies heavily between geographic regions, e.g. ranging from 0.1 to 0.5% in Central Europe to 15.0 to 28.0% in Southern Africa. These regional differences are likely the result of many factors, including: different availability of antiretroviral drugs, information campaigns (e.g. the government of South Africa pursued an "AIDS denialism" policy that caused deaths of about 330,000 people (Ano, 2006, 2008)),

the stigmatization of the HIV infection, and frequency of rapes (e.g. South Africa has an extremely high rate of rapes (Carries et al., 2007)). The high prevalence in sub-Saharan Africa probably results also from the fact that this region is the origin of HIV (Zhu et al., 1998). It is estimated that HIV entered the human population in 1931 [95% CI 1915-1941] through multiple infections from simian immunodeficiency virus (SIV)-infected nonhuman primates (Korber et al., 2000). Thus, approximately 50 years passed between the entry of the virus into the human population and the enrichment of AIDS cases in 1981. During these years, the virus could spread unrecognized, mainly because the late symptoms of AIDS coincide with symptoms of e.g. malnutrition and tuberculosis, which happen to be frequent problems in the infected population. As a consequence, the symptoms were attributed to rather well-known causes instead of a new infectious agent.

2.1.1 HIV Diversity

As briefly mentioned, HIV was introduced into the human population by multiple cross-species transmissions from SIV infected nonhuman primates. To date, we distinguish two types of HIVs: HIV type-1 (HIV-1), which is responsible for the current pandemic, evolved from an SIV variant present in chimpanzees, whereas HIV-2 is the result of a zoonotic infection from SIV in sooty mangabeys (Heeney et al., 2006). HIV-1 comprises four genetically distinct groups, namely M (main or major), N (non-M, non-O), O (outlier), and P, which were the results from at least four different zoonotic infections. Group P was only recently identified by Plantier et al. (2009) and originates from an SIV variant in gorillas. This work focuses on HIV-1 group M as it accounts for most infections worldwide and is therefore of premier interest.

HIV-1 group M is further divided into subtypes. This devision is based on clusters typically appearing in phylogenetic analyses of genetic sequences of HIV-1 group M (Robertson et al., 2000). These subtypes are named A to D, F to H, J, and K and have a different prevalence in different geographic regions of the world (Table 2.1), hence, suggesting that their population structure is the result of founder effects. However, it is possible that some phylogenetic clusters only appear because of incomplete sampling of the global viral population, e.g. newly sequenced HIV-1 strains from Central Africa fall in-between established subtype clusters (Rambaut et al., 2004).

In addition to the pure subtypes recombinant forms exist, as well. These are either well described established circulating recombinant forms (CRFs) or unique recombinant forms resulting from super-infection of a patient with two or more different subtypes.

Questions related to HIV-1 subtypes and its influence on disease progression (Kanki et al., 1999) and efficacy of antiretroviral treatment are still of major interest (Kantor et al., 2005). Subtype B is most prevalent in the developed, industrialized regions of the world, and therefore a representative of this clade was used for development of antiretroviral drugs. Thus, the genetic changes that distinguish B variants from non-B variants are believed to hamper the effectiveness of antiretroviral therapy (Descamps et al., 1998). Recently, it was shown that HIV infection-related complications may vary between different subtypes, e.g. the risk of developing HIV-induced dementia is increased in patients that are infected with subtype D compared to patients infected with subtype A (Sacktor et al., 2009).

HIV-1 subtype B is the best-studied variant of HIV-1 owing to the available resources

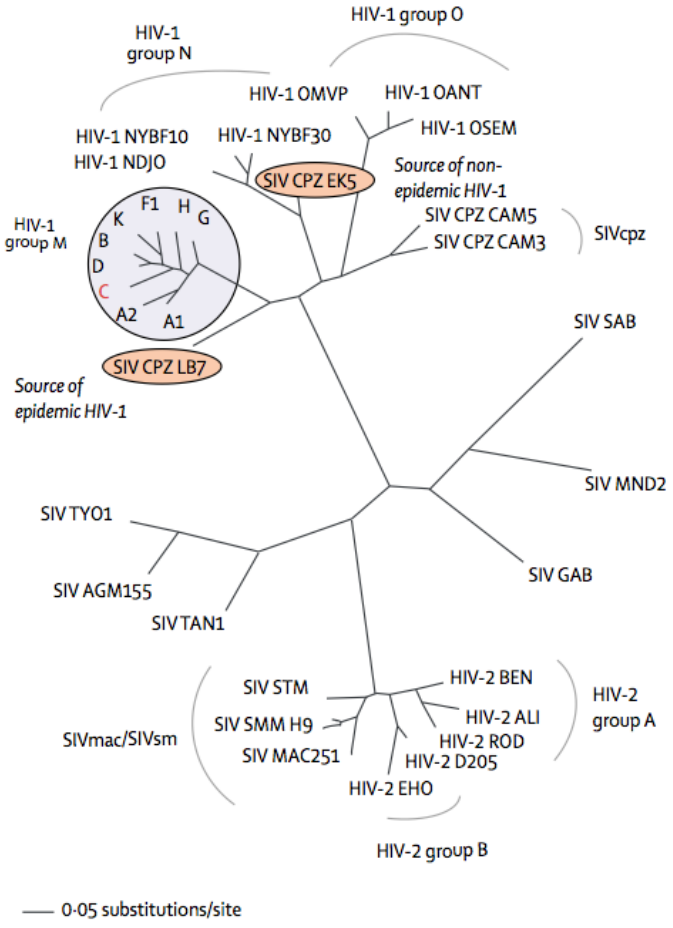


Figure 2.2: Phylogenetic relation of lentiviruses in men and non-human primates. Reprinted from The Lancet (Simon et al., 2006) with permission from Elsevier. The recently discovered HIV-1 group P (Plantier et al., 2009) is missing.

and infrastructure for research in its region of prevalence. The other HIV-1 subtypes clearly deserve more attention than they have received so far. However, data collection efforts in regions of their prevalence are less advanced than in Europe and North America. Also the data studied in this work originates mainly for collection efforts in North America and Central and Western Europe, hence, there exists a bias towards subtype B - simply due to practical limitations.

The diversity of HIV poses also one of the major challenges for HIV vaccine design (Walker and Burton, 2008). The sequence diversity within a single subtype, for example, can reach up to 20%. Clearly, the sequence diversity is an issue especially for antibody based vaccines that require conserved epitopes on surface proteins. For instance, in Africa with a multitude of different circulating viruses the sequence of the envelope protein can differ by up to 38% (Walker and Burton, 2008).

geographic region	number of infections	major risk groups	subtypes
Sub-Saharan Africa	22.400	HSex	A, D, F, G, C, H, J, K, CRF
South & South-East Asia	3.800	HSex, IDU	B, AE
Latin America	2.000	HSex, MSM, IDU	B, BF
Eastern Europe & central Asia	1.500	IDU	A, B, AB
North America	1.400	HSex, MSM, IDU	B
East Asia	0.850	HSex, MSM, IDU	B, C, BC
Western & Central Europe	0.850	MSM, IDU	B
North Africa & Middle East	0.310	HSex, IDU	B, C
Caribbean	0.240	HSex, MSM	B
Oceania	0.059	MSM	B

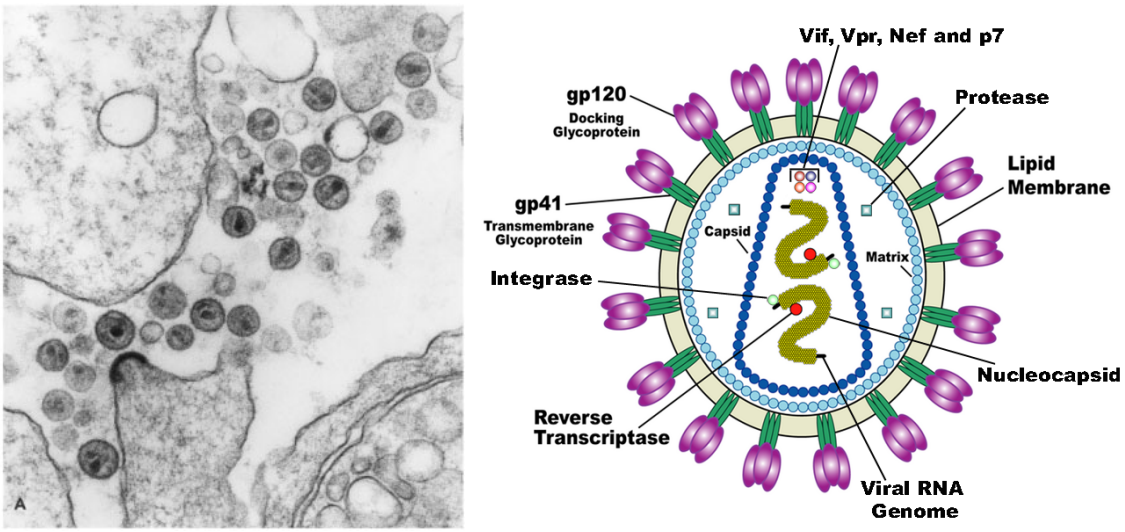
Table 2.1: Worldwide distribution of HIV-1 infections (as of December 2008), typical modes of transmission, and prevalent HIV-1 subtypes. HSex=heterosexual, MSM=Men who have sex with men, IDU=injection drug users. Numbers of infections are given in millions and originate from the WHO 2009 Report on the global AIDS epidemic. CRF=circulating recombinant form. Risk group and subtype distribution is based on (Simon et al., 2006).

2.2 The Human Immunodeficiency Virus Type 1

The human immunodeficiency virus (HIV) received its name officially in 1986. Extensive knowledge on HIV has been accumulated and a substantial amount of literature is available. In the following we provide a brief summary, for further details please refer to Fields et al. (2007) or similar reference literature.

HIV belongs of to the family of *retroviruses* and, more precisely, is a member of the genus of *lentiviruses*, which indicates a long incubation period. Electron microscopy of particles in infected cell cultures shows spherical entities with a diameter of 100 - 120 μm (Figure 2.3 a). A conceptual representation of the virus architecture is depicted in Figure 2.3 b). The characteristic of retroviruses is that they store their genetic information in ribonucleic acid (RNA) and thus require a mechanism to translate RNA to deoxyribonucleic acid (DNA), which is the carrier of genetic information in their hosts. Each viral particle contains two single stranded RNAs of approximately 10 kb length that are tightly bound to viral nucleocapsid proteins and two viral enzymes (reverse transcriptase and integrase) that are vital for a successful infection of the host cell. This complex is protected by a cone-shaped capsid comprising approximately 2,000 copies of the capsid protein. The viral cone – also known as the viral core – is clearly visible in the electron micrograph (Figure 2.3 a). The viral cone is surrounded by a spherical matrix that is in turn covered by a lipid membrane. Attached to the matrix, is the viral spike, which is responsible for target cell recognition and cell entry.

Figure 2.4 shows the organization of the HIV genome. HIV encodes a total of 15 viral proteins in overlapping reading frames. The majority of proteins is part of the three large precursor proteins that have to be cleaved into functional subunits. The polymerase gene



(a) Electron microscopy of viral particles

(b) Schematic representation of an HIV virion

Figure 2.3: Electron micrograph of a budding HIV particle and mature HIV particles with visible viral core from Fields et al. (2007) and reprinted with kind permission from Lipincott Williams & Wilkins (a). Schematic representation of a single HIV particle from Wikipedia (http://commons.wikimedia.org/wiki/File:HIV_Virion-en.png) (b).

(*Pol*) encodes for three proteins: the protease, the reverse transcriptase, and the integrase. The *Gag* gene is the precursor of four viral structural proteins: p24 (the viral capsid), p6 and p7 (the nucleocapsid proteins), and p17 (the viral matrix). The third precursor is encoded by the *Env* gene and comprises the two subunits of the viral spike, the glycoprotein gp120 and the transmembrane glycoprotein gp41. As indicated in Figure 2.3 b) the viral spike consists of 6 proteins: one trimer of gp41 and one trimer of gp120, which is heavily shielded by glycans. The smaller genes encode for transactivators (*Tat, Rev, Vpr*), which enhance gene expression, and other regulatory proteins (*Vif, Nef, Vpu*) helping the virus to be more efficient in its reproduction and counter defense mechanisms of the host cell. For instance, *Vif* disrupts the antiviral activity of the human enzyme APOBEC3G (Sheehy et al., 2002). APOBEC3G causes G-to-A hypermutations in the genome of HIV and other retroviruses, and thereby ultimately destroys the coding and replicative capacity of the invading pathogen.

2.2.1 HIV Replication Cycle

Figure 2.5 depicts the replication cycle of HIV: starting from cell entry to maturation of new infectious viral particles. The turnaround time is estimated with 1.5 days from entry to the production of new infectious virions. The life cycle of HIV is complex and our current understanding of each step is the result of extensive research. Nevertheless, many details are not fully understood, yet. Thus, we focus the description of the replication cycle on the targets for modern antiretroviral therapy. Further fascinating aspects of the interplay between viral regulatory proteins and the host as well as the interplay of the host's immune system with the viral infection are omitted. For a full description of the

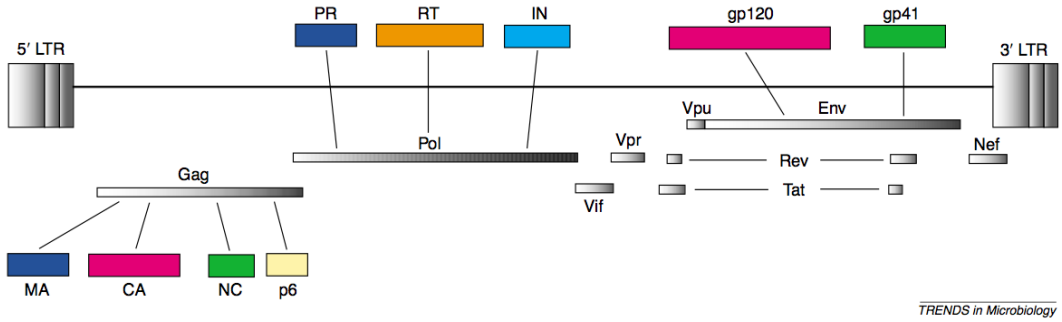


Figure 2.4: Diagram of the genome organization of HIV. Reprinted from Freed (2004) with permission from Elsevier.

molecular life cycle see e.g. Fields et al. (2007).

HIV enters the host cell by first binding one of its gp120 surface proteins to the CD4 receptor of the target cell. The importance of the CD4 receptor in HIV cell entry was identified shortly after the isolation of HIV (Dalgleish et al., 1984). The CD4 receptor is expressed on $CD4^+$ T cells, macrophages, microglia, and dendritic cells that thereby are the main targets of HIV. After the anchoring step, the gp120 subunit of the viral spike undergoes a conformational change with the consequence of exposing an epitope that, in turn, allows binding to a chemokine receptor (also called a coreceptor). The most important coreceptors used by HIV *in vivo* are the chemokine receptors CCR5 and CXCR4 (Berger et al., 1999). After coreceptor binding, another conformational change occurs in the gp41 subunit of the envelope protein. This change brings viral and host cell membranes in close proximity with the result of membrane fusion (Esté and Telenti, 2007). Recent results support the hypothesis that HIV primarily enters the target cell by endocytosis followed by fusion in the endosome and not by fusion directly at the plasma membrane (see Uchil and Mothes (2009) for a review). However, the successful fusion is followed by release of the viral core into the cytoplasm of the target cell. Right after the viral core is uncoated, the viral enzyme reverse transcriptase (RT) transcribes the viral RNA into a double stranded DNA (dsDNA). Together with viral and host proteins the dsDNA forms the preintegration complex (PIC), which is guided to the nuclear pore. Once the PIC has entered the nucleus of the host cell, the viral enzyme integrase, which was part of the PIC, catalyzes the integration of the viral DNA into the host chromosome. The integration of the viral DNA into the host chromosome marks a central point in the HIV infection, as from now on the cell is irreversibly infected.

The provirus, i.e. the integrated viral DNA, exploits the molecular machinery of the host cell for the production of new infectious viral particles. To be precise, the viral DNA serves as a template for the DNA-dependent RNA-polymerase II (*pol II*) that produces messenger RNA (mRNA). This mRNA is (partially) spliced and exported from the nucleus where it is translated into the different viral (poly-)proteins. In addition, full RNA transcripts of the complete integrated viral genome are generated and can be integrated into new virions.

The polyproteins *Env* and *Gag/Gag-Pol* are transported via different pathways to the viral membrane where they participate in forming new viral particles. The *Env* precursor protein, gp160, is glycolyzed in the endoplasmatic reticulum of the host cell, where the

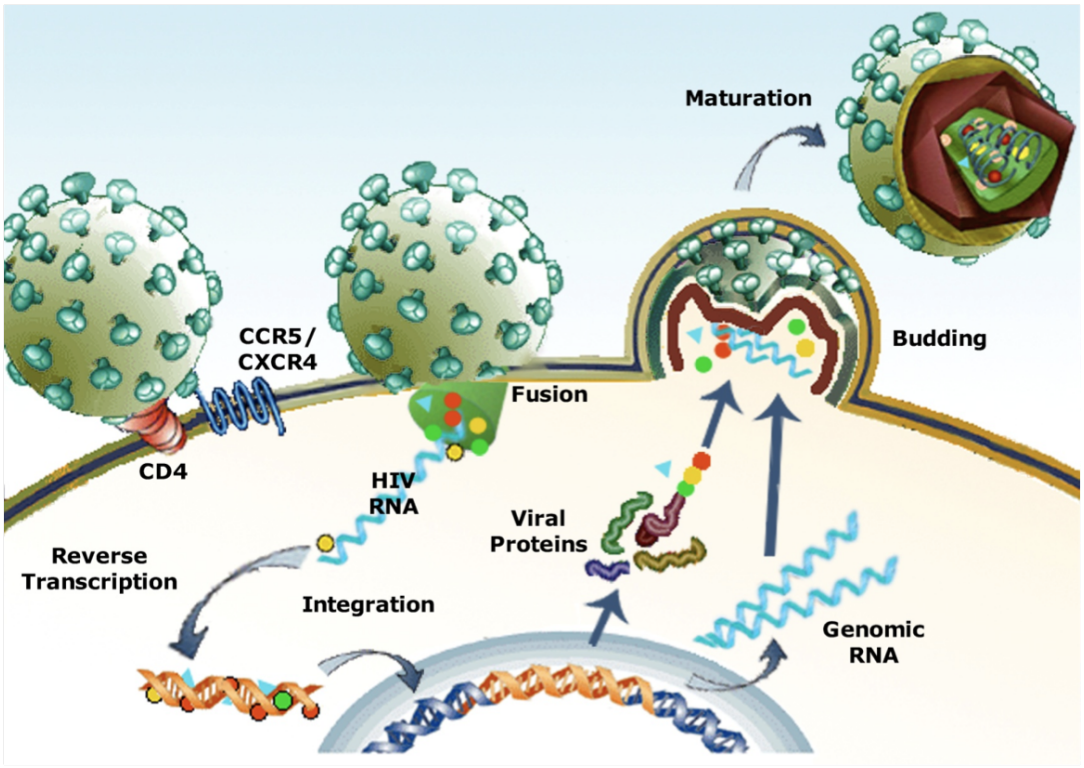


Figure 2.5: HIV replication cycle. The basic steps of the HIV replication cycle: viral entry, reverse transcription, integration, and formation of infectious particles. See main text for further details. Figure reprinted with kind permission of Saleta Sierra, Department of Virology, University Cologne.

monomers also undergo oligomerization. The predominant form is a trimer. The gp160 complexes are further transported to the cell's Golgi complex where they are cleaved into their subunits gp120 and gp41. There, cleavage is mediated by cellular enzymes and not by the viral protease. The heavy glycosylation of gp120 provides the means for HIV to shield its surface protein with a sugar coat against the immune response of the host.

The *Pol* polyprotein is only expressed as part of the larger *Gag-Pol* protein, which in turn is a result of a -1 translational frameshift (Jacks et al., 1988). The ratio of *Gag* to *Gag-Pol*, which is approximated to be 20:1 (20 copies of the *Gag* polyprotein for each *Gag-Pol* polyprotein), is crucial for the successful replication of HIV (Shehu-Xhilaga et al., 2001). The HIV protease (PR) is part of *Gag-Pol* and only active as a dimer. In fact, enzyme activation is initiated when the PR domains of two *Gag-Pol* precursors dimerize and the complex begins with intramolecular activity followed by intermolecular activity (Pettit et al., 2004). The shorter *Gag* protein is rapidly targeted to the inner surface of the cell membrane after its synthesis. At the membrane the precursor protein is then cleaved by the viral PR during or after budding from the host cell into its four subunits, which then form the viral matrix, capsid, and nucleocapsid, respectively. Only fully maturated viruses are able to successfully infect new host cells.

2.3 Clinical progression of an HIV infection

The main infection route of HIV is via sexual transmission across a mucosal surface. The actual risk of transmission varies and greatly depends on applied sexual practices. Male-to-male transmissions are one order of magnitude more likely than male-to-female and female-to-male transmissions. Further prominent routes of infection are mother-to-child transmission (via birth and/or breast feeding) and needle sharing among injection drug users. In the beginning of the HIV pandemic, also tainted blood transfusions were responsible for new infections.

The two main markers that are used for monitoring HIV infections are the viral load (copies of HIV RNA per milliliter (ml) blood plasma) and the $CD4^+$ T cell count per microliter (μl) blood. During the acute phase following primary infection, the viral load increases rapidly and reaches a peak of 10^6 to 10^7 copies/milliliter blood. HIV directly infects and causes the death of cells that are critical for effective immune response, and thus the $CD4^+$ T cell count decreases rapidly during this phase from over 1000 (as observed in healthy individuals) to around 500. Symptoms of acute HIV infection (for example fever, head ache, weight loss) appear about two weeks after exposure to the pathogen and are rather nonspecific and therefore, in general, are not distinguishable from infections with other viruses like influenza. Consequently, those symptoms are not used to diagnose an HIV infection. Moreover, the nonspecific symptoms result in an unrecognized infection, which in turn, leads to transmission of the virus to further individuals.

It is now known that, regardless of the route of transmission, HIV rapidly establishes a persistent infection of the gut-associated lympohid tissue, which is also the principal site of its replication. Arthos et al. (2008) showed that the HIV surface protein gp120 binds via one of its loops (V2) to the receptor integrin $\alpha_4\beta_7$, which in turn mediates the migration of leukocytes to the gut (von Andrian and Mackay, 2000). Furthermore, the gp120-$\alpha_4\beta_7$ complex facilitates the activation of lymphocyte function-associated antigen 1 (LFA-1) on $CD4^+$ T cells. LFA-1 is a central element for establishing virological synapses (Bromley et al., 2001) and consequently paves the way for HIV cell-to-cell transmission. Right after establishing the infection in the gut, the gut-associated lymphoid tissue undergoes a substantial depletion of $CD4^+$ T cells. It is estimated that half of the $CD4^+$ memory cells are killed during the initial attack (Mattapallil et al., 2005). Picker (2006) proposed that this depletion represents an irreversible damage to the immune system that ultimately results in AIDS. Another characteristic of the acute phase is the establishment of a latent viral reservoir. More precisely, HIV infects resting memory $CD4^+$ cells that, despite the integrated provirus, remain replication-competent. The provirus remains dormant in this reservoir, i.e. no viral proteins are actively transcribed, thus the infected cells are not especially targeted by the immune system. In this state, the provirus can survive for decades until eventually HIV replication is initiated and new viral particles are produced. Of note, the latency of the virus is regulated by epigenetic mechanisms (Kauder et al., 2009). Epigenetics refers to the change of phenotype or gene expression based on mechanisms other than changes in the genetic code. For instance, attachment of methyl groups to cytosines in the DNA stably alters gene expression profiles.

Towards the end of the acute phase, the viral load drops to a level of 10^3 to 10^4 and the immune system partially recovers from the primary attack as indicated by an increase

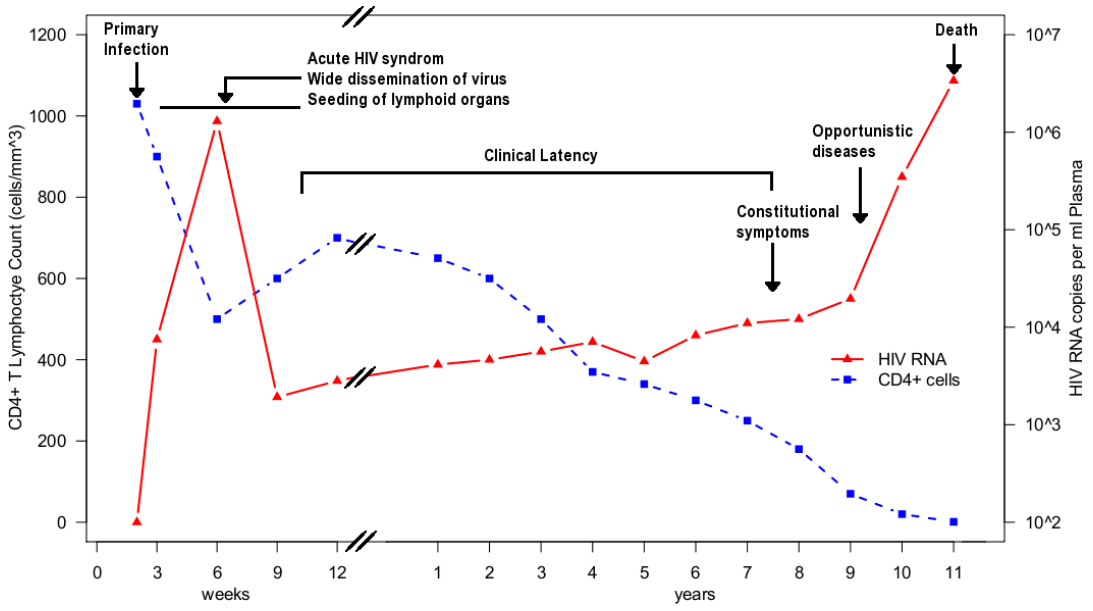


Figure 2.6: Typical progression of an untreated HIV infection based on Pantaleo et al. (1993).

in $CD4^+$ T cells. The acute phase is followed by a phase of clinical latency where the viral load slowly increases and the CD4 count continuously depletes. It is thought that HIV kills $CD4^+$ T cells directly mainly by inducing apoptosis or necrosis in infected cells. Nevertheless, it is believed that the continuous decrease of $CD4^+$ T cells is a result of the combination of the persistent immune hyper-activation (Hazenberg et al., 2003) followed by activation-induced cell death (Green et al., 2003) of T cells and the massive depletion of mucosal $CD4^+$ cells during the acute phase (Grossman et al., 2006).

At the point where the patient's immune system is too weak, opportunistic diseases and malignancies occur. Most HIV/AIDS-related complications are the result of bacterial or viral challenges to the already depleted immune system. The diseases range from cancers, which occur rarely among uninfected people, like Karposi's sarcoma as a result of an infection with human herpes virus type-8 to pneumonia induced by *Pneumocystis jirovecy*. The progression of an untreated HIV infection is depicted in Figure 2.6.

AIDS follows the phase of clinical latency and is defined by either reaching a CD4 count of less than 200 cells per μl or the manifestation of specified opportunistic infections. If untreated, AIDS is succeeded by death of the patient from these infections. Progression of the primary infection to AIDS varies among patients, ranging from only six months (Markowitz et al., 2005) to more than 25 years. One predictor of the rate of progression of the disease is the level of viral load established after one year without treatment (Lyles et al., 2000). Two groups of patients are able to control the replication rate of the virus such that the viral load does not exceed 5000 copies per ml (termed long-term non-progressors) or even 50 copies per ml (termed elite or natural controllers). These patients are recruited for genome-wide association studies that aim at uncovering the factors for their superior control of the infection (Fellay et al., 2007).

2.4 HIV Treatment and Drug Resistance

Since the discovery of HIV a large array of antiretroviral drugs has been developed, targeting several stages of the viral life cycle. The first drug zidovudine (abbreviated ZDV or AZT) was approved by the U.S. Food and Drug Administration (FDA) already in 1987 – only four years after the isolation of the virus. ZDV is a nucleoside reverse transcriptase inhibitor (NRTI) that targets the viral reverse transcriptase (RT), and thereby inhibits the process of generating DNA from the viral RNA. The synthetization of ZDV dates back to 1964.

The introduction of the first antiretroviral drug was a milestone in HIV therapy – this triumph, however, was only of short nature. Larder et al. (1989) and Larder and Kemp (1989) reported the emergence of resistant variants of HIV after prolonged treatment with ZDV. A major cause for the emergence of resistance is the viral RT itself, since the mechanism of reverse transcription is error prone and the enzyme lacks a proof-reading mechanism. Gao et al. (2004) estimated a mutation rate of HIV with 5.4×10^{-5} mutations per nucleotide per round of replication (initially Mansky and Temin (1995) reported 3.4×10^{-5} mutations per nucleotide per round, their result, however, was based on shorter fragments of viral RNA). Thus, the RT eventually generates a viral variant that is less susceptible to the drug and therefore is selected evolutionarily under treatment. By now, resistance mutations in all viral target proteins against all available compounds have been reported (Johnson et al., 2008).

2.4.1 Antiretroviral drugs

Since the introduction of zidovudine a multitude of anti-HIV drugs has been approved by the FDA and EMEA (European Medicines Agency) for treating HIV infections. Table 2.2 lists all currently FDA approved drugs. The targets of those drugs and their mode of action will be explained in the following paragraphs. Here we proceed in order of the steps of the viral replication cycle that they block.

Despite the large number of already existing anti-HIV drugs, multiple drugs in established and novel drug classes are under investigation. Table 2.3 lists an excerpt of currently investigated drugs.

Entry and Fusion Inhibitors intercept the viral replication at the first possible point: entry of the viral core into the cytosol of the host cell. Of all approved anti-HIV drugs, entry inhibitors are the only drugs that target a host protein rather than a viral protein. The development of entry inhibitors targeting human receptors originates from the observation that 4%-16% of the European population (higher rates in the north and lower rates in the south) have a homozygous $\Delta 32$ mutation in the CCR5 gene rendering the resulting CCR5 receptor nonfunctional (Novembre et al., 2005). However, the population affected by the genetic defect was also reported to be virtually immune against HIV infections, while on the other hand, no severe side effects resulting from the nonfunctional receptor are known (Novembre et al., 2005). In addition to maraviroc, which is a CCR5 inhibitor and currently the only approved entry inhibitor, more CCR5 and also CXCR4 inhibitors are under investigation (Esté and Telenti, 2007). Prior to the use of coreceptor blockers, it

Generic name	Abbreviation	Trade name	FDA approval
Nucleoside and nucleotide reverse transcriptase inhibitors (NRTIs)			
zidovudine	ZDV, AZT	Retrovir	1987
didanosine	ddI	Videx	1991
zalcitabine	ddC	Hivid	1992-2006
stavudine	d4T	Zerit	1994
lamivudine	3TC	Epivir	1995
abacavir	ABC	Ziagen	1998
tenofovir	TDF	Viread	2001
emtricitabine	FTC	Emtriva	2003
Non-nucleos(t)ide reverse transcriptase inhibitors (NNRTIs)			
nevirapine	NVP	Viramune	1996
delavirdine	DLV	Rescriptor	1997
efavirenz	EFV	Sustiva	1998
etravirine	ETR, ETV	Intelence	2008
Protease inhibitors (PIs)			
saquinavir	SQV	Fortovase, Invirase (SQV + RTV)	1995
ritonavir	RTV	Norvir	1996
indinavir	IDV	Crixivan	1996
nelfinavir	NFV	Viracept	1997
fos-/amprenavir	FPV/APV	Lexiva/Agenerase	2003/1999
lopinavir	LPV	Kaletra (LPV+RTV)	2000
atazanavir	ATV	Reyataz	2003
tipranavir	TPV	Aptivus	2005
darunavir	DRV	Prezista	2006
Fusion inhibitors (FIs)			
enfuvirtide	ENF, T-20	Fuzeon	2003
Entry inhibitors (EIs)			
maraviroc	MVC	Selzentry	2007
Integrase inhibitors (InIs)			
raltegravir	RAL	Isentress	2007

Table 2.2: Antiretroviral drugs approved by the FDA. Table is based on information from http://aidsinfo.nih.gov.

Generic name	Abbreviation	Phase	Comment
NRTIs			
apricitabine	ATC	III	
elvucitabine	ACH-126443	II	
racivir	RCV	II	
NNRTIs			
rilpivirine	TMC278	III	
lersivirine	UK-453061	II	
MIs			
bevirimat	BVM	II	
vivecon	MPC-9055	II	
EIs			
AMD11070	AMD070	II	CXCR4 antagonist, trial was halted
vicriviroc	VCV	III	CCR5 antagonist
ibalizumab	TNX-355	II	monoclonal antibody targeting CD4
	PRO 140	II	monoclonal antibody targeting CCR5
InIs			
elvitegravir	EVG	III	high cross resistance with RAL

Table 2.3: Some investigational antiretroviral drugs. Table is based on information from http://aidsinfo.nih.gov and http://www.aidsmeds.com.

is necessary to determine which coreceptor is used by the virus for entering (see Lengauer et al. (2007) for a review).

A further mechanism of preventing HIV from entering a target cell is to inhibit fusion of virus and host cell membranes. The currently only licensed fusion inhibitor (FI), Enfuvirtide (abbreviated ENF or T-20), is a synthetic peptide that mimics the amino acids 127-162 of gp41. ENF binds to a subunit of gp41 and therefore prevents the required conformational change that facilitates the fusion of host and viral membrane. Drug resistance mutations are usually located in the ENF-binding site on gp41 (direct resistance) or confer resistance indirectly via mutations in other regions of gp41 and even in gp120 (Miller and Hazuda, 2004). The latter mechanism is less well understood. The two major drawbacks of ENF are its price (approximately 2,200 Euros for one month of treatment compared to e.g. 400 Euros for one month of ZDV treatment in Germany; price estimates are based on retail prices at http://www.docmorris.de and the recommended daily dose at http://aidsinfo.nih.gov) and the way of administration: ENF has to be injected twice daily, and thereby causing a number of skin irritations.

Reverse Transcriptase Inhibitors interfere with the process of generating a DNA copy of the viral genome. The reverse transcriptase (RT) functions as a heterodimer comprising the p66 and p51 subunits. The p66 subunit is the full product of the RT region of the *Pol* gene, while the p51 subunit is obtained from the p66 unit by removal of a fragment at the C-terminus. This cleavage process is catalyzed by the viral protease. The p51 subunit plays only a structural role while the larger unit has two enzymatic centers: a DNA polymerase

and a ribonuclease H (RNaseH). The DNA polymerase copies either viral RNA or DNA, while the RNaseH is responsible for the degradation of tRNA primers and genomic RNA present in DNA-RNA hybrid intermediates. There are two classes of reverse transcriptase inhibitors that are distinguished by their mode of action.

The group of nucleoside/nucleotide reverse transcriptase inhibitors (NRTIs) are nucleoside and nucleotide analogs that are incorporated by the viral RT into the newly synthesized DNA strand. These analogs lack a free 3'-hydroxyl group and therefore terminate the transcription process after their incorporation into the DNA. The cost of NRTIs for one month of treatment varies from 300 Euros to 500 Euros (all prices apply to Germany only as they are based on information from http://www.docmorris.de). Three different mechanisms, associated with specific sets of mutations, used by HIV to lower the effect of NRTIs have been observed (see Sluis-Cremer et al. (2000) for a review). These mechanisms involve improved discrimination between NRTIs and their real dNTP counterpart, enhanced removal of the chain terminating NRTI at the 3' end, and alteration in RT-template primer interactions.

Non-nucleoside reverse transcriptase inhibitors (NNRTIs) form the second group of RT inhibitors. NNRTIs are small molecules that inhibit the RT by binding to a hydrophobic pocket in the proximity of the active site of the enzyme. Once the inhibitor is bound, it impairs the flexibility of the RT resulting in its inability to synthesize DNA. Resistance to NNRTIs occurs by mutations that reduce the affinity of the inhibitor to the protein. Usually, a single mutation selected by one NNRTI is sufficient to confer complete resistance to all compounds of the drug class (Clavel and Hance, 2004). This phenomenon is termed cross-resistance. Thus, newer inhibitors like etravirine (ETR) are designed such that they still exhibit high affinity in the presence of mutations induced by other NNRTIs. One month of NNRTI treatment costs about 450 Euros for older drugs and 650 Euros for recently approved products.

Figure 2.7 b) shows the structure of a functional heterodimer with typical NRTI and NNRTI resistance mutations highlighted in the p66 subunit.

Integrase Inhibitors aim at preventing the enzyme from integrating the viral DNA into the host chromosome. The integrase (IN) functions as a tetramer. Each monomer, which is cleaved out by the PR from the C-terminal portion of the *Gag-Pol* polyprotein, has three domains. The N-terminal domain contains a HH-CC zinc finger motif that is partially responsible for multimerization, optimal activity, and protein stability. The DDE motif in the core domain forms the catalytic triad. The C-terminal domain binds non-specifically to DNA with high affinity. So far it was not possible to crystallize the entire 288-amino-acid long protein due to its low solubility and propensity to aggregate. As an intermediate solution, the three domains have been crystallized individually. Moreover, structures of the N-terminus plus the core domain (Wang et al., 2001) and the core domain plus C-terminal domains (Chen et al., 2000) have been generated. The integration of the viral DNA requires three subsequent steps. During the 3' processing step the integrase removes a dinocleotide from the long terminal repeat of each HIV-DNA strand. This step is followed by a process termed strand transfer occurring in the nucleus where the integrase cuts the cellular DNA and covalently links the viral DNA 3' ends to the target DNA. The final step, the required gap repair, is believed to be carried out by host DNA repair enzymes (Yoder and Bushman,

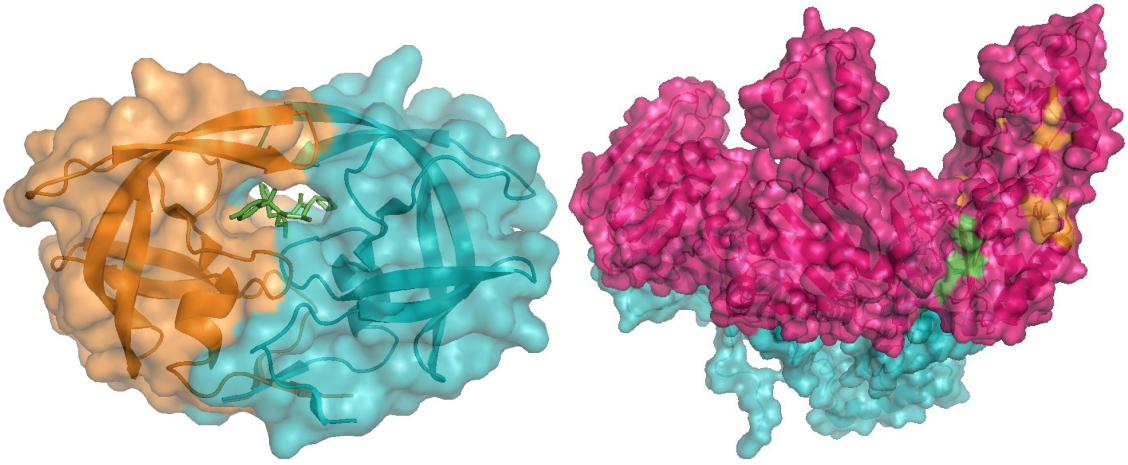
2000).

Currently only one FDA approved integrase inhibitor (INI) is available. Raltegravir (RAL) is a strand transfer inhibitor that interferes with the process by binding to the DDE motif in the catalytic domain (Hazuda et al., 2000). Successful inhibition of the integration process leaves the viral DNA in the nucleus, where it is recircularized by the host's repair enzymes. Hence, the HIV life cycle is interrupted. The mechanism of resistance of HIV against INIs is still subject to investigation. Some clinically relevant mutations are found to increase the rate of dissociation of the inhibitor from the IN-DNA complex (Grobler et al., 2009). The price for one month of raltegravir treatment can be assessed with 1,100 Euros.

Protease and Maturation Inhibitors interfere with the process of forming new infectious viral particles. The viral enzyme mainly engaged in virion maturation is the viral protease (PR). The PR cleaves the larger precursor proteins *Gag* and *Gag-Pol* into smaller functional units. Unlike its cellular relatives the viral PR acts as a true homodimer. Each monomer comprises 99 amino acids and the active site is formed by a cleft in between these symmetrically arranged monomers (Figure 2.7 a). The cleft accommodates the substrate to be cleaved and the two flexible flaps stabilize the substrate in the active site. The active site has room for a peptide of approximately seven amino acid length.

Because of its important role in virion maturation, the viral PR was soon subject of investigation for potential inhibitors. Protease inhibitors (PI) are small molecules that bind to the active site of the protease and therefore compete with its natural substrates. In fact, the design of protease inhibitors constitutes a good example of structure-based drug design, since the chemical structure of the inhibitors were chosen to mimic the structure of the viral peptides that are naturally recognized and cleaved by the protease. There are two mechanisms that result in resistance of HIV against PIs. The first one is the exchange of amino acids in the PR such that the affinity to the inhibitor is decreased while the natural substrates can be bound efficiently (Clavel and Hance, 2004). Modifications of the affinity to the natural substrate alter also the efficiency of the protease. Thus, the second mechanism introduces compensatory mutations aiming at reestablishing the efficiency of the enzyme while maintaining resistance against the inhibitor. These compensatory mutations can occur either in the protease or in its substrate, i.e. at cleavage sites (Nijhuis et al., 2007). As in the case of NNRTIs, the new generation of PIs shows antiviral activity in the presence of mutations selected earlier by other PIs. The expenses for one month of PI treatment range from 800 Euros to 1200 Euros.

Another class of drugs that interfere with the production of mature virions are maturation inhibitors (MIs). Unlike PIs they bind to the substrate of the protease instead of to the protease itself (Salzwedel et al., 2007). Currently no MI is approved by the FDA. The most advanced MI, bevirimat, prevents cleavage by the protease at the junction between the capsid and a spacer in the *Gag* precursor protein. Unfortunately, it was shown that a natural and frequently occurring polymorphism in *Gag* inactivates the antiviral activity of bevirimat (McCallister et al., 2008; Baelen et al., 2009), and therefore gives the inhibitor an unfavorable perspective (Verheyen et al., 2009).



(a) HIV protease (b) HIV reverse transcriptase

Figure 2.7: Structural models of viral targets for antiviral therapy rendered with the software PyMOL. (a) Protease with bound inhibitor *Amprenavir*. The two monomers of the functional homodimer are colored differently (PDB entry 3EKV). (b) Structural model of the functional RT dimer comprising the subunits p66 (mangenta) and p51 (cyan) from PDB entry 1RTJ. Typical NRTI resistance mutations are highlighted in orange, and NNRTI resistance mutations in green.

Future Developments. Novel drugs are under investigation in nearly all established drug classes described in the previous paragraphs. However, also novel targets or mechanisms are under ongoing investigation. The studied approaches range from new ways of attacking old targets, e.g. small peptides that prevent formation of an active RT dimer by blocking the binding of the p51 and p66 subunits (Agopian et al., 2009), aptamers targeting the RNaseH activity of the RT (Andreola et al., 2001), or integrase binding inhibitors with activity at the 3' processing step (Thibaut et al., 2009), to more exotic approaches. Sarkar et al. (2007) reported a successful excision of the integrated provirus from the host chromosome. A successful drug based on this mechanism would eventually allow to cure HIV infections, rather than (merely) delaying disease progression. A further possibility to eradicate the latent reservoirs of HIV is to trigger the transition from the latent phase to the active phase by targeting the epigenetic regulation (Kauder et al., 2009). Finally, targeting the frame shifting that maintains the ratio of *Gag* to *Gag-Pol* required for the generation of infectious virions is an option (Gareiss and Miller, 2009).

2.4.2 The Era of Highly Active Antiretroviral Therapy

The rapid resistance development of HIV against individual drugs required a new pharmaceutical strategy. Soon after the release of ZDV other nucleoside analogs were marketed and dual therapy comprising two NRTIs was the first attempt to control viral replication (Hammer et al., 1996). The approach of combining several antiretroviral compounds benefited the most from the release of drugs in other drug classes: NNRTIs and PIs. This marks also the start of the era of highly active antiretroviral therapy (HAART) in

1995. HAART combines a minimum of three drugs from at least two different drug classes (Clavel and Hance, 2004). A typical HAART combines two NRTIs plus either one PI or one NNRTI (Dybul et al., 2002). The rationale of HAART is to maximally suppress viral replication and thereby delay the progression of the infection to AIDS and death. The use of different drug targets provides the means of erecting a higher barrier for HIV to escape the regimen by developing resistance mutations. The success of HAART is based on the fact that HIV has to acquire multiple resistance mutations against the different drugs in the regimen. Here, the use of multiple drug classes ensures that different sets of resistance mutations are required for a successful escape. Thus, the combination of a successfully suppressed viral load, i.e. few viruses that can perform the "experiment", and a high genetic barrier to resistance, i.e. requirement of multiple mutations, substantially slows down the progression of the infection.

The success of HAART was manifest in the immediate decline in HIV related mortality following its introduction (Mocroft et al., 1998; Crum et al., 2006). Despite its success, HAART presents a burden to the patients as many pills have to be taken on a daily basis. Moreover, a strict adherence to the treatment schedule is required for optimal suppression of viral replication. In addition, every drug comes with a extensive list of side effects ranging from headache and diarrhoea to lipodystrophy (fat redistribution) and even peripheral neuropathy (nerve damage) and therefore can seriously impair the patient's quality of life. Given that currently there is no cure for HIV and antiretroviral treatment is a life long struggle, pharmaceutical companies are aiming at improving pharmacokinetics and the side effect profile of the drugs with the aim to render treatment more bearable. Nowadays, some antiretroviral combination therapies are available as a "once-daily" treatment delivered by a single pill. For example, the drug *Atripla* comprise three antiretroviral compounds (FTC+TDF+EFV) and is typically used for first-line treatment (one month of treatment costs approximate 1,300 Euros).

A further milestone in HAART was the discovery that the protease inhibitor ritonavir interferes with the liver enzyme cytochrome P450 (Kumar et al., 1996). This enzyme is involved in the metabolic processing of most protease inhibitors. Thus, the use of a small dose of ritonavir inhibits the liver enzyme, and helps to maintain optimal levels of other protease inhibitors in the patient's blood for a longer period of time. The *boosting* of protease inhibitors with ritonavir is standard as of 2001 – following the introduction of Kaletra (LPV+RTV) – and is usually denoted by PI/r. Currently, pharmaceutical companies are searching for further compounds achieving the same effect as ritonavir but with fewer side effects.

Despite of today's potency of HAART, drug resistance is still an issue. By acquiring drug resistance mutations the virus regains the ability to replicate in the presence of the drugs leading to a high viral load, which marks virological failure. Furthermore, virological failure usually precedes immunological failure marked by substantial decrease of $CD4^+$ T cells. A factor that contributes to resistance development in patients undergoing HAART is incomplete adherence (Harrigan et al., 2005; Glass et al., 2008). Patients not taking their drugs at all do not impose enough selective pressure on the virus. Conversely, if the drugs are taken correctly and optimal drug levels are always reached, then the virus has little chance to develop mutations. On the other hand, drugs taken at irregular intervals leave the opportunity of developing mutations that are subsequently selected due to the

selective pressure. Unfortunately, drug resistance may also appear in patients with perfect prescription refill rates. Harrigan et al. (2005) found at least one abnormally low drug plasma concentration in 36% of the patients with 95% prescription refill rate (i.e. almost perfect adherence to treatment). This points to either host genetic factors lowering the amount of available drug in the patient's body or to incomplete adherence during the daily schedule, e.g. all drugs for a day are taken in the morning. A major factor towards non-adherence are again the side effects of the compounds. Patients occasionally remove drugs from an effective HAART without consulting with their physician in order to achieve remedy for troubling side effects.

2.5 HIV Therapy Management

The circumstance that HIV evolves into drug resistance is further complicated by the fact that resistance mutations selected by one drug also confer resistance against other drugs from the same class. This phenomenon of *cross-resistance* is most pronounced among NNRTIs where, prior to the release of etravirine, a single mutation could render the complete drug class useless. The second generation of NNRTIs and PIs has therefore been designed in a way that they retain antiviral activity in the presence of commonly selected resistance mutations by older drugs.

In order to ensure an effective treatment the extent of drug resistance has to be appraised prior to onset of the regimen. Resistance of HIV against antiretroviral drugs can be assessed by two different approaches. The first approach, *phenotyping*, measures the resistance of the virus in an experimental assay (see e.g. Walter et al. (1999) and Petropoulos et al. (2000)). In such assays the replication of the patient's virus (clinical isolate) is compared to the replication of a reference wild type strain in the presence of a varying concentration of the drug. The concentration that cuts the replication rate in half is termed 50% inhibitory concentration (IC_{50}). The fold-change (FC) in resistance, also often referred to as resistance factor (RF), is then simply the quotient between the IC_{50} of the clinical isolate and the reference virus:

$$\mathrm{FC} = \frac{\mathrm{IC}_{50}(\text{clinical isolate})}{\mathrm{IC}_{50}(\text{reference isolate})}.$$

Figure 2.8 depicts an example of dose-response curves generated by a phenotypic resistance test. The experiment has to be carried out for every single drug, thus rendering the whole process time- and cost-intensive – results are available within weeks and costs are in the range of 1,000 Euros. Moreover, owing to the nature of the assay the test is restricted to few specialized laboratories with sufficient security level and expertise.

The second method focuses on the genetic sequence of the viral drug targets. In *genotyping* the viral sequence is obtained using standardized fast and cheap methods (result available within days at costs of 100 Euros) that can be carried out in virtually every laboratory. The outcome of *genotyping* is simply a list of mutations compared to a wild type virus and clearly much harder to interpret in terms of drug resistance as the single number per drug generated by *phenotyping*. In order to assist the interpretation of the viral genotype the international AIDS society (IAS) maintains an annually updated list of resistance mutations (Johnson et al., 2008). This list, however, mainly informs about the mere existence of mutations, the necessary knowledge about how many of the mutations

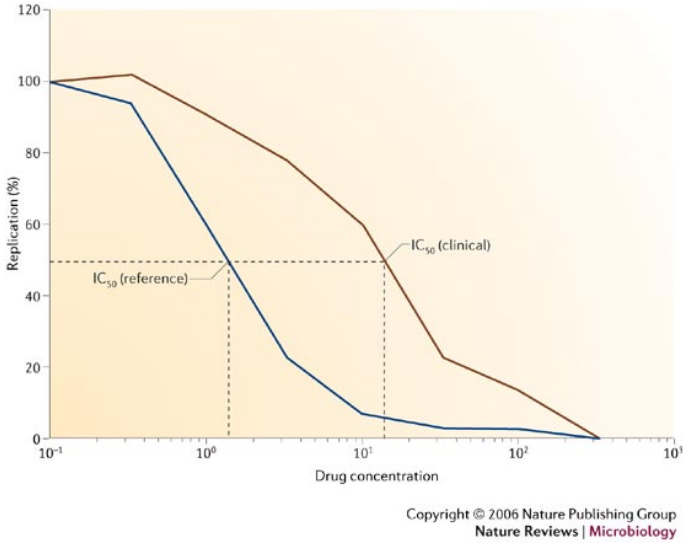


Figure 2.8: Dose-response curve of a phenotypic drug resistance test. Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Microbiology (Lengauer and Sing, 2006), copyright (2006).

and in which combination are required to confer intermediate or complete resistance is not provided. The missing semantics of the resistance mutations is introduced by rules-based interpretation systems developed by virologists and clinicians (see Lengauer and Sing (2006) for a review).

The benefit of resistance tests (genotypic or phenotypic) has been demonstrated in prospective clinical trials. For example, Durant et al. (1999) showed that treatment decisions based on a genotype have a significantly better outcome than uninformed readjustments of medication. Moreover, Oette et al. (2006) showed that genotypic resistance analysis is beneficial even in treatment-naïve patients, i.e. patients that have never received any antiretroviral drug. The observed improvement is a direct consequence of transmitted drug resistance mutations, where patients are infected with already resistant viral strains. The advantage of resistance testing and the broad availability of genotyping led to routine sequencing of relevant parts of the viral genome. The most relevant part comprises the 99 amino acids of the viral protease and up to the first 240 amino acids of the RT. These data are now generated routinely in laboratories for genotype based resistance tests and are stored together with frequently obtained markers of the treatment success (viral load and $CD4^+$ T cell count) and detailed information about the patient's regimen (drugs, start, stop, cause of stop) in databases (Rhee et al., 2003; Roomp et al., 2006). These data collections provide a rich basis for further developments in terms of resistance interpretations as we will show in the remainder to this thesis.

2.5.1 Rules-Based Interpretation Systems

The need for adding semantics to the resistance mutations promoted the development of many rules-based interpretation systems. The rules are generated by expert panels comprising virologists and clinicians on the basis of genotype-phenotype relationships, publica-

tions, clinical response (in the form of mutations observed in patients failing antiretroviral therapy), and personal experience. The predictions by different rules-based interpretations are often discordant owing to their history of origins, which is manifest in the differing sets of rules and design principles. Frequently the discordant predictions lead to confusion by users consulting multiple decision support systems (De Luca et al., 2003). Moreover, the degree of disagreement can also depend on the HIV-1 subtype (Snoeck et al., 2006) thereby reflecting the general lack of knowledge on non-B subtypes. The rule sets are maintained regularly in meetings of the expert boards, and the rules have to be updated because of the ongoing release of new antiretroviral drugs. Rule sets for novel compounds have to be added as well as the impact of mutations selected by new compounds on the established drugs has to be studied. Thus, over time the rules undergo refinement that reflects the state-of-the-art knowledge on HIV drug resistance.

Nowadays, antiretroviral therapy comprises multiple drugs. The decision support systems, however, provide only classifications for individual drugs. In order to overcome this limitation, the verbal classification provided by the tools is often mapped to a rating between 0 and 1 assessing the susceptibility of the virus against an individual drug. The ratings of individual drugs are summed to form a genotypic susceptibility score (GSS) that usually ranges between 0 and the number of drugs in the regimen (De Luca et al., 2003). For treatment-naïve patients clinicians usually aim at achieving a GSS of 3.0 (i.e. three fully active drugs) while for treatment experienced patients with heavily mutated viruses a GSS of 2.0 is recommended.

The rules-based systems enjoy great popularity among clinicians and virologists involved in HIV patient care, mainly owing to the fact that they are not providing "black box" predictions. More precisely, the basis of the classification, i.e. the rule sets, are freely available (for most systems) and thereby allow users to follow the rationale behind the decision. In the following we provide a brief overview about the most popular rules-based decision support systems.

ANRS 07/2009 V18 The interpretation system provides a three-level classification of drug resistance and is developed by the French ANRS (National Agency for AIDS Research) AC 11 Resistance group. Their set of rules is now available in version 18 on the web site: http://www.hivfrenchresistance.org/ (Meynard et al., 2002).

Rega Version 8.01 Is a rule set maintained by the Rega Institute of the Catholic University in Leuven. It also provides a three-class rating. In contrast to the ANRS system it applies a weighted score for the conversion from the verbal classification (i.e. susceptible, intermediate, and resistant) into the scores that are then used to compute the GSS. Precisely, the score for boosted PIs is 0.75 and 1.5 instead of 0.5 and 1 for intermediate and fully susceptible viruses, respectively, and the score corresponding to intermediate resistance of INIs, EIs and some NNRTIs is 0.25 instead of 0.5. The specifications are available at http://www.rega.kuleuven.be/cev/ (Van Laethem et al., 2002).

HIVdb Version 6.0.5 The HIVdb system is maintained by the Stanford university and available at http://hivdb.stanford.edu. Unlike in the other systems, each mutation receives a score between -10 and 60 and therefore the systems resembles rather a linear

model with expert-derived weights than a rules-based system. Negative scores indicate resensitization effects, i.e. a virus with such mutations is more susceptible against that particular drug in the presence of the mutation. Based on the weights the resistance against each drug is assigned to one of five possible classes (Rhee et al., 2003). Drugs scored with 0-9 are estimated to be "Susceptible", 10-14 indicates "Potential low-level resistance", while 15-29 and 30-59 correspond to "Low-level resistance" and "Intermediate resistance", respectively. Values of 60 or greater indicate "High-level resistance". Moreover, the web service provides implementations of other rule sets like ANRS and Rega and additionally offers the option to use a personalized rule set.

HIV-grade Version 12-2008 HIV-grade is an interpretation system maintained by HIV-grade e.V. an association of German clinicians and virologists. Like HIVdb it provides the possibility to compare different systems (all those mentioned above including GENO2PHENO described in Section 2.5.2). HIV-grade differs from the other systems by providing special rules for some drugs. These special rules consider which other drugs are part of the combination. In particular, the additional rules are aiming to capture mutations with resensitization effect and require that the drug selecting for the resensitization mutation is maintained in the regimen for keeping up the selective pressure. The susceptibility for each drug is provided in a three-class system. The web service is available at http://hiv-grade.de.

AntiRetroScan version 2.0 The AntiRetroScan algorithm is maintained by the Italian Antiretroviral Resistance Cohort Analysis (ARCA) consortium (Zazzi et al., 2009). Like HIVdb it provides five resistance classes and like the Rega rules it applies different weights for different drug classes that should be used to compute a GSS. The rule set is available at http://www.hivarca.net/includeGenpub/AntiRetroScan.htm.

Proprietary Systems In addition to the free decision support systems introduced in the previous paragraphs, there also exists a number of proprietary rule sets that are part of diagnostic kits. Those kits usually include a sequencing machine and primers needed for the sequencing. Moreover, the interpretation software has to be approved by the FDA. An example of a proprietary system is the ViroSeq (Version 2.8) rule set, which is developed by Celera Diagnostics and provides three output categories (http://www.celeradiagnostics.com/cdx/ViroSeq). The tool accompanies the ViroSeq HIV-1 Genotyping System developed by Abbott. Likewise, the GuideLines Rules (Version 14) are updated annually by an independent expert panel and accompany the TRUGENE HIV-1 Genotyping Assay offered by Siemens (http://www.medical.siemens.com/).

2.5.2 Predicting the Drug Resistance Phenotype from Genotype

Currently, there exists a multitude of expert-based decision support systems. The systems frequently disagree on the classification results, which is no surprise given the different design principles and experts involved in the numerous boards. However, there has been the desire to base the classification of drug resistance on a more objective foundation. Obviously, an objective criterion of drug resistance is the phenotypic drug resistance as measured

in vitro by experimental assays. There are two frequently used tools that predict phenotypic drug resistance from the viral genotype. One system, VircoTYPE, is proprietary and maintained by the company Virco BVBA (Mechelen, Belgium), the other, GENO2PHENO, is a freely available web service offered via http://www.geno2pheno.org (Beerenwinkel et al., 2002, 2003a). Both systems use statistical learning to infer drug resistance models from matched genotype-phenotype pairs. Application of these models allows to infer the more expensive and objective information on the basis of the cheap and easily available genotype.

The first version of GENO2PHENO used decision trees (Breiman et al., 1984) to classify the virus as resistant or susceptible with respect to several drugs. The training data for the initial models comprised approximately 450 genotype-phenotype pairs (Beerenwinkel et al., 2002). The decision trees achieved moderate prediction performance but demonstrated in an appealing way, how rules of drug resistance can be automatically derived from data. Precisely, decision trees are a white box classifier, i.e. in addition to the classification result the user can learn what input led to the final decision, as opposed to black box models like neural networks and support vector machines (SVMs) with non-linear kernels (Boser et al., 1992). In a later version of GENO2PHENO a SVM-based prediction was added to the web service and the models were trained on approximately 650 genotype-phenotype pairs (Beerenwinkel et al., 2003a). The SVM exhibited better prediction performance than the decision trees and allowed to solve a regression problem, i.e. predicting the FC in resistance rather than only a class label. Because linear SVMs cannot operate on categorical data, e.g. amino acid sequences, the viral genotype was represented as binary vector, with one indicator for every possible amino acid at each sequence position. Hence, the encoding required 20 times the length of the amino acid sequence. The model interpretation of SVMs, however, is not as trivial as in the case of decision trees. Based on the work by Sing et al. (2005b) the linear SVM models for the individual drugs could be interpreted and a list of scored mutations (main contributers to observed resistance) is now provided with every prediction (for details see Section 3.1). The current version of GENO2PHENO is trained on approximately 1,000 genotype-phenotype pairs per drug.

VircoTYPE uses also a linear model but, unlike GENO2PHENO, uses linear regression with pairwise interaction terms for mutations. The statistical learning task was carried out on a median of 46,100 genotype-phenotype pairs per drug (Vermeiren et al., 2007). Again, the use of a linear model facilitates model interpretation. In contrast to GENO2PHENO, VircoTYPE is a commercial tool that charges the user for every prediction.

2.6 Further Advancements

Since the first methods that predict *in vitro* drug resistance have been published, this task was subject of numerous publications. The wealth of publications was also boosted by a freely available genotype-phenotype dataset published by Stanford's HIVdb (Rhee et al., 2006). The publications range from use of established statistical learning models like simple linear models (Wang et al., 2004), linear models including interaction terms (Vermeiren et al., 2007), and rules-learning algorithms (Kierczak et al., 2009) to the development of novel techniques like non-parametric methods (DiRienzo et al., 2003) and machine learning approaches aiming at the discovery of patterns of resistance mutations (Saigo et al., 2007).

Other approaches extended the genotype-phenotype relationship by incorporating descriptors of the antiretroviral drug in the prediction model (Drăghici and Potter, 2003; Lapins et al., 2008). Additionally, methods from docking and molecular dynamics simulation were used to compute the binding energy of PIs to variants of the protease. The binding energy turned out to be a good predictor of resistance without requiring a training step nor training data (Jenwitheesuk and Samudrala, 2005). These systems, however, represent a proof-of-concept and are not used as clinical decision support tools.

Despite all the approaches for predicting *in vitro* drug resistance, the overall aim is still to infer clinical *in vivo* response of the patient to the drugs. Thus, the estimation of relevant cutoffs is an essential part for establishing a useful prediction system. The expert derived rules-based systems, on the other hand, mainly aim at interpreting resistance mutations in terms of clinical response and consequently have no need for relevant cutoffs. Nonetheless, all approaches have in common that they only classify single drugs as resistant or susceptible, but what lies at the heart of the problem is the prediction of response to a combination of drugs. In fact, HIV experts have discouraged inferring response from genotype via the intermediate step of first predicting the resistance phenotype (Larder et al., 2007; Brun-Vézinet et al., 2004).

In contrast to this criticism, we could show that predicted phenotypes – using the VircoTYPE system – combined with clinically relevant cutoffs are at least as predictive as several expert derived rule sets. Additionally, predictive accuracy could be further enhanced for all systems by weighting individual drugs using statistical learning methods instead of the simple summation usually applied to derive a GSS or PSS (Altmann et al., 2009b). Moreover, statistical learning allowed to include predictions by multiple interpretation systems to infer virological response. In fact, combinations of predicted phenotypes with expert predictions showed a slight improvement, which could also be confirmed on a large independent test set. In contrast, combinations of only rules-based predictions showed no added value over the use of a single system. In this study we also evaluated the importance of predictions for single drugs in terms of response to antiretroviral treatment. Indeed, the benefit of statistical learning originated from the different weighting of drugs and confirms earlier results (Swanstrom et al., 2004). This weighting is presumably the result of the abundance of the drug in the training data, its potency, the accuracy of the prediction algorithm for that drug, pharmacokinetics (e.g. the *forgiveness* of a drug when a dosage was missed or taken too late), and further factors adding to the success of the drug *in vivo*.

Given a sufficient amount of data, it is possible to develop models that directly predict response to antiretroviral combination using genetic information and the drugs in the treatment. Such models thereby circumvent the need to first assess resistance against single compounds and then subsequently combine the single predictions to a score for the complete regimen. A further advantage is that such direct models can learn negative and positive pharmacokinetic interactions between drugs from the data. Indeed, there exists a large number of interactions between the different drugs (see Boffito et al. (2005) for a review), ranging over different drug classes and even to non-anti-HIV drugs (e.g. all drugs metabolized by cytochrome P450 are affected by ritonavir boosted PIs). The worst interactions between drugs in terms of drug resistance are effects that lower the bioavailability of one or more compounds of the combination treatment: frequent suboptimal levels of

a drug in the blood plasma provide the right mixture of selective pressure and extent of viral replication to eventually generate resistant viruses. An online repository providing lectures and overviews on pharmacokinetic interactions is maintained by the Liverpool HIV Pharmacology Group (http://www.hiv-druginteractions.org/).

Plasma levels of drugs can also be influenced by the genetic predisposition of the host. For example, Mahungu et al. (2009) showed that a mutation in cytochrome P450 has a positive influence on the plasma level of the NNRTI nevirapine. Hence, consideration of host genetics during the selection of compounds for a combination therapy can yield further improvements. On the other hand, severe adverse effects can be increased depending on host genetics, e.g., the patient immunotype HLA-B*5701 is strongly associated with hypersensitivity to the NRTI abacavir (Mallal et al., 2008). Still, despite its evident benefit, pharmacogenetic testing is currently not widely applied.

A further host-specific aspect not considered by current tools is the presence of co-infections with other viruses (e.g. HBV) or other (chronic) diseases that require medication. Again, compounds used to treat other diseases may limit the efficacy of anti-HIV drugs or might add limitations in the selection of anti-HIV drugs.

Further important information is neglected by the restriction of the resistance analysis to the baseline genotype alone: archived resistance mutations in the viral reservoirs (latent viruses). Resistance mutations selected by earlier treatments can disappear from the viral population as soon as the selective pressure is removed. For instance, protease resistance mutations disappear in the absence of protease inhibitors because the protease is more efficient without those resistance mutations. Increased efficiency translates into a higher number of infectious viruses that eventually outnumber the protease resistant variants. The mutations, however, persist in viral reservoirs and are rapidly reselected after reappearance of the selective pressure by reusing the drug. Consequently, just studying the most recent genotype might miss important mutations. And indeed, the patient's treatment history has long been recognized as valuable information (Bratt et al., 1998), and recently it has been shown that inspection of previous genotypes helps to explain response to antiretroviral combination therapy (Zaccarelli et al., 2009).

A quantitative analysis of the viral population with new sequencing techniques revealed advantages in predicting the coreceptor usage phenotype from genotype (Däumer et al., 2008). In terms of drug resistance, the interest is mainly on resistance mutations that are only present in a minority of the viral population at the onset of the treatment. Minorities might increase the risk of selecting that resistance mutation early after treatment start. Hence, the question is, if minorities are harmful for the success of treatment at all, and if so, what threshold of the minority leads to rapid selection of resistant variants. Using standard Sanger sequencing (also called bulk-sequencing) a minority in the population can only be detected, if it accounts at least for 20% in the overall sample. In contrast, ultra-deep sequencing allows to determine the sequence of individual viruses. Hence, here the resolution is mainly bound by the cost of the experiment. Initial studies using ultra-deep sequencing demonstrated that resistance mutations found in minorities of the viral population correlate well with previous antiretroviral therapies and can therefore provide information on archived resistance mutations even in the absence of treatment records (Le et al., 2009).

Eventually, every known drug combination fails due to acquired drug resistance muta-

tions. The fact that anti-HIV therapy is a life-long effort paired with the limited number of antiretroviral drugs and the cross-resistance between them, raises the need to consider different strategies of effectively using the available array of compounds. For example, one could follow a "hit-hard" strategy and apply one drug from each class, which will probably lead to a very long lasting successful treatment (e.g. six years), but once it fails there are resistance mutations against drugs from all classes. Consequently, only a very limited number of drugs is available after treatment failure. On the other hand, one can restrict the treatment to two drug classes, like in standard HAART, which remains effective for e.g. three years. The fact that other drug classes have been spared, allows for multiple repetition of the medium term success. More specifically, one could optimize the order in which drugs of one class should be administered for minimizing the impact of cross-resistance. This planning of sequences of treatments, or therapy sequencing, aims at prolonging the patient's life.

Summing up, improvement over state-of-the art treatment guidance can be achieved by:

1. Construction of models that predict *in vivo* response to antiretroviral combination therapy and thereby capture interactions between drugs and between drugs and mutations *in vivo*.

2. Inclusion of features of viral evolution for assessing the risk of the virus to escape a putative treatment by developing further mutations.

3. Incorporation of data on the patient's treatment history for assessing the risk of resistance mutations archived in viral reservoirs.

4. Consideration of host specific characteristics ranging from genetic factors to co-infections for addressing the interactions between drugs and host (pharmacogenetics) and virus and host (e.g. immune control)

5. Investigation of the impact of minorities in the viral population on the treatment outcome and resistance development using novel sequencing techniques.

6. Generation of personalized treatment schedules that ensure the optimal use of the available array of antiretroviral drugs.

This thesis addresses some of these issues. Dealing with points four and five, for example, requires novel data: presence of co-infections or genotyping of the patient and sequencing the viral quasi species, respectively. These data are not routinely collected in treatment databases, and the corresponding issues are therefore not addressed. The objective of the thesis is the inference of virological response *in vivo* directly from the viral genotype, the applied drug combination, features derived from the viral genotype, and further available information (Chapters 4 and 5). The models are restricted to RTIs and PIs (excluding next generation drugs within these classes), simply due to lack of data on novel drugs in databases collecting clinical response data. This poses no major limitation, as RTIs and PIs remain the major building block for today's antiretroviral therapy of HIV infections. Novel drugs are typically spared for later treatment lines often termed salvage treatments. Finally, we undertake first steps towards therapy sequencing by introducing a fast method for estimating the viral evolution during combination therapy (Chapter 6).

3 Extensions to GENO2PHENO

The web service GENO2PHENO for predicting phenotypic drug resistance from the viral genotype has been in operation since December 2000. Updates and continuous research is required for maintaining its benefit for virologists and clinicians treating HIV patients. In this chapter we describe the web tool GENO2PHENO$_{[\text{integrase}]}$ that predicts from the viral genotype phenotypic resistance against two integrase inhibitors (Section 3.1). Moreover, we present research aiming at improving the usefulness of predicted fold-change in IC_{50} by deriving clinically relevant cutoffs of the continuous measure (Section 3.2). Finally, in Section 3.3 we explore approaches for improving prediction of resistance to individual antiretroviral drugs by using abundantly available viral genotype data in addition to the genotype-phenotype pairs.

3.1 GENO2PHENO$_{[\text{integrase}]}$

The viral integrase comprises 288 amino acids and is a product of the C-terminal portion of the large precursor protein encoded by the *Pol* gene. Only recently, in 2008, the first integrase inhibitor raltegravir (RAL) was approved by the FDA. Knowledge of drug resistance against RAL and elvitegravir (EVG), an integrase inhibitor in phase III clinical trials, accumulates only slowly. Moreover, due the recent release of the drug, routine resistance testing is not established yet, mainly owing to the fact that the integrase of the patient's virus is assumed to be wild type and thus susceptible to the novel drug. As a consequence, genotype-phenotype data is rare. However, some researchers investigate mutations, which were reported to emerge during integrase inhibitor containing regimens, with phenotypic resistance assays. These genotype-phenotype data are partially available via Stanford's HIVdb[1].

Material and Methods

Using this freely available resource we compiled a dataset comprising 113 and 126 genotype-phenotype pairs for RAL and EVG, respectively. The raw data were preprocessed exactly like in the case of GENO2PHENO: the decadic logarithm was applied to the fold-change values, and each of the 288 amino acid positions was represented by 20 binary variables indicating the presence (or absence) of a specific amino acid at that position. In accordance to the GENO2PHENO web service we applied linear support vector regression (SVR) for computing the continuous $\log_{10}(\text{FC})$ from genotype. The cost parameter C of the SVR was optimized in a five-fold cross-validation. Performance was measured using Pearson's correlation coefficient (r) between predicted and measured $\log_{10}(\text{FC})$ in a leave-one-out cross-validation (LOOCV) with the optimized cost parameter. In addition, we trained one

[1] http://hivdb.stanford.edu/cgi-bin/IN_Phenotype.cgi

linear SVR model for each drug on the complete set for investigating the contribution of each mutation to the overall observed resistance. More precisely, linear support vector machines (SVMs) allow to rewrite the decision function for classification and regression as a simple linear model with trivial access to the contribution of every mutation (Guyon et al., 2002). The training of a SVM results in a set of support vectors $\vec{x_k}$ with label y_k and a set of non-zero weights a_k. In case of a classification task, the label for a new sample $\vec{x}$ is then computed by:

$$f(\vec{x}) = \text{sgn}\left(\sum_k y_k \alpha_k K(\vec{x_k}, \vec{x})\right), \tag{3.1}$$

where $K(\vec{x_k}, \vec{x})$ is the kernel function. Essentially, a kernel function expresses the similarity between two vectors. Hence, during the decision process the similarity of the input vector to each support vector $\vec{x_k}$ is computed and multiplied with its label y_k and its influence α_k derived during training. Popular kernel functions are for instance, the polynomial kernel with degree d, $K(\vec{x_k}, \vec{x}) = (1 + \langle \vec{x_k}, \vec{x} \rangle)^d$ and the radial basis function kernel $K(\vec{x_k}, \vec{x}) = \exp(-\beta \|\vec{x} - \vec{x_k}\|^2/(2\sigma^2))$. However, when using a plain linear kernel we simply set

$$K(\vec{x_k}, \vec{x}) = \langle \vec{x_k}, \vec{x} \rangle,$$

which allows us to rewrite the decision function as

$$f(\vec{x}) = \text{sgn}\left(\vec{w} \cdot \vec{x} + b\right), \text{ with}$$

$$\vec{w} = \sum_k \alpha_k y_k \vec{x_k} \text{ and } b = \langle y_k - \vec{w} \cdot \vec{x_k} \rangle.$$

Thus, the weight vector $\vec{w}$ is simply a linear combination of the support vectors and offers a straightforward interpretation.

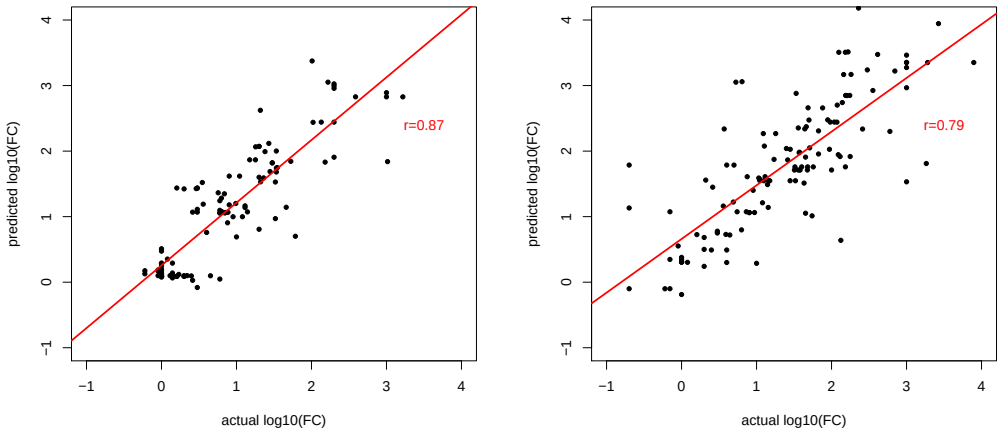
Results

Figure 3.1 depicts the scatter plots between predicted and observed $\log_{10}$(FC) values. For both drugs the correlation coefficients are high, whereas resistance to RAL ($r = 0.87$) was predicted more accurately than resistance to EVG ($r = 0.79$).

The influence of each mutation on the measured phenotypic resistance against RAL and EVG is displayed in Figures 3.2 a) and b), respectively. As expected, most mutations lead to an increase in FC; a few mutations, however, slightly increase susceptibility to the drugs. Figure 3.2 c) shows a scatter plot between the SVR weights obtained from the RAL and EVG models. Here it becomes evident that most mutations confer resistance to both drugs. A few exceptions are mutations at position 66, which influence the EVG phenotype but not the RAL phenotype, and mutation 151I that increases RAL resistance, but decreases EVG resistance. The offset (i.e. b in the linear model) is close to 0 for both models, thus indicating no resistance for wild type viruses.

Discussion

A clear limitation of this study is the low number of training instances, which, in addition, originate from a total of eleven different research groups using different experimental assays for measuring phenotypic resistance. These factors clearly increased the variation



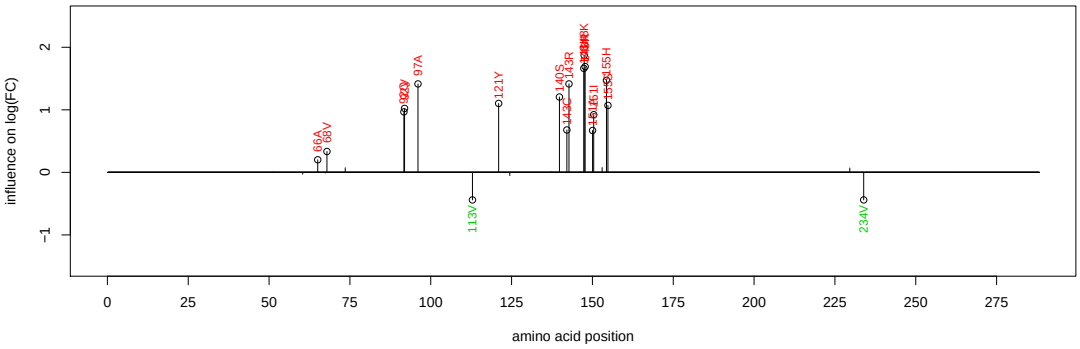
(a) RAL (b) EVG

Figure 3.1: Scatter plots between predicted and observed $\log_{10}$(FC) values for (a) RAL (n=113) and (b) EVG (n=126).

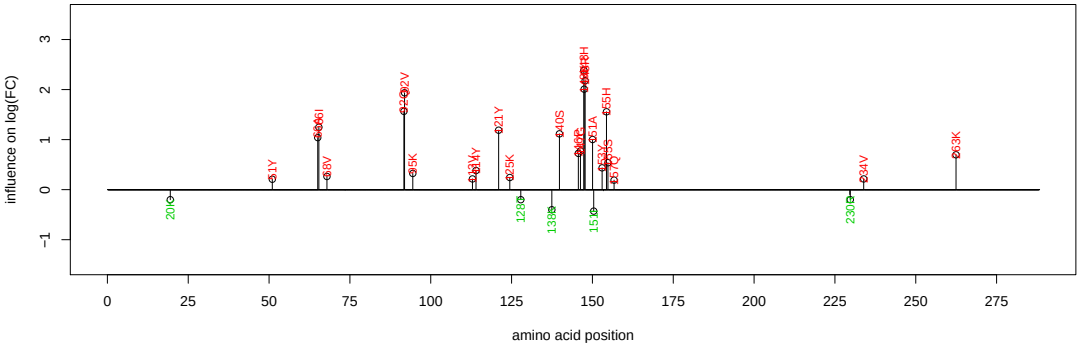
in the phenotypic measurements we used for training the SVR models. Despite this shortcoming, the achieved prediction accuracy was surprisingly good for both drugs. This, however, might be a direct consequence of the origin of the data: researchers were interested in the effect of a few mutations, which frequently emerge during treatment with integrase inhibitors. Thus, they conducted site-directed mutagenesis to introduce these mutations, individually or in combination, into a wild type virus. Hence, our model summarizes the knowledge derived from these experiments, but it does not allow to identify, yet un-described integrase resistance mutations as it was done with the training data for GENO2PHENO (Sing et al., 2005b). Furthermore, as a consequence of the bias towards well-known resistance mutations, some training instances had identical genotypes. Of course, identical samples can artificially boost the performance assessed in cross-validation settings. In order to study the impact of duplicate samples on the prediction accuracy, we repeated the LOOCV study with a smaller set of unique samples (n=76 for RAL and n=90 for EVG). The observed correlation was slightly worse for RAL ($r = 0.85$) and even slightly better for EVG ($r = 0.81$) than with the full dataset. Nonetheless, analysis of the feature importance could confirm the high cross-resistance potential between the two inhibitors (McColl et al., 2007).

The prediction of phenotypic resistance to the integrase inhibitors RAL and EVG based on SVR models is implemented in the freely available web service GENO2PHENO[integrase][2]. Along with the predicted FC values, a list of mutations contributing to the drug resistance is provided. Currently, prediction of resistance to INIs is decoupled from prediction to resistance to RTIs and PIs, since, as of now, determining the genotype of the viral integrase and the *Pol* fragment containing protease and reverse transcriptase are separate working steps.

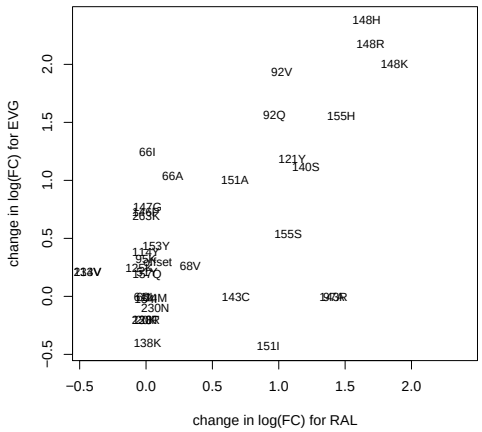
[2] http://integrase.geno2pheno.org



(a) RAL feature importance



(b) EVG feature importance



(c) Correlation of mutations

Figure 3.2: Contribution of each mutation to the predicted phenotypic resistance to RAL (a) and EVG (b). All mutations that cause a change in $\log_{10}$(FC) of more than 0.15 are labeled with the corresponding substitution in red (increase) or green (decrease). Correlation of mutations (c): the x-axis (y-axis) displays the effect on $\log_{10}$(FC) by a specific mutation on RAL (EVG) resistance.

3.2 Estimating Clinically Relevant Cutoffs for GENO2PHENO

The appealing benefit of the phenotypic drug resistance test, and consequently also of the predicted FC in drug resistance, is the output of a scalar value that objectively quantifies drug resistance. These values, however, have to be interpreted in terms of clinical impact. Intuitively, high values predict unfavorable outcome, while low values indicate good response to the drug. Still, where exactly the cutoffs are located, which classify a drug as susceptible, intermediate, and resistant, is unclear. The computation of clinically relevant cutoffs is therefore an essential task for converting an *in vitro* measure into a clinically useful tool. Moreover, the observed ranges of FC vary heavily between drugs. For example, FC for ZDV in the training data for GENO2PHENO has a median of 29 [interquartile range (IQR): 1.5; 225], while ABC has a median of only 3.9 [IQR: 1.7; 8.2]. In GENO2PHENO this scaling issue is addressed by normalizing the raw predicted FC with the predicted FC as observed in treatment-naïve patients (Beerenwinkel et al., 2003a). Briefly, the $\log_{10}(\text{FC})$ of one drug in a group of untreated patients, i.e. patients, who are assumed not to have resistance mutations, is predicted with GENO2PHENO and the mean $\mu_{\text{naïve}}$ and the standard deviation $\sigma_{\text{naïve}}$ are computed. The $\log_{10}(\text{FC})$ for a new prediction is now normalized by computing the z-score

$$z = \frac{\log_{10}(\text{FC}) - \mu_{\text{naïve}}}{\sigma_{\text{naïve}}}.$$

Hence, drug resistance is expressed as the distance (measured in standard deviations) from a treatment inexperienced group, and therefore makes FCs for different drugs more comparable. The problem, however, which values correspond to intermediate or complete drug resistance remains unsolved by this transformation.

The first generation of cutoffs in GENO2PHENO was based on the observation that on a logarithmic scale the distribution of the FC values displays a bimodal distribution (Figure 3.3) – more precisely, a mixture of two Gaussian distributions, with one Gaussian representing the susceptible subpopulation and the other the resistant subpopulation. Hence, the density of the $x = \log_{10}(\text{FC})$ can be modeled as

$$\alpha \cdot \phi(x; \mu_1, \sigma_1) + (1 - \alpha) \cdot \phi(x; \mu_2, \sigma_2),$$

with $\phi(x; \mu, \sigma)$ denoting a normal distribution with mean μ and standard deviation σ (Beerenwinkel et al., 2003a). The model parameters can be efficiently estimated using the expectation-maximization (EM) algorithm (Dempster et al., 1977), and the intersection (between μ_1 and μ_2) of the two (weighted) Gaussians represents a logical candidate for a cutoff between susceptible and resistant. Beerenwinkel et al. (2003a) exploited the bimodal nature of the distribution further for calculating the probability of an FC value of belonging to the susceptible subpopulation, that is where prob(sus|FC) > prob(res|FC) (Beerenwinkel et al., 2003b). To this end, the zero x_0 of the log-likelihood function $l(x) = \frac{prob(sus|x)}{prob(res|x)}$ is computed and the log-likelihood function is approximated by its tangent $L(x)$ at x_0. Finally, the probability of an FC value to belong to the susceptible subpopulation can be approximated with $\text{prob(sus|FC)} \approx \frac{1}{1+\exp(-L(x))}$, where $x := \log_{10}(\text{FC})$. This quantity is also referred to as the *activity* of a drug d against a virus.

Despite its appealing mathematical properties, the probability of belonging to the susceptible subgroup was not favored by the clinicians and virologists, who considered the

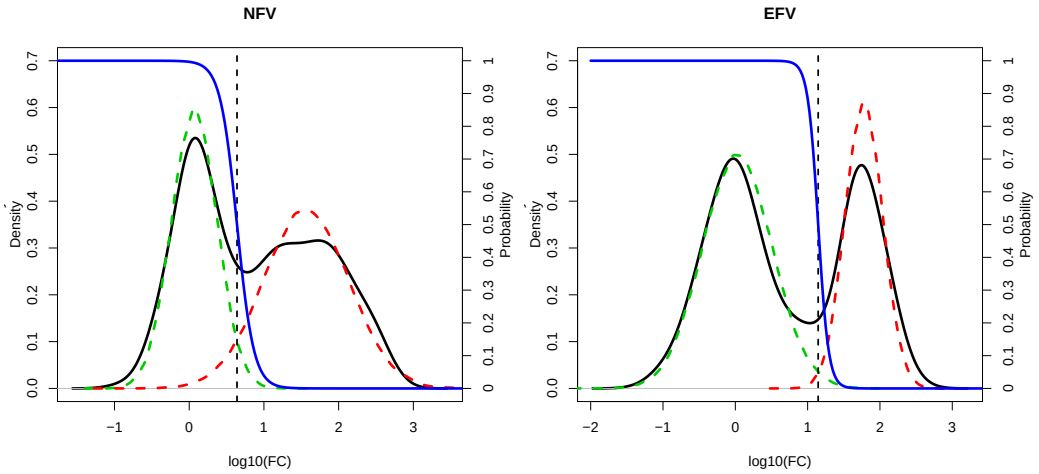


Figure 3.3: Distribution of $\log_{10}$(FC) for the PI NFV and the NNRTI EFV. The Gaussian representing the susceptible (resistant) subpopulation is depicted by a green (red) dashed line. The intersection of the Gaussians is denoted by the vertical dashed line, and the approximation of prob(sus|FC) is given by the blue line.

measure being too conservative. That is, viruses are being declared resistant against a compound although the drug still shows some clinically useful activity in the patient. The current set of cutoffs used in GENO2PHENO are so-called clinically relevant cutoffs, meaning, they reflect the clinical usefulness of predicted *in vitro* resistance. For estimating those cutoff values special treatment changes were manually examined. In these treatment changes only a single drug was added to the regimen and the achieved change in viral load can therefore be fully attributed to this new drug in the combination. Hence, clinically relevant cutoffs can be computed by correlating the predicted drug resistance against the compound to its effect on viral load decline (Däumer et al., 2007). These special treatment changes, however, occur only rarely in databases collecting routine treatment information, and consequently, the estimated cutoffs are unreliable. As a result, the approach was restricted to the drug LPV for which most data was available. A further method for deriving clinically relevant cutoffs (employed for the remaining drugs) was based on the correlation of predicted drug resistance to a drug with the viral load measured during a failing treatment containing that drug. In short, a linear model between $\log_{10}$(FC) and $\log_{10}$(VL) was fitted and the FC corresponding to a VL of $10^{3.41}$ copies per ml (i.e. two standard deviations from the mean of VL in treatment-naïve patients) was selected to be the upper cutoff (Däumer et al., 2007). The lower cutoff could not be estimated with this method and was set to the biological cutoff (i.e. two standard deviations from the mean IC_{50} observed in treatment-naïve patients).

The company Virco with its proprietary FC prediction system VircoTYPE was facing a similar problem. For deriving clinically relevant cutoffs Winters et al. (2008) compiled one dataset per drug comprising only regimens containing that drug. Briefly, the treatments are grouped with respect to the estimated background activity of the other drugs (using a preliminary cutoff), thus allowing for studying the relation between change in viral load and predicted FC of the target drug. The lower and upper cutoffs are defined as the FC

values at which the drug exhibits a 20% and 80% loss in activity compared to a wild type virus, respectively. The thresholds are optimized iteratively per drug, that is, after values for all drugs have been computed sequentially, the next round of cutoff optimization uses the improved values to achieve a better estimate of the background activity. The iterative procedure was stopped when the computed clinical cutoffs remained stable. Moreover, stability of the derived cutoffs was estimated using 1000 bootstrap replicates of the initial data (Winters et al., 2008).

For allowing a more systematic derivation of clinical relevant cutoffs for the GENO2PHENO system we investigated an approach akin to the one applied by Virco. In contrast to that approach, all cutoffs are estimated simultaneously using global probabilistic optimization algorithms like simulated annealing (Kirkpatrick et al., 1983) and genetic algorithms. In the following, we briefly describe our approach.

Material and Methods

Data

The basis for the optimization is a set of treatment change episodes (TCEs) extracted from the EURESIST integrated database. Briefly, a TCE comprises the viral genotype (at most three months prior to treatment start), the prescribed drug combination, and a viral load before (at most three months) and after treatment start (within 4-12 weeks). Treatment success was defined as a reduction of the VL below 500 copies per ml blood or by at least a 100-fold reduction compared to the pre-treatment VL measurement. Details on this *standard datum definition* and the EURESIST database are provided in Chapter 5. Application of the definition to data stored in the current version of the EURESIST database resulted in 5,012 TCEs. Of those, 4,514 instances were randomly selected and formed the development set. The remaining 498 TCEs were used as an independent test set. Success rate was 72.5% in both datasets. Furthermore, FC in drug resistance against the drugs applied in a treatment was predicted with GENO2PHENO.

Global Probabilistic Optimization

Simulated annealing (SA) and a genetic algorithm (GA) were used to compute two cutoffs for every compound. Below (above) the lower (upper) cutoff activity of the drug was rated 1.0 (0.0), representing full (no) activity. Between the cutoffs the activity was set to 0.5.

SA starts with a random set of cutoffs, and a subset of those are randomly modified in each iteration of the algorithm. The cutoffs are used to compute the phenotypic susceptibility score (PSS) of the applied regimen, i.e. simply the sum of individual drug activities. If a modified set of cutoffs improves the correlation between the PSS and the change in VL (ΔVL) on the development set, then it is retained. If, on the other hand, the modified set lowers the correlation, then it is only retained with a small probability. This probability depends on the magnitude of decrease of the correlation and on the runtime of the algorithm. Acceptance of slightly worse solutions is a tool for avoiding to get stuck in local maxima. Precisely, the probability p_{accept} of accepting a worse set of cutoffs is

$$p_{\text{accept}} = \exp\left(\frac{\delta}{T(i)}\right), \quad \text{with} \quad T(i) = \frac{1}{50}\left(1 - \frac{i}{I}\right),$$

where δ is the difference in performance between the old set and the new set (i.e. negative), $T(i)$ is the temperature of the system, I is the maximal number of iterations (here 5000) and i is the current iteration. Based on the current best set of cutoffs, a neighboring set of cutoffs was tested. In order to reduce computational complexity, the continuous IC_{50} values were searched in steps of 0.05. A neighboring set of cutoffs is defined as a set in which no cutoff differs by more than 0.15 (i.e. three grid steps) from the corresponding value in the current best set. Finally, it was ensured that lower cutoffs were smaller than upper cutoffs. Among all neighboring sets of cutoffs, one set was randomly selected for being tested in the next iteration.

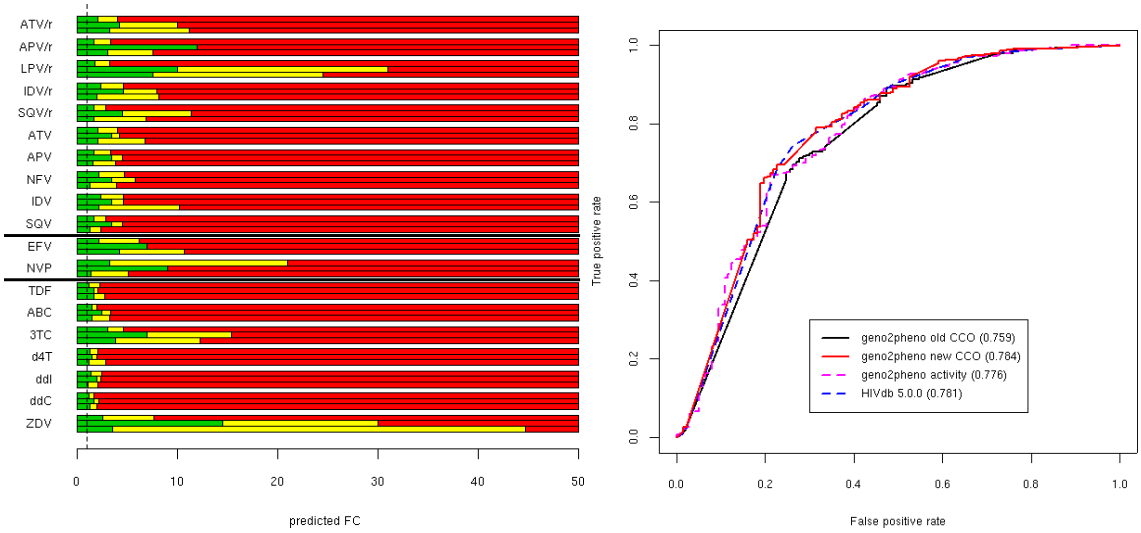
In contrast to SA, GA maintains multiple sets of cutoffs (the population) with each set containing cutoffs for all drugs (one individual). In every iteration of the algorithm (generation), the chance to copy a set of cutoffs depends on its performance (fitness). During the copying process the cutoffs are randomly modified (mutated) and even parts of individuals might be exchanged (cross-over). The algorithm terminates when a stopping criterion is fulfilled. Usually, the stopping criterion comprises a maximum number of generations and a maximum number of generations without substantial improvement (stall generation). For GA we used the implementation of the genetic algorithm and directed search toolbox of the programming environment Matlab.

For robust estimates of the cutoffs, SA and GA were carried out on 100 bootstrap replicates of the original development set and the mean of the cutoffs obtained from the bootstrap replicates represents the final set of cutoffs. This set was used to compute PSS values on the validation set where the PSS was correlated to ΔVL and to the binary outcome – performance was measured as Pearson correlation (r) and area under the receiver operating characteristics (ROC) curve (AUC). Briefly, ROC curves depict classifier performance by giving a true-positive rate (TPR; percentage of correctly predicted successes) for every false-positive rate (FPR; percentage of failing therapies that were predicted to be successful). The AUC summarizes the performance and is a convenient measure for comparing scoring systems without the need to provide a particular cutoff (Brun-Vézinet et al., 2004). The AUC is a value between 0 and 1 corresponding to the probability that a randomly selected success receives a higher score than a randomly selected failure (Fawcett, 2006). The achieved performance was compared to the old cutoffs, to the original activity scores, and to Stanford's HIVdb.

In addition to the original approach, we investigated slight modifications. Namely, the use of AUC instead of correlation as the target function, a linear interpolation of the intermediate region (instead of simply using 0.5), and the combination of both.

Results

Figure 3.4 a) depicts the cutoffs based on the activity score, the old manually derived clinical cutoffs, and the set of new automatically derived clinical cutoffs. This new set of cutoffs was derived with GA and correlation between PSS and ΔVL as a target function without interpolation between the cutoffs. Of note, cutoffs based on the bimodal activity model tend to be lower for most drugs than the (old or new) clinical cutoffs, and thereby confirm the subjective impression of virologists and clinicians. When used with linear interpolation between the lower and upper threshold, the new cutoffs achieve an AUC



(a) clinical cutoffs (b) ROC curve

Figure 3.4: New and old clinical cutoffs (CCO) in comparison (a). Cutoffs based on the bimodal activity model are shown in the upper bar. The old clinical cutoffs are depicted in the middle bar, and the newly derived cutoffs are depicted below. Green corresponds to susceptible, yellow and red indicate intermediate and full resistance, respectively. Performance of different sets of cutoffs (b). Performance achieved with the different sets of cutoffs on the validation data in comparison to a HIVdb 5.0.0 based prediction.

(r) of 0.784 (0.469), compared to 0.759 (0.433) with the old cutoffs, 0.776 (0.461) with the activity score, and 0.781 (0.457) with HIVdb using the five state interpretation. GA provided an improved set of clinical cutoffs for GENO2PHENO that performed slightly better than (as good as) HIVdb with respect to correlation (AUC). Figure 3.4 b) depicts the ROC curves on the validation set computed with the old and new clinical cutoffs and the reference methods HIVdb and activity scores. Based on a method by DeLong et al. (1988) the significance of the difference between two AUC values could be assessed. Only improvements over the old clinical cutoffs by any other method showed a statistical trend (new clinical cutoffs: $p = 0.0877$; activity score: $p = 0.1097$) or even statistical significance (HIVdb: $p = 0.0204$). The p-values for differences between the remaining pairs of methods were all larger than 0.19.

Table 3.1 lists the achieved performance on the validation set using either 0.5 for the intermediate resistance or the linear interpolation. Of note, for simulated annealing it beneficial, in general, to use interpolation during the estimation of the cutoffs, in addition, using the correlation between PSS and ΔVL as target function optimized both r and AUC while AUC as target function results in worse r on the validation set. For GA such conclusions cannot be drawn. Here, using AUC instead of correlation as optimization function achieved slightly better performance in terms of AUC. The upper thresholds for some drugs, however, were extremely large, e.g. ZDV and 3TC received approximately an upper cutoff of 100-fold (data not shown). That is, for these drugs, viruses were regarded as resistant only for FC values of 100 and more. Overall, cutoffs derived with GA yielded

			validation							
			simulated annealing				genetic algorithm			
			r		AUC		r		AUC	
			D	I	D	I	D	I	D	I
development	r	D	0.455	0.464	0.763	0.774	0.472	0.469	0.780	0.784
		I	0.465	0.467	0.759	0.766	0.468	0.471	0.774	0.784
	AUC	D	0.421	0.396	0.759	0.756	0.467	0.463	0.786	0.789
		I	0.446	0.436	0.752	0.751	0.479	0.466	0.784	0.790

Table 3.1: Performance of different sets of automatically derived clinical cutoffs. Performance was measured in correlation (r) and AUC on the independent validation set. The rows denote the setting in the training stage, i.e. the different target functions and whether interpolations between thresholds was used (I for interpolation, D for discrete). The columns correspond to different settings in the validation setting. The first (last) 4 columns denote performance of cutoffs derived with simulated annealing (genetic algorithm).

better performance on the independent validation set than the cutoffs derived with SA when the same training setup was used. The method of DeLong et al. (1988) indicates that the observed improvements in AUC are significant (p-values range from 0.0502 to 0.0041).

Clearly, SA and GA are not the only probabilistic global optimization methods. Other methods like ant colony optimization (Dorigo and Gambardella, 1997) can also be used to simultaneously optimize the cutoffs. However, the genetic algorithm performs well, and it is likely that modifications of the target function and the computation of the PSS play a more pivotal role in the final quality of the cutoffs than the optimization method.

3.3 Improvements Using Semi-Supervised Learning

The advantage of data-driven methods for developing decision support systems, namely that information is only derived from data without interference of expert opinions, is also their major weak point: a sufficient amount of data is required during the training step of the models. This weakness becomes more evident when new drugs are released. Rules-based systems can easily be extended on the basis of first reports of the drug in clinical trials or extended access programs. For the phenotypic resistance test, which is an important step during data generation, the drug has to be available to the laboratory conducting the assay. This, however, is usually only the case when the drug is already FDA approved. Consequently, there exists a serious time lag between approval of a drug and the availability of sufficient data for training statistical models.

In a recent work we investigated the use of semi-supervised learning (SSL) methods for improving the prediction of drug resistance (Perner et al., 2009). Briefly, SSL uses labeled data (genotype-phenotype pairs) and unlabeled data (just genotypes from routine diagnostic) to derive improved models. The idea behind SSL is that unlabeled data can provide information on the distribution of the data in the input space. This information

can be exploited by the machine learning algorithm to avoid the separation of clusters, since points in the same cluster are assumed to belong to the same class (cluster assumption). Moreover, SSL methods generally assume that samples that are close in the input space are also close in the output space (smoothness or manifold assumption). Zhu (2007) provides an introduction into SSL and a survey on available SSL methods.

We examined two types of unlabeled data: the first category originates from patients that were treated with the particular drug (for which we are building a model) at the time the viral sequence was obtained (S_{drug}), while the second category required only the use of one drug of the same drug class (S_{class}) at the time of genotyping. The second set of unlabeled data can be expected to be larger but probably less informative for the learning as the virus was (at the time of genotyping) not exposed to the drug in question. We found that if all labeled data (up to 1,000 genotype-phenotype pairs) were used, then the SSL learning methods showed little benefit over the standard supervised learning methods regardless of the type of unlabeled data. If, however, only little labeled data were available (e.g. 10% of all labeled data) then the SSL methods showed a significant improvement (Figure 3.6; upper row). Unfortunately, the benefit could not be witnessed for all drugs/drug classes and all tested SSL methods (data not shown). In fact, a violation of the SSL assumptions by the underlying data, led to inferior models. For example, the performance of the 3TC model was heavily corrupted by SSL methods. In the case of 3TC, one amino acid exchange is sufficient to confer complete resistance, while for other NRTIs several mutations are necessary. NRTIs are usually given to the patients in pairs. Thus viruses that were exposed to 3TC were also exposed to other NRTIs with more complicated resistance patterns. As a consequence, the data density does not reflect the labeling of 3TC resistance, which is a clear violation of the smoothness assumption.

The transductive SVM (tSVM) is a SSL version of the SVM and, just like its supervised counterpart, provides a decision boundary (see Eqn. 3.1) after the training phase. Briefly, the standard soft margin SVM optimizes the following function:

$$\min_{\vec{w}} \frac{1}{2}\|\vec{w}\| + C\sum_{i=1}^{N}\xi_i, \text{subject to } \xi_i \geq 0, y_i(\vec{w}\cdot\vec{x_i}+b) \geq 1-\xi_i, \forall_i \tag{3.2}$$

where N is the number of labeled training instances, $\vec{x_i}$ and y_i are the features and the label of the ith instance, respectively, $\vec{w}$ and b define the hyperplane, ξ_i are the slack variables that allow for misclassification and C is the cost parameter for misclassified examples. The tSVM aims at determining a separating hyperplane under consideration of the M unlabeled samples $\{\vec{x}^*_1, \ldots, \vec{x}^*_M\}$, therefore Eqn. 3.2 is extended in the following way:

$$\min_{\vec{w}} \frac{1}{2}\|\vec{w}\| + C\sum_{i=1}^{N}\xi_i + C^*\sum_{j=1}^{M}\xi^*_j, \text{subject to}$$
$$\xi_i, \xi^*_j \geq 0, y_i(\vec{w}\cdot\vec{x_i}+b) \geq 1-\xi_i, y^*_j(\vec{w}\cdot\vec{x}^*_j+b) \geq 1-\xi^*_j, \forall_{i,j} \tag{3.3}$$

where the additional parameters ξ^*_j and C^* are the slack variables and the misclassification cost parameter for the unlabeled instances, respectively. Thus, the optimization problem in Eqn. 3.3 differs from Eqn. 3.2 in that the tSVM has to find a labeling $y^*_1, \ldots, y^*_m$ for the unlabeled data and a hyperplane $< \vec{w}, b >$ simultaneously. An approximative optimization procedure, which is required due to the complexity of the optimization problem, has been

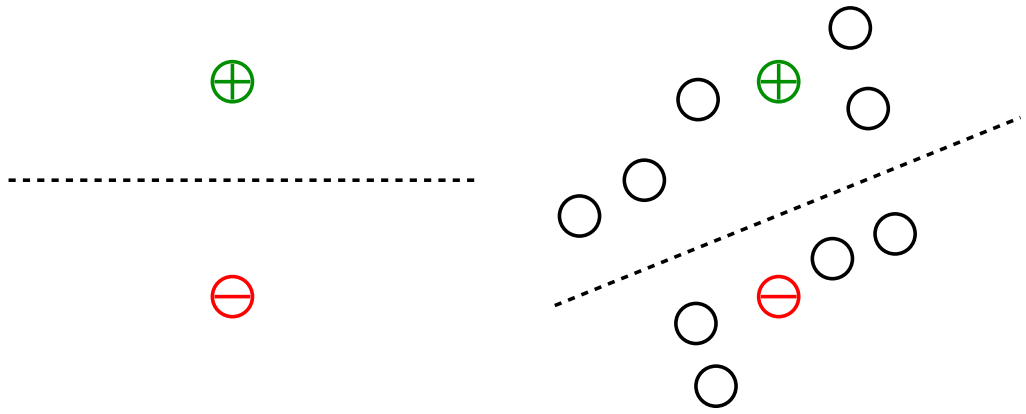


Figure 3.5: Decision boundaries fitted by SVM-based models. Decision boundary derived only from labeled instances (colored circles) using a supervised SVM method (left). The transductive SVM takes also the unlabeled samples (black circles) into account when fitting the decision boundary (right).

implemented in the software library SVM^{light} by Joachims (1999). The approach begins with a labeling of $\vec{x}_1^*, \ldots, \vec{x}_m^*$ based on the classification of an inductive SVM and a low weight C^* for the penalty for misclassified unlabeled data points. Then the labels of two randomly selected samples (one positive and one negative) are swapped. If the objective function is improved by that exchange of labels, then the switch is made permanent. This process is repeated until there are no more switches possible that yield an improved objective function. At this point the penalty for misclassified unlabeled data points C^* is increased and further labels are swapped to greedily improve the objective function. The iterative procedure stops when C^* exceeds a user defined value. Figure 3.5 depicts the concept of transductive SVMs.

The resulting hyperplane (i.e. decision boundary) has the same form as the one derived with the standard supervised SVM approach. Consequently, the boundary derived with an tSVM can be used to assess the class of new unseen samples. The ability of classifying unseen samples is not granted for SSL methods. In fact, many methods only provide a classification for all unlabeled samples that are available during the training phase. Hence, classification of a new sample requires retraining of the entire model with all previous training data and the new unlabeled sample. For instance, low-density separation (Chapelle and Zien, 2005), one of the SSL methods we studied, cannot be used on unseen samples. This inability is a direct result from the fact that low-density separation (LDS) constructs a kernel which is based on all samples (labeled and unlabeled) in training data. LDS is only capable of making predictions for samples that were used to build the kernel.

Nevertheless, the benefit of using tSVMs in the transductive setting (i.e. only prediction of unlabeled samples used in the training phase) that we previously observed (Perner et al., 2009), could be transferred to the inductive setting (i.e. prediction of new unseen samples without retraining). Figure 3.6 depicts the learning curves (size of labeled training data versus model performance) for the two drugs LPV and ZDV using the tSVM implementation SVM^{light} by Joachims (1999) and both types of unlabeled data. A standard supervised SVM served as reference method. The results show that with only very little labeled data

(e.g. 2.5%) the SSL models almost reach the performance of the supervised method using all available labeled data.

Concluding, SSL does not yet provide a useful tool for improving the prediction of resistance for the first drug in a novel drug class (e.g. integrase inhibitors), simply due to the fact that only little relevant unlabeled data exist at the time when the drug is released. For instance, there are currently about 3800 integrase sequences stored in the Los Alamos HIV Sequence Database (http://www.hiv.lanl.gov/), compared to about 75.000 sequences comprising the protease. Moreover, since prior to the release of *raltegravir* there were no integrase inhibitors available, the majority of the available sequences will not contain any resistance mutations. The question, whether SSL provides a benefit for a new drug in an established drug class, depends on the resistance profile of the class and how the drugs are used. For instance, the NNRTI and NRTI models suffered from the circumstance that multiple drugs with different resistance patterns were simultaneously targeting the same viral protein, while PIs generally benefited from SSL. Likewise, future INIs might benefit from sequence data generated after exposure to *raltegravir*.

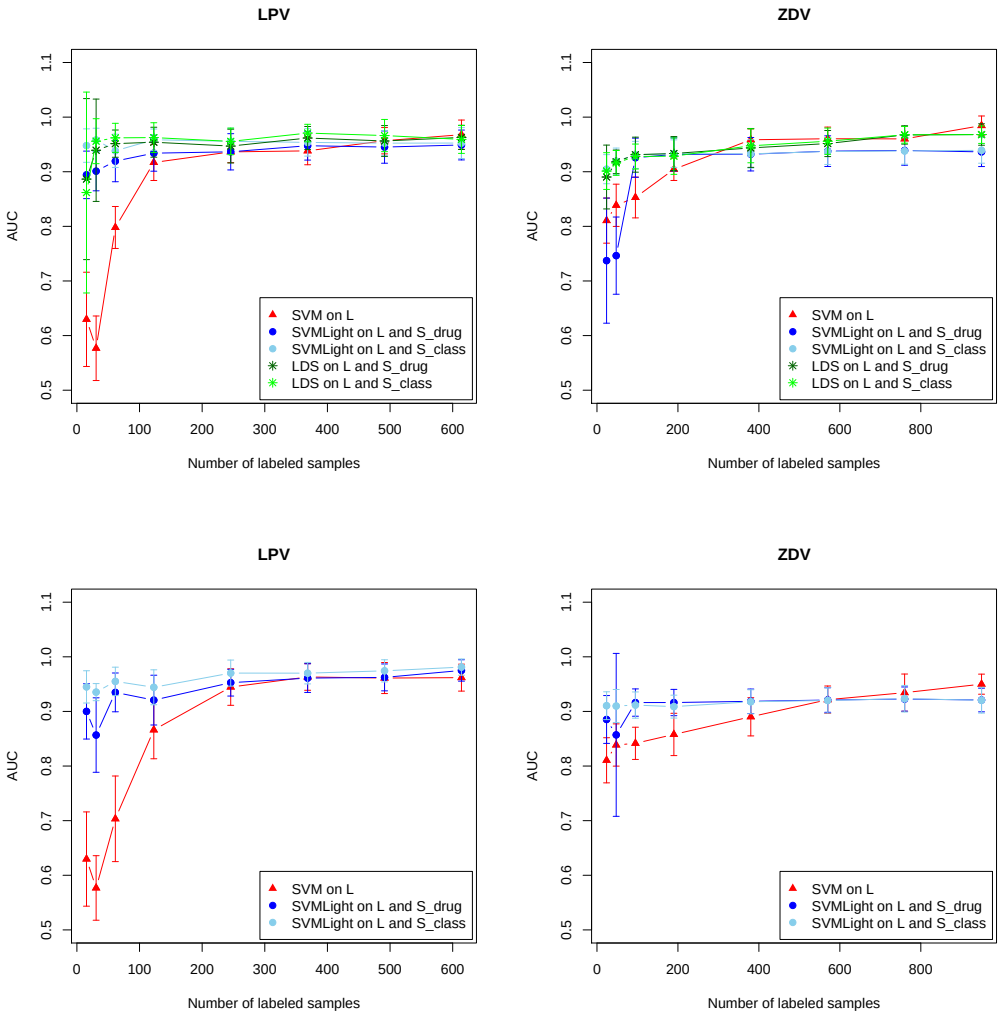


Figure 3.6: Learning curve for the drugs LPV and ZDV. Model performance is given as the area under the ROC curve (AUC) and based on a 10-fold cross-validation. For LPV (ZDV) a total of 682 (1055) labeled samples (L) were available. For the tSVM 1442 (2717) or 4435 (7887) sequences, which were exposed to LPV (ZDV) or at least one PI (NRTI) at the time of sequencing, respectively, were used. The performance was estimated using 2.5%, 5%, 10%, 20%, 40%, 60%, 80%, and 100% of the labeled data. The upper row depicts the performance of tSVM and LDS in the transductive setting (i.e. test instances were used as unlabeled training data). By contrast, the lower row depicts the performance of tSVM in the inductive setting.

4 Predicting Response to Combination Therapy

In Chapter 2 we introduced the basis of modern anti-HIV therapy. In order to ensure the selection of potent regimens, several genotype interpretation systems are used to infer *in vitro* drug susceptibility and/or *in vivo* response to antiretroviral treatment on the basis of HIV-1 genotype. Most of these tools use a set of rules carefully crafted by experts and classify the virus as susceptible, intermediate, or resistant to each of the individual compounds. Few tools are fully data driven rather than based on expert knowledge. The most prominent examples are GENO2PHENO (Beerenwinkel et al., 2003a) and VirtualPhenotype (Vermeiren et al., 2007), which both apply methods from statistical learning for predicting *in vitro* resistance on the basis of genotype. Although all of these methods are designed to infer susceptibility to individual compounds (Vercauteren and Vandamme, 2006), recently developed decision support tools are being explored to infer virological response directly to a typical 3-4-drug HAART regimen. In one study (Larder et al., 2007), artificial neural networks were used to predict the change in viral load, given the sequence, regimen, and additional host-specific features. In Section 4.2 we introduce the software pipeline geno2pheno-THEO, which predicts the probability of reaching an undetectable viral load during the course of the regimen given the applied drug combination and the genetic makeup of the viral population. Geno2pheno-THEO uses covariates that encode the estimated viral evolution during treatment in addition to drug combination and viral genotype. These evolutionary features are based on works by Beerenwinkel et al. and are briefly summarized in Section 4.1. The resulting statistical model is evaluated on a large external database in a comparison to well-established HIV genotype interpretation systems (Section 4.3).

4.1 Features of Viral Evolution

The features estimating viral evolution that are used to improve the prediction of response to antiretroviral treatment are based on works by Beerenwinkel et al. and are briefly summarized in the following sections.

4.1.1 Activity Score

The activity score for combination treatments is based on the activity score for individual drugs as computed by GENO2PHENO. Precisely, let D be the set of all drugs and $T = \{d_1, \ldots, d_n\} \subseteq D$ be a combination of n drugs. Furthermore, T_c represents all drugs of T belonging to drug class $c \in \{\text{NRTI}, \text{NNRTI}, \text{PI}\}$. The activity of a treatment T against

virus seq is defined as:

$$\text{activity}(T, \text{seq}) = \sum_c \max_{d \in T_c} \text{activity}(d, \text{seq}),$$

where $\text{rf}_d(\text{seq})$ denotes the resistance factor (RF) of the virus seq against drug d, and $\text{activity}(d, \text{seq}) = \text{prob}(\text{sus}|\log \text{rf}_d(\text{seq}))$ as defined in Section 3.2. Inhibitors sharing drug target and mechanism of action are competing, e.g. for the binding site, thus the activity score is restricted to the most active representative of a class (Jordan et al., 2002). No such restrictions hold for drugs from different drug classes, and consequently the activity score is additive for inhibitors from different drug classes. Obviously, the number of different drug classes constitutes the upper bound of the activity for treatments. Thus, the activity score is in [0,3] for our case with at most three drug classes.

Maximizing over drugs in the same class might at first glance seem limiting, since there is no immediate benefit of administering multiple drugs from the same class. The benefit becomes evident though, when considering the change of the activity score during the course of treatment. The selective pressure posed by the treatment leads to mutations in the viral genome, which in turn might affect drugs from the same class differently and thus the most active drug from one class changes over time. For instance, at start of a regimen comprising ZDV and 3TC the RT of a virus shows only the T215Y mutation, which solely reduces the susceptibility of ZDV. Thus, the activity of 3TC is unharmed and consequently the activity of the treatment is 1. During the time of therapy, the virus develops another RT mutation. For instance, M184V that completely inhibits the activity of 3TC but does not affect ZDV at all. The activity score for the combination treatment is now equal to the activity of ZDV, in contrast to a 3TC monotherapy where the activity score would have decreased immediately to 0.

This example illustrates that, in addition to the resistance to drugs at treatment start, the long-term success of an antiviral therapy depends greatly on the ability of the virus to escape from selective pressure presented by the treatment. An estimate of the virus' ability to escape from the treatment might therefore be an useful information for predicting the success of combination treatments. Such an estimate can be derived by studying changes of activity in the mutational neighborhood of the virus.

Due to the high dimension of the sequence space, i.e. there are 20^{319} possible amino acid sequences of length 319, the neighborhood of the virus is explored using a beam search. Precisely, starting from the original sequence, all variants with one additional mutation are generated and the activity of the treatment against all variants is computed. The resulting activity score for these *in silico* mutants is used to rank the variants. Only for the top b mutants showing the least activity all variants with one additional mutation are generated and their activity score is computed. The search stops when d mutations are introduced into the original sequence. The breath b and depth d are the two parameters of the beam search. The activity score at depth r is defined by:

$$\text{activity}_r(T, \text{seq}) = \min_{\text{seq}' \in B_r} \text{activity}(T, \text{seq}'),$$

where B_r denotes the neighborhood of the original sequence with r additional mutations. B_r is generated from the b viruses with the lowest activity score in B_{r-1} by adding an additional mutation.

For further details on an the activity score see the work by Beerenwinkel et al. (2003b).

4.1.2 Mutagenetic Trees

The beam search approach assumes that every mutation is equally likely, and that only the increase in resistance is important. This, however, is an oversimplification. For example, there exist well-known resistance pathways for the first approved anti-HIV drug zidovudine. The thymidine analog mutation (TAM) pathway 1 and 2 comprise the RT mutations 41L, 210W, and 215F/Y and 67N, 70R, and 219E/Q, respectively. The mutations belonging to one pathway are mainly accumulated in a preferred order (hence pathway). The molecular causes for the preferred order or for the existence of two distinct pathways are still unknown. However, these pathways occur so frequently that they were identified by experts without the use of computational analyses.

Beerenwinkel et al. (2005b) introduced mixtures of mutagenetic trees to estimate mutational pathways from cross-sectional data. Briefly, a mutagenetic tree is a weighted directed tree with vertices representing the appearance a new mutation (event). Edge weights are conditional probabilities between vertices with the constraint that the parent event has to be present before the child event can occur. Thus, every path from the root of the tree to a vertex represents the ordered accumulation of mutations. Basically, mutagenetic trees are restricted Bayesian networks and constitute an extension of the oncogenetic tree model introduced by Desper et al. (1999) that was restricted to undirected trees. A further extension of the original approach is the ability to generate mixtures of mutagenetic trees.

Formally, a mutagenetic tree is a quartuple $T = (V, E, r, \rho)$, with $V = \{1, \ldots, l\}$ being a set of vertices representing the events, E being the set of edges, $r \in V$ being the root vertex that encodes the absence of any events, and $\rho : E \to [0, 1]$ being the mapping from edges to conditional probabilities. A single tree is constructed by first building the complete digraph $G = (V, V \times V, w)$ on all vertices. The weights w are defined as

$$\forall u, v \in V : \; w(u, v) = \log Pr(u, v) - \log(Pr(u) + Pr(v)) - \log Pr(v),$$

where $Pr(u)$ and $Pr(u, v)$ denote the marginal probability of event u and the joint probability of events u and v, respectively, as estimated from the data. The edge weights are the logarithm of the independence of the mutations $\frac{Pr(u,v)}{Pr(u)Pr(v)}$ multiplied by the direction of the dependence $\frac{Pr(u)}{Pr(u)+Pr(v)}$. The mutagenetic tree is defined as the branching in G that maximizes the sum of its edge weights. Using the algorithm by Edmonds (1967) the maximum weighted branching can be computed in $O(|V||E|)$ time, for a fully connected graph this corresponds to $O(|V|^3)$. The weighting function $w(u, v)$, however, displays an undesired behavior for edges leaving the root node, since the less likely events receive a higher score $w(r, v) = -\log(1 + Pr(v))$. Thus, for edges leaving the root node the alternative weighting function $w(r, v) = \log Pr(v)$ is used.

The likelihood of a mutational pattern x is the probability that a given mutagenetic tree T generates x: $L(x|T) = Pr(x|T)$. Let $S \subseteq V$ be the set of events defined by the mutational pattern x. If there is a subset of edges $E' \subseteq E$ such that exactly all vertices of S can be reached in the subtree (V, E'), then x can be generated by T with likelihood:

$$L(x|T) = \prod_{e \in E'} Pr(e) \prod_{e \in \{S \times V \setminus S\}} (1 - Pr(e)).$$

The first product computes the probability of reaching all events in S and the second product represents the probability of not going beyond the events in S. If there is no

appropriate subtree of T, then the pattern x cannot be generated by T. Hence, $L(x|T) = 0$. For instance, the right tree in Figure 4.1 does support the mutational pattern 215F, 41L but not the pattern 215F, 41L, and 70R, since the mutations 67N and either 219E or 219Q have to be acquired before the mutation at position 70. Thus, the likelihood of the pattern 215F, 41L, and 70R to be generated by that tree is 0. On the other hand, the likelihood of the 215F, 41L is $0.4 \times 0.77 \times (1 - 0.61) \times (1 - 0.45)$.

The mixture of mutagenetic trees is estimated from cross-sectional data in an EM-like learning algorithm (Dempster et al., 1977). Formally, a k-mutagenetic tree mixture model is defined as

$$M = \sum_{i=1}^{k} \alpha_i T_i \text{ with } \alpha_i \in [0,1] \wedge \sum_{i=1}^{k} \alpha_i = 1,$$

with $T_i = (V, E_i, r, \rho_i)$. Consequently, the likelihood of a pattern x given the mixture model is defined as

$$L(x|M) = \sum_{i=1}^{k} \alpha_i L(x|T_i).$$

In order to handle noisy real-world data, one component of the mixture is forced to attain a star-like topology modeling the independence of mutations. Briefly, during the learning phase we want to find the mixture of k mutagenetic trees that maximize the log-likelihood of the training data:

$$\sum_{n=1}^{N} \log \sum_{i=1}^{k} \alpha_i L(x_n|T_i). \tag{4.1}$$

The responsibility γ_{ni} of the tree T_i for the training sample x_n is defined as the probability of x_n being generated by T_i given the mixture model M. In the *M-like step* of the training algorithm the parameters of the current mixture model, i.e. the edge weights of the star component, the $k-1$ trees, and the mixture parameters α_i, are updated. In the *E step* the responsibilities are computed using the updated model. E and M step are iterated until the log-likelihood converges. At the initialization of the algorithm initial responsibilities are assigned according to a $(k-1)$-means clustering of the training data. Algorithm 1 is adapted from (Beerenwinkel et al., 2005b) and summarizes the complete training process.

For each antiretroviral drug one mixture of mutagenetic trees is trained. The training data for each drug comprises sequences from patients that were obtained during treatment with that drug. Figure 4.1 depicts a 2-mutagenetic tree mixture model learned for the drug ZDV. The left component is a star that supports every possible pattern and therefore models the noise of the data. The right component is a tree estimated from data. The mixture weights α are depicted above the components. The actual tree estimated from the data is able to generate 72% of the instances found in the training data. The mixture of mutagenetic trees can be used to derive two evolutionary features. The genetic progression score provides an estimate on the expected waiting time for a mutational pattern to occur and therefore summarizes how advanced the virus in terms of drug resistance is. The genetic barrier to drug resistance is the probability that the virus will not escape from drug pressure by developing further mutations. Both features are briefly described in the next sections.

Algorithm 1 k-mutagenetic tree learning $(X = (x_{nj})_{1\leq n\leq N, 1\leq j\leq l}, k)$

N is the number of training instances, and l is the number of events

1. Guess initial responsibilities:

(a) Run (k-1)-means clustering algorithm

(b) Set responsibilities

$$\gamma_{ni} = \begin{cases} \frac{1}{2} & \text{if } x_n \text{ is in cluster } i-1, \\ \frac{1}{2(k-1)} & \text{else.} \end{cases}$$

2. *M-like step.* Update model parameters:

$N_i = \sum_{n=1}^{N} \gamma_{ni}\ \forall i \in \{1,\ldots,k\}$

T_1 is a star with edge weights: $\beta = \frac{1}{lN_1}\sum_{j=1}^{l}\sum_{n=1}^{N}\gamma_{n1}x_{nj}$

$\forall i \in \{2,\ldots,k\}$:

(a) Estimate the joint probability for all pairs of events (u, v):

$$p_i(u,v) = \frac{1}{N_i}\sum_{n=1}^{N}\gamma_{ni}x_{nu}x_{nv}.$$

(b) Compute the maximum weight branching T_i from the complete digraph with weights w derived from p_i.

(c) Compute the mixture parameter $\alpha_i = \frac{N_i}{N}$.

3. *E step.* Compute responsibilities:

$$\gamma_{ni} = \frac{\alpha_i L(x_n|T_i)}{\sum_{m=1}^{k}\alpha_i L(x_n|T_m)}.$$

4. Iterate steps 2 and 3 until convergence, i.e. no changes in Eqn. 4.1.

return $T_1,\ldots,T_k$ and $\alpha_1,\ldots,\alpha_k$

The Genetic Progression Score

The genetic progression score (GPS) was introduced by Rahnenführer et al. (2005). Briefly, a timed mutagenetic tree can be obtained by assuming independent Poisson processes for the occurrence of events on the edges and for the sampling time of the virus. The difference of occurrence time of the event j and its parents is denoted by Z_j. Since event j was generated by a Poisson process, the waiting time Z_j is exponentially distributed with parameter λ_j. Furthermore, let the sampling time of the virus Z_S be exponentially distributed with parameter λ_S. The probability of event j can be computed as $Pr(j) = \lambda_j/(\lambda_j + \lambda_S)$. Thus, the expected waiting time for event j is defined as:

$$E[Z_j] = \frac{1}{\lambda_j} = \frac{1-Pr(j)}{Pr(j)\lambda_S} = \frac{1-Pr(j)}{Pr(j)}E[Z_S].$$

Estimation of the λ_S from real data is usually impossible, since the time of infection is usually unknown and the virus might have had events before starting the particular drug. Thus, we set $E[Z_S] = 1$, which defines a unit-less waiting time. We emphasize that the GPS is not intended to estimate absolute waiting times. Rather it provides a dimension-free

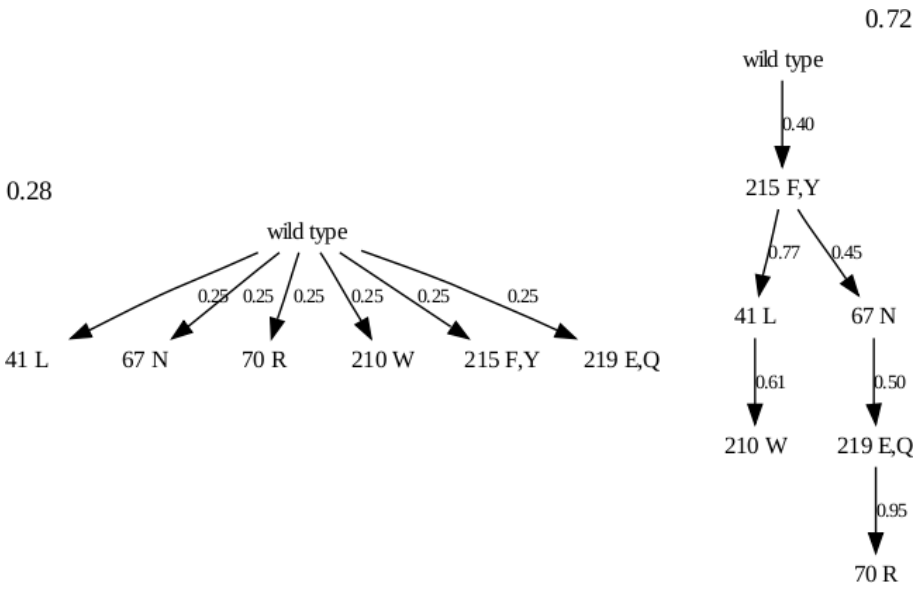


Figure 4.1: Example Mutagenetic tree mixture.

measure of genetic progression that allows for comparing mutational patterns of different viruses.

Unfortunately – apart from the case of a single event – there exists no formula for computing the waiting time for arbitrary pattern of events. However, within the timed mutagenetic tree model the expected waiting time can be estimated by simulating the waiting process. To this end, the estimated probabilities $Pr(j)$ are used to compute the required $\lambda_j = Pr(j)/(1-Pr(j))$. For a sufficiently large number of simulations one receives a stable estimate for the waiting time Z_x for the mutational pattern x with respect to the distribution induced by the mutagenetic mixture model M. This waiting time is the genetic progression score: $\mathrm{GPS}(x) = E_M[Z_x]$.

Of note, the GPS of a virus is computed for every drug (mutagenetic tree model) separately. Due to the definition of $E[Z_S] = 1$ the waiting times cannot be compared between drugs.

The Genetic Barrier to Drug Resistance

Instead of inspecting the past of the virus, i.e. to which state(s) in the (mixture of) tree(s) the virus already has progressed, we can use the evolutionary model for studying the future of the virus, i.e. to which states will it most likely move. Briefly, the genetic barrier is defined as the probability of not escaping from a drug by developing further mutations. With escape from a drug being defined as exceeding a predefined threshold of phenotypic drug resistance.

In the mutagenetic tree model it is sufficient to describe the given viral sequence as one of 2^l possible mutational patterns. Given a tree topology one can now compute the transition probability from the current pattern to any of the 2^l possible patterns. If the tree topology does not allow the transition – e.g. when loosing mutations, then the probability is set to 0. Furthermore, using the paired genotype-phenotype data, which was used to

train GENO2PHENO, one can compute the mean phenotypic resistance associated with each mutational pattern. Due to the limited number of genotype-phenotype pairs it is likely that some mutational patterns do not occur, in this case the weighted mean of all mutational patterns with non-zero transition probabilities is used, with the transition probabilities serving as weights. The genetic barrier to drug resistance is then simply defined as the sum of all transition probabilities to mutational patterns with phenotypic resistance below a certain threshold.

The genetic barrier to drug resistance combines evolutionary information, the transition probabilities, with phenotypic information.

4.2 geno2pheno-THEO

This section describes a joint work with Niko Beerenwinkel, Tobias Sing, Igor Savenkov, Martin Däumer, Rolf Kaiser, Soo-Yon Rhee, W Jeffrey Fessel, Robert W Shafer, and Thomas Lengauer. The work was published in *Antiviral Therapy* under the title "Improved prediction of response to antiretroviral combination therapy using the genetic barrier to drug resistance" (Altmann et al., 2007a).

With about 24 drugs available and novel drugs being approved almost every year, it becomes increasingly difficult for the treating physician to select an optimal drug combination. There are a variety of different therapeutic goals, including VL reduction, increase in $CD4^+$ T cell counts, minimization of adverse effects and preservation of future drug options. Importantly, different optimization criteria will tend to favor different therapies (Jiang et al., 2003). To date, most methods predict virological response to therapy based on the baseline genotype and the compounds in the applied combination. Specifically, artificial neural networks (Wang et al., 2003) and fuzzy rules combined with a genetic algorithm (Prosperi et al., 2004) were used in this manner to predict the change in VL. Related approaches include the application of case-based reasoning (Prosperi et al., 2005) and combinatorial optimization based on expert rules (Lathrop and Pazzani, 1999).

In this section, we report new ways of analyzing treatment change episodes (TCE) and demonstrate how these new approaches might be more useful than current methods for prediction of response to therapy. The improvement results from incorporating genetic analysis, phenotypic prediction, and a prediction of the probability that further evolution of resistance will occur. The applied methodologies involve various techniques of statistical learning. We dichotomize virological response and compare the performance of several classifiers that predict the therapeutic success or failure of each genotype-therapy pair. The novel features derived from genotype and drug combination encode information about the evolutionary potential of the virus and the predicted level of phenotypic drug resistance. Specifically, we consider predicted phenotypes (see Section 2.5.2), the activity score based on a heuristic search over *in silico* mutants (see Section 4.1.1), the genetic barrier and the genetic progression score (see Section 4.1.2). Analyzing over 6,300 TCEs observed in a clinical setting, we show that all new descriptors significantly improve prediction of therapy outcome, especially if they combine evolutionary and phenotypic information.

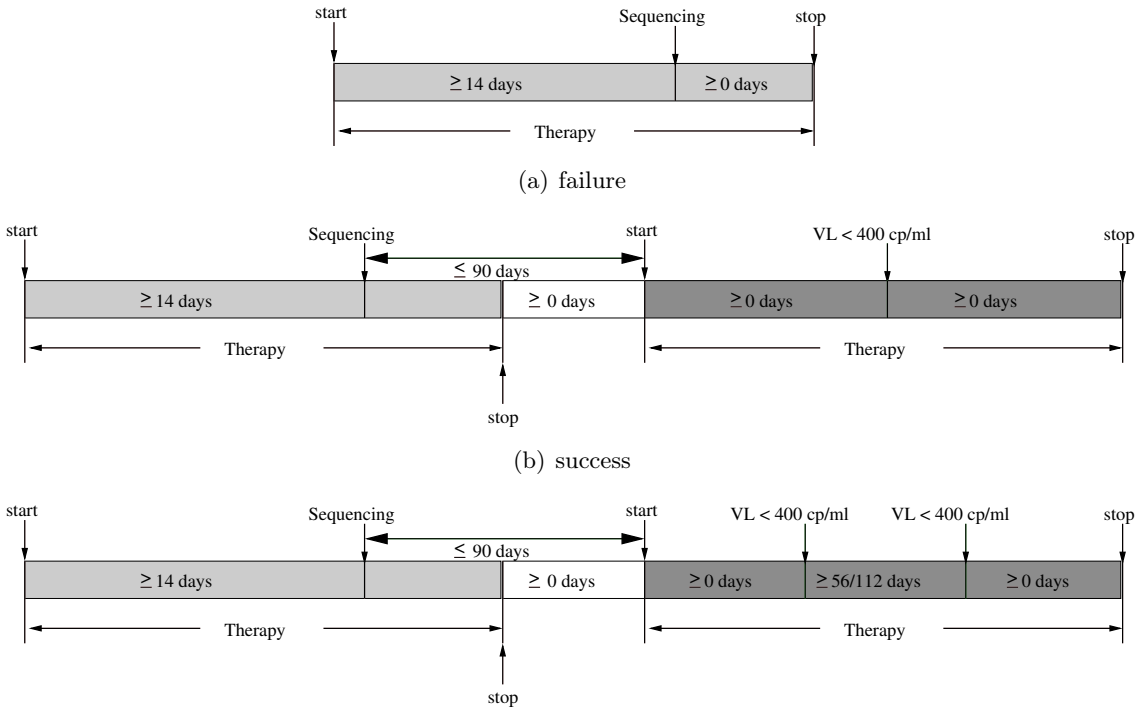


Figure 4.2: A treatment change episode (TCE) resulting in sequencing is always considered to be a failure (a). If the virus is undetectable during the treatment following a failure, then the TCE is called a success (b). In addition, the alternative definition for successful TCEs requires two subsequent viral load (VL) measurements below the threshold with at least eight (or 16) weeks in between (c).

4.2.1 Material and Methods

Treatment change episodes

A TCE (Larder et al., 2003) consists of a baseline genotype, a drug combination, and a binary outcome indicating success or failure of the regimen. For our analysis, valid successful or failing TCEs were defined as follows (Figure 4.2): any available genotype is considered as evidence of a failing regimen, because, in general, sequencing can only be performed if the VL exceeds ∼1,000 copies/ml. Successful regimens are defined by inspecting therapies that follow a genotype measurement. When multiple genotypes are available, the most recent sequence sample before the onset of therapy was used. If the VL decreases below 400 copies/ml at least once during the course of the follow-up therapy and if genotyping was performed no earlier than three months before starting the therapy, then the respective treatment is considered a success. This definition of success focuses on initial response. Sustained response is also investigated by using an alternative definition of success that requires a second follow-up VL value below the threshold at least 8 weeks (or 16 weeks) after the first (Figure 4.2 c).

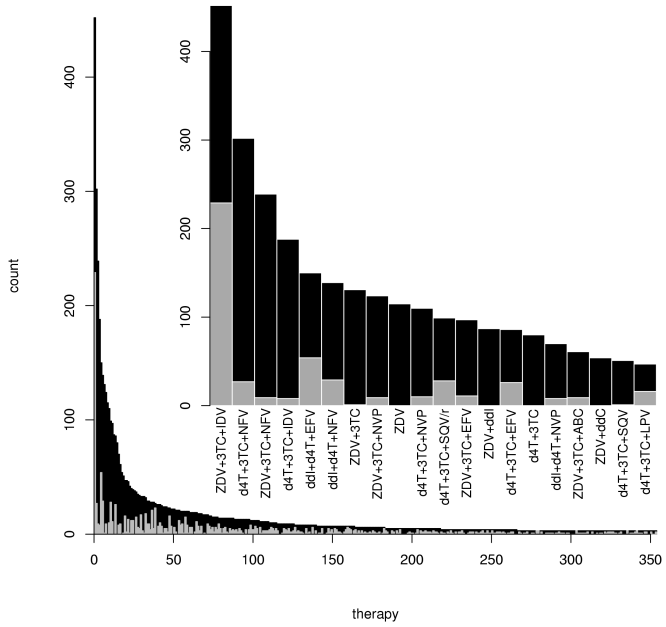


Figure 4.3: The 6,337 analysed TCEs comprise a total of 875 distinct combination therapies. The histogram summarizes all 354 drug combinations that occur at least three times in the dataset. Another 116 combinations appeared only twice and 405 only once. Grey bars indicate successful therapies defined as initial virological responses below 400 copies/ml. The inlet histogram shows the 20 most abundant drug combinations.

Datasets

We analysed data obtained from the Stanford HIV Drug Resistance Database (comprising data from clinical studies ACTG 320, ACTG 364, GART and HAVANA) and from two Northern California clinic populations undergoing genotypic resistance testing at Stanford University. From a total of 25,717 therapies, 10,288 sequences, and 6,706 patients, we extracted 6,337 TCEs, including 4,776 failures and 1,561 successes, according to the definition based on initial response. The median time period between treatment change and the first follow-up VL measurement $<$400 copies/ml is 38 weeks (interquartile range: 16-78 weeks). Figure 4.3 shows the distribution of drug combinations for this dataset (denoted A). The alternative definitions resulted in 1,082 and 900 successes for eight and 16 weeks sustained response, respectively. Because of lack of data on the drug enfuvirtide (at time of the study the only approved EI), we considered only regimens consisting of NRTIs, NNRTIs and PIs.

From dataset A, we generated two special subsets that are balanced with respect to sequences (BS) and with respect to therapies (BT), respectively. The dataset BS contains one successful and one failing regimen for each of 1,364 sequences exactly. Each genotype in this dataset gives rise to two TCEs, defined by the regimens before and after a successful treatment change. Similarly, in the set BT each regimen is paired with two different sequences, giving rise to exactly one successful TCE and one failure. This selection resulted in a total of 2,436 TCEs comprising 321 different regimens. If, for a fixed drug combination,

there were more sequences resulting in failure than in success, then failures have been selected randomly to match the number of successes (or *vice versa*).

Dataset A was split into two groups denoted A1 and A2 according to the clinical centre in which the TCE was observed. Group A1 comprises 816 successes and 2,098 failures from one health care provider, and group A2 includes the remaining 745 successes and 2,678 failures from clinical studies and other hospitals. By using, in turn, one of these subsets for training and the other for testing, we seek to identify possible biases that might be related to the specifics of individual clinics, such as preferred treatment protocols. Datasets BS and BT were split in the same manner.

In addition to the clinical data, we also used *in vitro* data on phenotypic drug resistance. For each drug, 880 sequences with fold-change (FC) determined by a recombinant virus assay (Walter et al., 1999) were available. These data were used to regress phenotype on genotype and to identify resistant states in defining the genetic barrier.

Features

The baseline approach for predicting TCE outcome uses one indicator variable for each resistance-associated mutation and one for each drug. For this approach and approaches based on mutagenetic trees (Section 4.1.2), we considered the resistance mutations presented in Johnson et al. (2005) resulting in 66 binary variables (49 mutation indicators, 17 drug indicators). We refer to this encoding of genotypes and therapies as the indicator representation. All other encodings contain these straightforward covariates and additional, more elaborate, features.

The phenotype representation includes, for each drug in the respective regimen, the predicted FC in susceptibility. These predictions are based on a linear support vector machine that has been trained on the 880 matched genotype-phenotype pairs as described previously (Section 2.5.2).

In the activity representation, the indicator vector is extended by the estimated activity of the drug cocktail against the virus population. For the beam search both parameters, both the breath and depth are set to 10.

The genetic barrier representation also adds both phenotypic and evolutionary information to the indicator representation, and it can be regarded as an advancement over the activity score. More precisely, we used mutagenetic trees, a family of probabilistic graphical models, to estimate the order and rate of occurrence of resistance mutations. Using the Mtreemix software[1] (Beerenwinkel et al., 2005c), for each drug, a mixture model of mutagenetic trees was learned from sequences derived under regimens comprising that drug. The number of trees was selected using the Bayesian information criterion (BIC) (Yin et al., 2006). A validation of mutagenetic tree models in terms of tree stability and goodness of fit has been presented in Beerenwinkel et al. (2005b). Viral escape is approximated by exceeding a predefined level of phenotypic resistance. These levels are defined by the cutoffs listed in Table 4.1 and are exactly the (old) clinical cutoffs for GENO2PHENO (Section 3.2). Unlike the sequence space search for low-activity mutants, the genetic barrier accounts for the fact that not all mutations are equally likely to occur. This is also an advantage over simple counting of resistance mutations, a frequently employed approximation to the

[1] http://mtreemix.bioinf.mpi-sb.mpg.de/

drug	ZDV	ddC	ddI	d4T	3TC	ABC	TDF	NVP	DLV	EFV
cutoff	30.0	2.2	2.4	2.0	15.4	3.4	2.1	9.0	9.7	7.0
trees	5	2	4	4	5	7	3	5	2	4

drug	SQV	IDV	RTV	NFV	APV	LPV	ATV
cutoff	4.5	4.6	2.6	5.8	12.0	10.0	4.2
trees	4	4	4	6	3	5	2

Table 4.1: Resistance cutoff used for each drug and number of trees in the mutagenetic tree mixture model.

genetic barrier.

Finally, the genetic progression score (GPS) involves only evolutionary information that is extracted from the mutagenetic tree models. The GPS of a genotype is defined as the expected waiting time for the mutational pattern to occur. Thus, the GPS also accounts for different probabilities of different mutations, but it does not include any phenotypic information.

Statistical Learning Methods

The five feature sets defined the input to several different machine-learning techniques, which were used to predict treatment response. We selected several standard classification methods (Hastie et al., 2001, pp.79-114), including linear discriminant analysis (LDA), which was used in Beerenwinkel et al. (2003b), together with the activity representation, least-squares regression, linear support vector machines (SVMs), decision trees (C4.5 software[2]), and logistic regression. We also included the more recent method of logistic model trees (LMT), which combine decision trees with logistic regression in the leaves of the tree (Landwehr et al., 2005).

Receiver Operating Characteristic Analysis

We used receiver operating characteristic (ROC) curves to compare the predictive power of classifiers. ROC curves arise from varying a parameter of the classifier that controls the trade-off between sensitivity and specificity. Each point on the ROC curve represents one classifier, and the curve allows for reading off its false positive and true positive rate. For example, the point (0.1, 0.8) represents a classifier that will falsely predict 10% of failing regimens as "success", but correctly detect 80% of the successful regimens. The comparison of ROC curves is preferred to comparing error rates, because it corrects for skewed class distributions (as in dataset A) and controls both sensitivity and specificity of the classifier (Brun-Vézinet et al., 2004). The area under the ROC curve (AUC) is used as a summary performance measure. The trivial classifier that makes random predictions produces a linear ROC curve with an AUC of 0.5. The maximum AUC a classifier can achieve is 1.0 and the larger this value, the better its prediction performance. AUC values were compared using Wilcoxon rank-sum tests. The ROCR software[3] was used for ROC analysis (Sing et al., 2005a).

[2] http://www.rulequest.com
[3] http://bioinf.mpi-sb.mpg.de/projects/rocr/

4.2.2 Results

Table 4.2 shows the performance of all learning techniques in combination with all feature encodings for the complete dataset A using the initial response definition of therapeutic success. Classifier performance was estimated by 10-fold cross-validation and is reported as the AUC. All of the proposed extensions of the indicator representation improved predictive power (Table 4.2, last column). This improvement was observed across all statistical learning methods. The additional features activity score and GPS yield the same predictive power on average. The largest improvement is achieved by the genetic barrier and by phenotype predictions, which both use phenotypic information for prediction. Compared with the feature encoding, the choice of the learning method has only a small effect. On average, the AUC differs by as much as 0.064 (7.7%) between different feature encodings, but only by 0.027 (3.2%) between different learning methods. The AUC of the phenotype representation decreases if combined with decision trees. However, as illustrated in Figure 4.4, the ROC curves reveal that this difference is mainly due to lower true positive rates at very high false positive rates, which might not be relevant for practical purposes. We restrict the further analysis of ROC curves to the LMT learning method, which, on average, outperformed all other techniques (Table 4.2).

In Figure 4.2, the ROC curves for the five different feature encodings used with LMT on dataset A are shown. The indicator representation (black line) is improved significantly by all four additional features ($p<0.0002$). This advancement is most prominent for the genetic barrier that incorporates phenotypic information indirectly ($p=0.0002$), followed by the predicted phenotype, the activity score and the GPS. If we accept a false positive rate of 10%, then the indicator representation will detect 65% of the successful TCEs correctly (represented by the point [0.1, 0.65] on the black line in Figure 4.2), whereas the genetic barrier representation achieves an accuracy of 76.6%. The accuracy can be further increased by accepting more false positives. For example, if 90% of the successes were to be detected, we would have to accept 19.5% false positives for the genetic barrier or the phenotype encoding, and 35.5% for the indicator representation. The differences between the four encodings are even more strongly articulated in the analysis of the two balanced subsets of the complete dataset.

In the dataset BS, each viral genotype is paired with a drug combination that gave rise to a successful TCE and with another TCE that resulted in failure. Thus, the genotype alone does not provide any information on the outcome of these TCEs. As might have been expected, the GPS does not improve the predictive power on this dataset, because this feature is derived only from the genotype (Figure 4.5 a). By contrast, the remaining three encodings, namely activity, genetic barrier, and phenotype, enhance the performance significantly ($p<0.004$). The genetic barrier encoding outperforms the activity encoding, and the phenotype representation provides the best encoding on this dataset. However, the difference between genetic barrier and phenotype is negligible ($p>0.9$).

In the dataset BT, the therapies (rather than the sequence, as in BS) are balanced. For every therapy there exists the same number of genotypes that gave rise to successful and to failing TCEs. Here, the drug combination alone does not provide any information on the outcome of the TCEs. Usage of the GPS on this dataset increases the performance of the indicator representation to the level reached with the activity representation (Figure 4.5 b).

	LDA	LSR	SVM	C4.5	LOGR	LMT	Mean
Indicator	0.825 (0.005)	0.825 (0.005)	0.819 (0.009)	0.845 (0.005)	0.820 (0.007)	0.868 (0.004)	0.834 (0.008)
+ Phenotype	0.912 (0.004)	0.911 (0.004)	0.910 (0.004)	0.839 (0.005)	0.912 (0.002)	0.905 (0.004)	0.898 (0.012)
+ Activity	0.868 (0.005)	0.870 (0.005)	0.864 (0.006)	0.842 (0.007)	0.868 (0.003)	0.898 (0.004)	0.868 (0.007)
+ Genetic barrier	0.891 (0.005)	0.892 (0.003)	0.891 (0.005)	0.875 (0.005)	0.891 (0.002)	0.916 (0.005)	0.893 (0.005)
+ GPS	0.856 (0.006)	0.856 (0.005)	0.864 (0.005)	0.861 (0.005)	0.868 (0.005)	0.899 (0.004)	0.867 (0.007)
Mean	0.870 (0.015)	0.871 (0.015)	0.870 (0.015)	0.852 (0.007)	0.872 (0.015)	0.897 (0.008)	

Table 4.2: Classifier performance on the full dataset A. The table displays, for all combinations of feature encodings (rows) and learning techniques (columns), the resulting area under the receiver operating characteristic (ROC) curve (AUC) and its standard error (in parentheses) computed using 10-fold cross-validation. C4.5, C4.5 software; GPS, genetic progression score; LDA, linear discriminant analysis; LMT, logistic model trees; LOGR, logistic regression; LSR, least-squares regression; SVM, linear support vector machines.

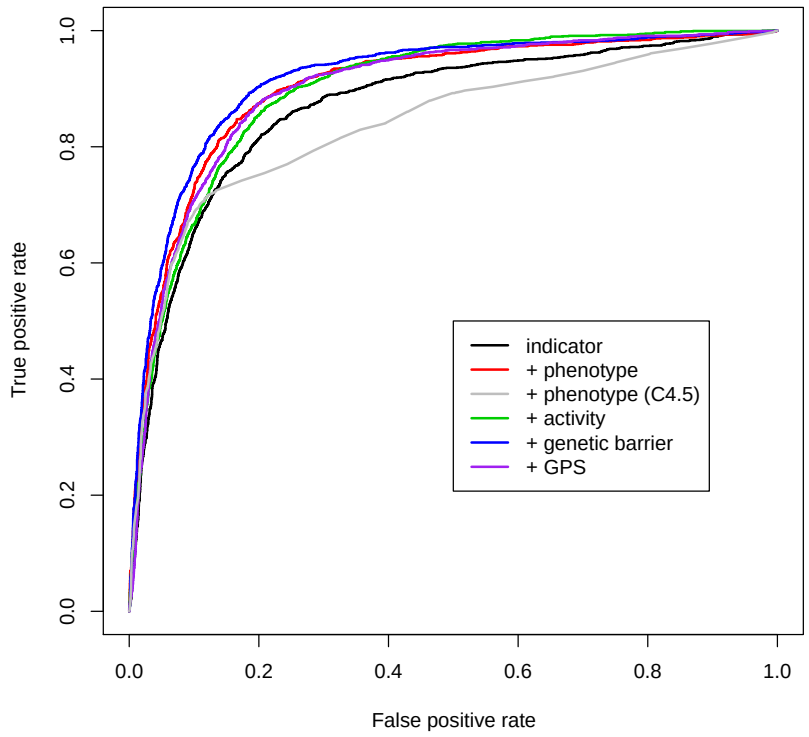
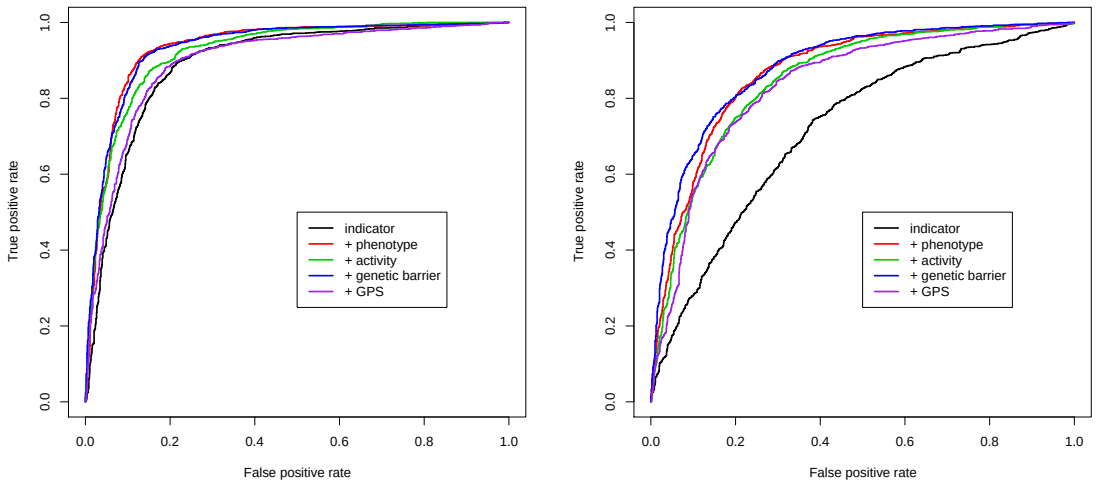


Figure 4.4: ROC curves for the complete dataset A using LMT (unless stated otherwise in the legend) and 10-fold cross-validation. Every feature encoding is represented by a receiver operating characteristic (ROC) curve, namely the baseline indicator representation (indicator), and the following additional features: predicted phenotypes, activity score, genetic barrier, and genetic progression score (GPS). Each point on the curve represents a classifier and allows determining its true positive rate and false positive rate. For example, the point (0.355, 0.9) on the black line represents a logistic model trees (LMT) classification model trained on the plain indicator representation with an expected 35.5% of false positives and 90% of true positives.



(a) Dataset BS (b) Dataset BT

Figure 4.5: ROC curves for dataset BS (a) and dataset BT (b) using LMT and 10-fold cross-validation. BS referes to "balanced with respect to sequences" and likewise BT refers to "balanced with respect to therapies". Fur further details see caption of Figure 4.4.

Similarly to dataset BS, application of the phenotypic and the genetic barrier representation results in maximal performance. Here, the genetic barrier encoding outperforms the phenotype encoding, but the difference does not reach statistical significance (p=0.1655). In the interesting region of low false positive rates (<30%), however, the difference can be substantial. For example, at a false positive rate of 10% the indicator approach recognizes 28% of the successful TCEs correctly, GPS yields a true positive rate of 54.7%, using the activity score increases true positives to 54.5%, the phenotype representation achieves almost 58%, and the genetic barrier recognizes as many as 64.9% of successes correctly. This rate more than doubles the precision of the indicator encoding alone.

In order to investigate possible biases in the estimated models resulting from the specific clinical centres which collected the TCE data, we used the split of dataset A into A1 and A2. The TCEs in A1 originate from a single health care provider, but those in A2 stem from several different clinical studies and hospitals and hence are expected to be less homogeneous. When A1 and A2 are used separately in the same cross-validation procedure described above for the pooled dataset A, then the predictive performance is similar in both cases (AUC of 0.898 for A1 and 0.906 for A2). If A1 (A2) is used for training and A2 (A1) for testing, we estimate an AUC of 0.837 (0.875). The performance loss due to separation of the data by clinics was slightly more pronounced for the balanced dataset BS than for BT (data not shown).

All of the reported results remain qualitatively unchanged when therapeutic success is defined by a more sustained response over at least eight or 16 weeks instead of by initial response. Using the alternative definitions we re-calculated the AUC values for all feature

encodings with LMT on all datasets. None of the derived performance measures showed any significant difference compared with the initial definition.

4.2.3 Discussion

Given the increasing number of possible drug combinations and the genetic diversity of HIV, it is unlikely that simple hand-crafted rules will capture the complex interplay between drug cocktails and mutational patterns that determine response to antiretroviral therapy. Thus, statistical and computational approaches are required for optimal use of the available drugs in each individual patient. Here we have analysed the ability of various statistical learning techniques in combination with different feature encodings to predict therapy outcome from the baseline genotype and the applied drug combination.

The viral genotype is only one of many patient-specific characteristics, such as immune status or genetic predisposition, and it is straightforward to extend our approach to situations where additional parameters are available (see Chapter 5). Nevertheless, the viral genotype has a prominent role among those covariates as it encodes the structure of the target proteins. The challenge in using these genetic data is to predict the evolution of the virus population under drug pressure and to understand the complex relationship between mutational patterns and *in vivo* drug resistance. We have addressed these issues by considering, in addition to drug and mutation indicators, features that make use of phenotypic and evolutionary information. Our computational experiments have identified the predicted phenotype and the genetic barrier to drug resistance as the most beneficial features, significantly boosting the predictive power of all classifiers. The choice of the learning technique had less impact, but LMT consistently showed an advantage over the other methods. In general, the representations that incorporate *in vitro* phenotype predictions yielded better performance than the GPS, a purely evolutionary feature, but both sources of information led to improvements among all tested classifiers and datasets.

The two subanalyses of the complete TCE dataset A showed opposing results. Whereas the performance was increased compared with dataset A for the balanced sequences (BS), it was decreased for the balanced therapies (BT). In the case of the BS data, classifiers need to learn the differences between drug combinations conditioned on the genotype. However, in effect, with this dataset the dependence on genotype is largely masked by the differing application profiles of regimen use accumulated in the dataset. For example, lopinavir appeared in only 38 failing regimens and in 284 successful follow-up therapies. Likewise, the combination of zidovudine and didanosine defined 60 failures, but not a single success. This is because the underlying clinical cohort data reflects the historical approval and use of drugs. For example, the combination of zidovudine and didanosine has been available for some time and many patients have received it until failure. This is easy to learn from the data, but the generalization that this combination always fails, although strongly supported by the data, is certainly wrong and thus misleading when evaluating treatment options containing zidovudine and didanosine. This problem is eliminated with the other subset, BT, where classifiers learn differences between mutational patterns conditioned on the regimen. Because, in learning therapy outcome, we aim to generalize the mutation-drug interactions and not to reconstruct historical drug-use patterns, we regard the performance using dataset BT as a more genuine estimate of our ability to predict treatment response.

We have addressed the concern that the learned models might be biased by the specific sampling of patients by analysing TCEs from different clinics separately. We found a small decrease in predictive performance that was more pronounced when the more homogeneous subset A1 was used for training. This finding highlights the importance of including data from many different clinical centers. The performance loss was greater for the BS dataset than for the BT dataset indicating that the patterns of drug use can vary among clinics. Larger studies that include sufficient data from several clinics need to investigate this source of bias in more detail in the future.

In the present study, we have dichotomized therapy response into "success" and "failure". However, with response defined as the change in VL after a certain time, regression methods can be used for learning in a similar way, and this is unlikely to affect the results presented here. Furthermore, most classifiers predict a score (in fact, often a probability) rather than only the class label. Thus, for a given genotype, they can be used for ranking drug combinations with respect to the expected therapeutic success. In the future, ranking systems could evolve into valuable tools for supporting the complex decision-making process of clinicians. However, such systems will have to provide an interface that allows physicians to incorporate their prior knowledge on a given case, and to constrain the range of possible regimens by explicitly excluding or including specific drugs. In this way, it is possible to spare drugs or drug classes, limit total pill burden, or react to adverse effects, while being able to identify the most promising options from the remaining regimens. The development of improved therapy rankers will crucially depend on their public availability, which allows experts to identify strengths and weaknesses of different approaches. For this reason, we have implemented the prototypical therapy ranker THEO (THErapy Optimizer, Figure 4.6). It is freely accessible for research purposes as part of the geno2pheno web site[4]. In order to produce a ranking of therapies, geno2pheno-THEO (g2p-THEO) applies the classifier trained for predicting therapy response to the sequences of PR and RT provided by the user and a predefined set of therapies (consisting of any combination of either two PIs or two NRTIs plus one NNRTI or one PI). The score produced by the classifier for every combination of drugs is the expected therapeutic success, which is used to rank the considered therapies. In particular, g2p-THEO applies the LMT classifier, which was trained on dataset BT using the genetic barrier encoding as input, to derive the scores needed for the ranking. This selection of feature encoding and statistical learning method was based on the cross-validation results obtained on dataset BT. Figure 4.6 depicts the results of a g2p-THEO analysis of a genotype from a patient failing treatment after first-line therapy with abacavir, lamivudine and efavirenz (for a list of mutations, see caption of Figure 4.6). The first two regimens proposed by g2p-THEO are double PI therapies (saquinavir and lopinavir/ritonavir, and amprenavir and lopinavir/ritonavir) with a predicted probability of virological success of over 70%.

As discussed above, the statistical model used by g2p-THEO has some limitations. Thus, suggestions made by g2p-THEO must be used with care. Moreover, the ranking displayed by g2p-THEO might appear counterintuitive, because it is based on confidence scores provided by the statistical learning method and the single criterion for therapy success is reduction of VL below a threshold. Hence, regimens that are highly active tend to occur among the top-ranked therapies. Although these regimens are most likely to reduce the

[4] http://www.geno2pheno.org

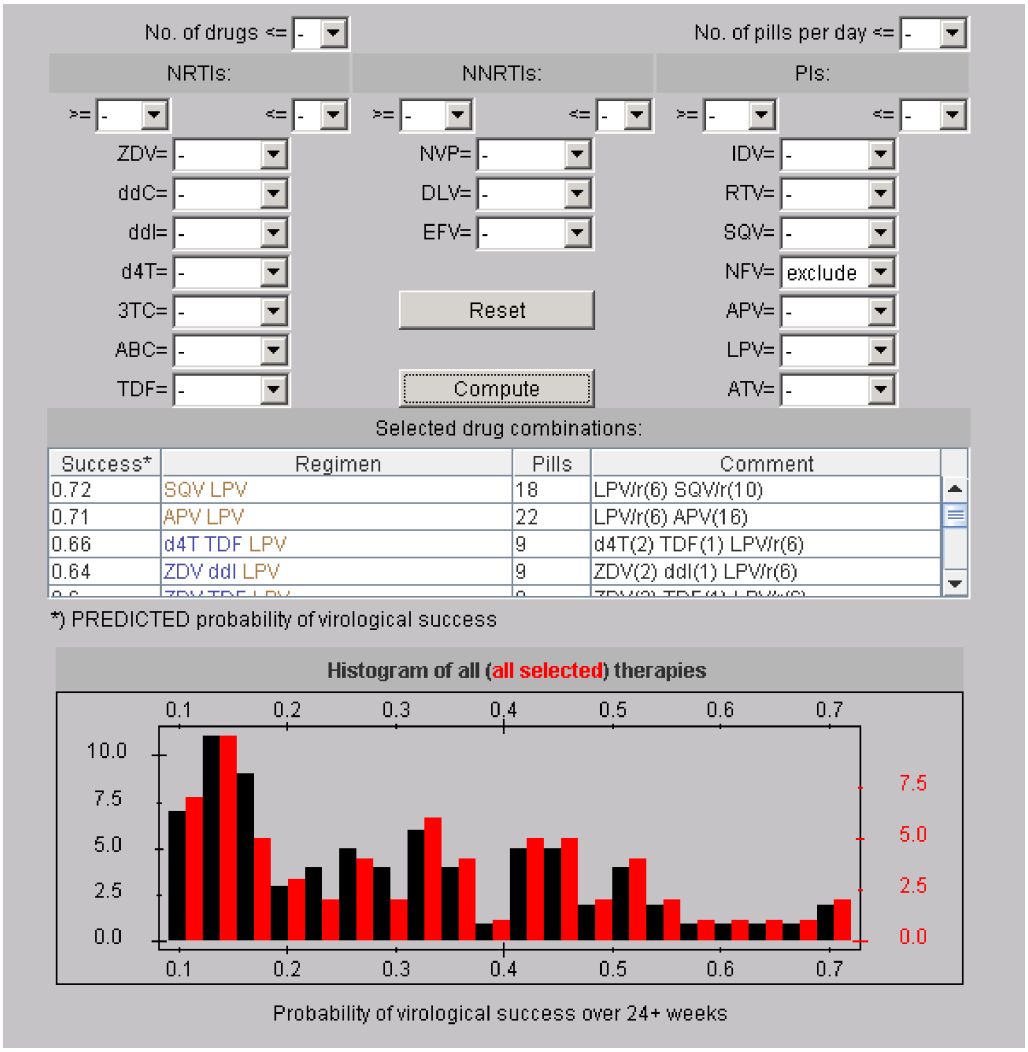


Figure 4.6: The g2p-THEO applet for selecting and evaluating drug combinations. The applet allows for limiting the number of drugs that can be part of a therapy (No. of drugs). Furthermore, the daily burden (No. of pills per day) can be limited. The number of compounds from each drug class can be set (for example, 1≤NRTIS≤3), and the use of single drugs can be enforced (for example, 3TC=include) or excluded. Pushing the "Compute" button ranks all remaining therapies according to the underlying prediction model. In the resulting table, the components of the regimens are listed (Regimen), the number of pills, how many pills of every compound are included in the regimen (Comment), and the score calculated by the model (Success). Additionally, the predictions are compared in a histogram between all selected (according to the constraints specified by the user; red bars) and all possible therapies (black bars). The figure shows the result of a g2p-THEO analysis of a genotype from a patient failing treatment after first line therapy with abacavir, lamivudine and efavirenz. The sequence (subtype B) contained the following mutations (compared with the reference strain HXB2): PR V3I, E35D, S37N, L63P and A71T and RT L74V, K102N, K103N, Y115F, E122K, D123E, D177E, I178V, V179D, M184V, G190A, T200A, R211K, L214F, H221Y, T286A, E297R and S322T.

VL below the threshold necessary for classifying a TCE as a success, less active regimens might be a better choice for preserving future drug options and might therefore be ranked higher by clinicians (Jiang et al., 2003). Thus, we emphasize that the purpose of therapy rankers will always be to support, and not to replace, the complex decision-making process of clinicians.

In future work, further improvement of the genetic barrier representation might be achieved by estimating mutagenetic trees for combinations of drugs instead of single drugs. Such trees would model evolutionary pathways to multi-drug resistance and handle the interplay of drugs within the applied regimen. However, this approach is currently infeasible due to the lack of sufficient data. We also emphasize that the drug-wise computation of the genetic barrier, as employed here, facilitates the incorporation of a new drug, because only one additional mutagenetic tree needs to be learned. Moreover, to a first approximation, this tree can be learned from *in vitro* and clinical study data even prior to approval. The most promising avenues to increased prediction accuracy and improved therapy ranking are via larger datasets, the use of additional parameters (such as $CD4^+$ T cell counts, plasma levels of drugs, viral replication capacity, and host genetic factors), analysis of previous sequences of the viral population that represent important background information, and a profound understanding of viral evolutionary escape from drug pressure.

In the following section, we provide a retrospective validation of g2p-THEO using data from European patient cohorts. These type of validations are used to ensure that interpretation approaches are not overfitted to a specific patient cohort. Furthermore, validations facilitate the comparison between different approaches on independent data, and thereby allow for objectively rating their performance.

4.3 Clinical Validation

This section describes a joint work with Martin Däumer, Niko Beerenwinkel, Yardena Peres, Eugen Schülter, Joachim Büch, Soo-Yon Rhee, Anders Sönnerborg, W. Jeffrey Fessel, Robert W. Shafer, Maurizio Zazzi, Rolf Kaiser, and Thomas Lengauer. The work was published in the *Journal of Infectious Diseases* under the title "Predicting the Response to Combination Antiretroviral Therapy: Retrospective Validation of geno2pheno-THEO on a Large Clinical Database" (Altmann et al., 2009a).

In this section, we present the external validation of g2p-THEO in a dataset containing 7600 treatment-sequence pairs collected in a Europe-wide effort (Aharoni et al., 2007). This validation is of similar nature as the split of dataset A into A1 and A2 (based on the center where the data was collected) presented in the previous section. Virological response was dichotomized, and performance was compared with three state-of-the-art expert-based interpretation tools. In subsequent analyses, various techniques of statistical learning were applied to (1) assess the putative improvement in prediction accuracy incurred by applying models for specific drug combinations and (2) investigate the reliability of g2p-THEO when applied to unseen combinations of compounds - that is, those combinations that are not contained in the training dataset.

4.3.1 Material and Methods

Treatment Change Episodes (TCEs)

The present study used the previously introduced definition of a TCE (Altmann et al., 2007a), on which g2p-THEO is based (Figure 4.2). As opposed to the previous analysis we do not make use of the sustained response. Although current assays have a threshold of sensitivity of 40 or 50 copies, the 400-copy threshold was used to include data obtained by earlier assays.

Datasets

The statistical model applied by g2p-THEO was trained on data obtained from the Stanford HIV Drug Resistance Database (Rhee et al., 2003) (comprising data from clinical studies ACTG 320, ACTG 364, GART, and HAVANA) and from two northern California clinic populations undergoing genotypic resistance testing at Stanford University. From a total of 25,717 therapies, 10,288 sequences, and 6706 patients, 6359 TCEs were extracted (4776 failing and 1583 successful therapies). This dataset is hereafter called "Stanford-California" (A in the previous section). Overrepresentation of certain compounds in failing or successful therapies within the Stanford-California dataset led the statistical learning models to often base their decisions only on the drug combination, irrespective of the genotype. To eliminate this artifact, g2p-THEO was trained not on the full dataset but on a subset that contained the same number of failure- and success-associated genotypes for every drug combination (see Section 4.2 for details). The number of genotypes per drug combination ranges from two to 446 (2478 TCEs in total). Hereafter, this dataset is called "Stanford-CaliforniaBT" (for "balanced therapies"). In some analyses, both the Stanford-California and the Stanford-CaliforniaBT datasets were evaluated again after removal of ZDV+3TC+IDV combination therapy, which was overrepresented because of the inclusion of the large ACTG 320 dataset. For further analysis, a subset of Stanford-California containing only drug combinations with $\geq$ 20 successes and $\geq$ 20 failures was selected. Only six treatments met this requirement (Table 4.3); hereafter, this dataset is called "Stanford-California6".

Using the same definition, an independent TCE dataset (EuResistDB) was extracted from the EuResist integrated database (version 2007/05/29), comprising data from Germany (Arevir) (Roomp et al., 2006), Italy (ARCA) (De Luca et al., 2006), and Sweden (Karolinska Institute). For more details on the EuResist database see Chapter 5. From a total of 58,195 therapies, 19,258 sequences, and 16,999 patients, we obtained 7603 TCEs (6217 failing and 1386 successful therapies). For further analysis, we generated the subset EuResistDB6, comprising the same six drug combinations as in Stanford-California6 (Table 4.3). Because the EuResist integrated database includes antiretroviral treatments started from 1990 to 2007, the main analysis was repeated on the TCE subset derived from treatments started after 31 December 2000, to minimize the contribution of obsolete therapy records featuring e.g. non-boosted protease inhibitors.

	Stanford-California6			EuResistDB6		
Regimen	Failure	Success	Total	Failure	Success	Total
ZDV+3TC+IDV	223	229	452	108	5	113
d4T+3TC+SQV/r	71	28	99	27	2	29
ddI+d4T+EFV	96	54	150	132	25	157
d4T+3TC+EFV	60	27	87	114	16	130
ddI+d4T+NFV	110	28	138	131	22	153
d4T+3TC+NFV	275	28	303	209	16	225
All regimens	835	394	1229	721	86	807

Table 4.3: Distribution of failing and successful treatment change episodes for the 6 most common regimens in the Stanford-California and EuResistDB datasets (Stanford-California6 and EuResistDB6 subsets).

	Score		
	1.0	0.5	0.0
ANRS	susceptible	possible resistance	resistant
Rega	susceptible	intermediate	resistant
Stanford HIVdb	susceptible, potential low-level resistance	low-level resistance, intermediate resistance	high-level resistance

Table 4.4: Mapping of the classification given by different expert-based interpretation systems to continuous values.

Interpretation systems

ANRS (Meynard et al., 2002, version 2006/07), Rega (Van Laethem et al., 2002, version 7.1.1), and Stanford HIVdb (Rhee et al., 2003, version 4.3.0) are expert-based interpretation methods. These algorithms apply carefully handcrafted interpretation rules or tables derived by expert panels from the analysis of available *in vitro* and *in vivo* resistance data. In addition, some rules of the ANRS algorithm were derived from the statistical association between baseline genotypic data and virological response. The classification generated by the interpretation systems was normalized into a score by mapping the verbal classification to a score according to Table 4.4. Thus, the rating numerically represents the activity of a drug against the virus on a scale ranging from 0.0 (inactive) to 1.0 (fully active). Individual scores for NNRTIs and for boosted PIs computed using the Rega algorithm were converted to 1.0, 0.25, and 0.0 and to 1.5, 0.75, and 0.0, respectively, as indicated by the algorithm developers. The treatment score or genotypic susceptibility score (GSS) (De Luca et al., 2004) was then defined as the sum of single-drug scores for the compounds included in the regimen.

g2p-THEO is a data-driven interpretation system that directly computes a rating for a combination therapy. This value can be interpreted as the probability of the viral load being reduced to below the limit of detection during the course of therapy. g2p-THEO represents the HIV-1 genotype by 49 indicator variables, each of them indicating the presence (1) or absence (0) of a resistance mutation (Johnson et al., 2005) (the complete

list: for protease, 10F/ I/R/V, 16E, 20M/R/I, 24I, 30N, 32I, 33F/I/V, 36I/L/V, 46I/L, 47V/A, 48V, 50L/V, 53L, 54L/V/M/T/A/S, 60E, 62V, 63P, 71V/T/I/L, 73S/ A/C/T, 77I, 82A/F/T/S, 84V, 88D/S, 90M, and 93L; for reverse transcriptase, 41L, 44D, 62V, 65R, 67N, 70R, 74V, 75T/M/S/A/I, 77L, 100I, 103N, 106A/M, 108I, 115F, 116Y, 118I, 151M, 181C/I, 184V/I, 188C/L/H, 190S/A, 210W, 215F/Y, and 219Q/E). Similarly, treatment is encoded using 17 indicator variables, each representing the presence (1) or absence (0) of a compound in the regimen (for the list of considered compounds, see Table 4.1). In addition, viral evolution during the course of therapy is represented by the genetic barrier to drug resistance (Beerenwinkel et al., 2005a) for all drugs in the regimen. The genetic barrier is the probability that the virus will remain susceptible under drug pressure, given as a numerical value between 0.0 (no genetic barrier [i.e., the virus is expected to become resistant]) and 1.0 (insurmountable genetic barrier [i.e., the virus is expected to remain susceptible]). Together with the indicator variables, the genetic barrier (one value per drug) is used as input to the logistic model tree (LMT) (Landwehr et al., 2005) applied by g2p-THEO to compute a score for a combination therapy.

Receiver Operating Characteristic (ROC) Curves

ROC curves depict classifier performance by giving a true-positive rate (TPR; percentage of correctly predicted successes) for every false-positive rate (FPR; percentage of failing therapies [i.e., with a genotype obtained during treatment] that were predicted to be successful [i.e., to decrease viral load to <400 copies/ml]). The area under the ROC curve (AUC) summarizes the performance and is a convenient measure for comparing scoring systems without the need to provide a particular cutoff (Brun-Vézinet et al., 2004). Briefly, the AUC is a value between 0.0 and 1.0 corresponding to the probability that a randomly selected success receives a higher score than a randomly selected failure (Fawcett, 2006). The ROCR software (version 1.0-2) (Sing et al., 2005a) was used for the ROC analysis.

Comparative Analysis

The statistical model applied by g2p-THEO was used to predict the outcome of the genotype-therapy pairs in the external EuResistDB dataset. The performance was compared with that of the three expert-based interpretation tools: ANRS, Rega, and Stanford HIVdb. The EuResistDB dataset was used as an independent test set. One hundred bootstrap replicates of the EuResistDB dataset were used for computing standard deviations.

Stability

To further analyze the robustness of the approach, the prediction of response to drug combinations that were not present in the training data was simulated by a variant of cross-validation on the training data. In standard cross-validation, the available data are split randomly into n equally sized non-overlapping subsets. Then, $n-1$ pooled subsets are used as a training set, and the remaining subset is used as a test set to compute the performance of the model. This procedure is repeated n times. Hence, every subset is used as a test set once. To simulate the prediction of unseen drug combinations, the splits were not carried out randomly, and every subset contained only TCEs with the same

Analysis, setting	Kernel	ϵ	C
Stability analysis			
All	Linear	0.1	1
Regimen-specific models			
Regimen specific	Linear	0.1	C optimizing AUC in 5-fold CV
All	Linear	0.1	4 (optimized AUC in 10-fold CV)

Table 4.5: Support vector machine settings. C is the cost factor, CV refers to cross-validation.

drug combination. Results derived by this "therapy-fold" cross-validation were compared with results of standard cross-validation, with the number of folds equal to the number of different drug combinations. A substantial loss in performance (e.g., a loss of 0.1 in AUC) for the therapy-fold cross-validation compared with the normal protocol indicates unstable behavior with respect to unseen drug combinations, because it indicates the requirement to include examples of the drug combination that should be predicted in the training data for maintaining the performance. The experiment was repeated on the special subset Stanford-CaliforniaBT and on the complete EuResistDB dataset. Because of limited computational resources, the LMT applied by g2p-THEO was replaced by the faster linear support vector machines (SVMs) (Chang and Lin, 2001), with similar predictive performance (see e.g. Table 4.2). SVM settings are listed in Table 4.5.

Regimen-specific Models

For some drug combinations, many samples were available. Models trained exclusively on TCEs for one drug combination are expected to predict response to that regimen more accurately than models trained on all available TCEs. The Stanford-California6 subset contained sufficient data for training 6 regimen-specific models and for measuring their performance. The performance of the individual models was assessed by 10 repetitions of five-fold cross-validation on data comprising only one drug combination. The performance of the full model was assessed by 10 repetitions of a variant of five-fold cross-validation in which all TCEs with other drug combinations were added to the four subsets forming the training data. SVMs with a linear kernel were used as a statistical learning method. SVM settings are listed in Table 4.5. Performance in an independent test set was assessed by predicting the outcome of TCEs in EuResistDB6 with the full model and the six regimen-specific models (10 repetitions). Results were compared with those obtained using g2p-THEO and the three expert-based interpretation tools.

4.3.2 Results

Comparative Analysis

Figure 4.7 depicts the ROC curves for the three expert-based interpretation tools and g2p-THEO. The curves for Stanford HIVdb, ANRS, and Rega did not differ substantially, resulting in comparable TPRs and FPRs for every GSS cutoff. In contrast, in the FPR range from 0% to 40%, the curve for g2p-THEO was distinctly located above all the other

curves. For higher FPRs, all curves proceeded with similar slopes. The AUC value for g2p-THEO was significantly larger than those for the expert-based approaches ($p < 0.001$; paired Wilcoxon test). ROC curves allow for a detailed analysis of specific points on the curve. For example, at a FPR of 20% (close to a GSS cutoff of 2.5 for Rega and ANRS and of 2.0 for Stanford HIVdb), Stanford HIVdb, ANRS, and Rega yielded TPRs of 44.2% (±3.2%), 47.8% (±2.7%), and 46.5% (±2.2%), respectively. In contrast, at the same FPR g2p-THEO achieved a TPR of 64.0% (±1.6%). On the other hand, at a false-negative rate (FNR) of 10% (close to a GSS cutoff of 1.0 for all systems), Stanford HIVdb, ANRS, and Rega yielded true-negative rates of 54.1% (±2.0%), 51.0% (±2.1%), and 55.9% (±0.8%), respectively, compared with 54.7% (±1.8%) for g2p-THEO. Restriction of the EuResistDB dataset to therapies started after 31 December 2000 led to an even more pronounced difference in AUC between g2p-THEO (0.824±0.006) and the expert-based approaches (for Stanford HIVdb, 0.728±0.008; for ANRS, 0.733±0.008; for Rega, 0.754±0.007).

Stability

Table 4.6 summarizes the results of the stability analysis. The standard cross-validation performance on the Stanford-California dataset was in line with previously computed results (Table 4.2). For both Stanford-California datasets containing the highly prevalent ZDV+3TC+IDV regimen, the therapy-fold cross-validation yielded slightly worse results than the standard protocol (difference in AUC of ~ 0.03). In contrast, in the EuResistDB dataset and both Stanford-California datasets without ZDV+3TC+IDV, no loss in performance was observed.

Regimen-specific Models

Table 4.7 shows the mean AUC and standard deviation for the regimen-specific models and the full model for the six most common drug combinations in the Stanford-California dataset. The results for the combination of d4T+3TC+SQV/r were the worst for both models, and there was no benefit in using the regimen-specific model. However, for the remaining five drug combinations, the benefits of regimen-specific models ranged from 0.017 to 0.048 and reached statistical significance. Within the EuResistDB6 dataset, an insufficient number of successful TCEs was available for ZDV+3TC+IDV and d4T+3TC+SQV/r (Table 4.3). The full model outperformed the regimen-specific model only for d4T+3TC+EFV. For the remaining three drug combinations, the benefit of the regimen-specific model was more pronounced than in the cross-validation setting and ranged from 0.046 to 0.301 for d4T+3TC+NFV, a combination for which the full model actually failed to make useful predictions. The LMTs in g2p-THEO also outperformed the regimen-specific model for d4T+3TC+EFV. In the remaining three cases, the benefit of the regimen-specific models ranged from 0.011 to 0.082. All regimen-specific models outperformed the expert-based methods.

Figure 4.8 depicts the ROC curves for the regimen-specific models (g2p-THEO-SVM-RS), the full model (g2p-THEO-SVM), g2p-THEO, and the expert-based methods applied to the EuResistDB6 dataset. The full model performed better than the expert-based methods in the area below a FPR of 32% but worse in the remaining region. As in the previous ROC plot (Figure 4.7), g2p-THEO performed better than the expert-based

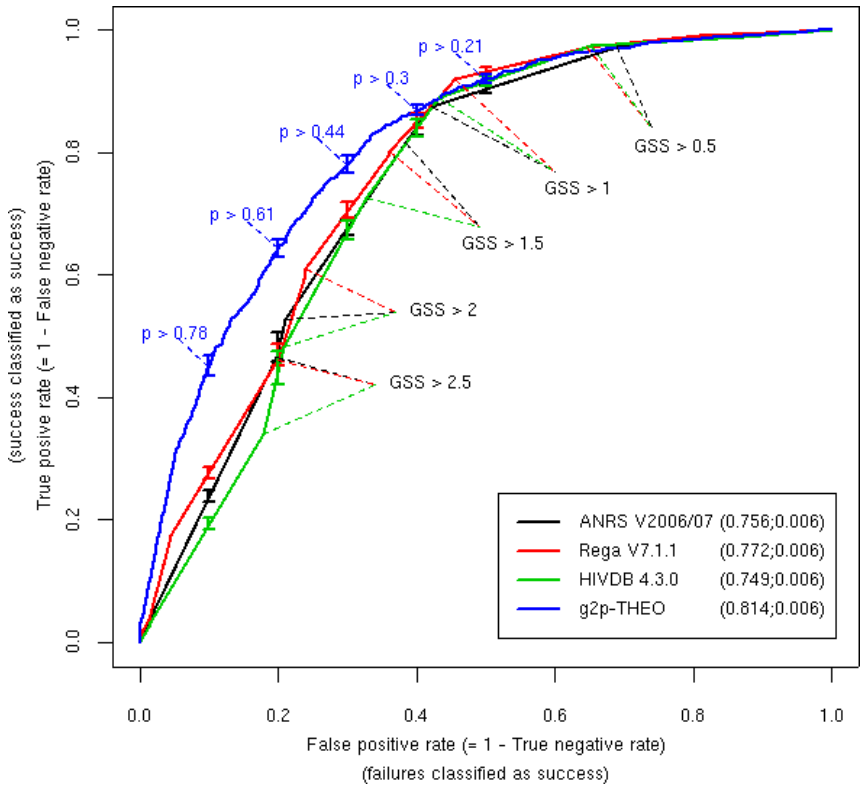


Figure 4.7: Receiver operating characteristic (ROC) curves for the EuResistDB dataset. Every method is represented by a single ROC curve, namely Stanford HIVdb, ANRS, Rega, and geno2pheno-THEO (g2p-THEO). Each point on the curve represents a classifier with a different cutoff and allows the true-positive rate (TPR) and false-positive rate (FPR) for that cutoff to be determined. Whiskers indicate the standard deviations of the TPRs at a specific FPR. The genotypic susceptibility score (GSS) and predicted success probability (p) cutoffs leading to specific TPR and FPR values are indicated within the plot for expert-based approaches and g2p-THEO, respectively. For each method, the area under the ROC curve and its standard deviation are given parenthetically in the box.

interpretation tools in the area below a 50% FPR and performed as well in the remaining region. However, the regimen-specific models outperformed the other methods over the whole range of FPRs. More specifically, they yielded a TPR of 58.4% at a FPR of 20%, compared with 39.8% for the expert-based methods and 53.6% for g2p-THEO.

4.3.3 Discussion

Validation of HIV genotype interpretation systems is a crucial step in translating computer-based methods into clinically effective treatment decision support tools. In the present study, using an external dataset of $\approx$ 7600 TCEs extracted from the EuResist integrated database, the recently developed g2p-THEO system was shown to outperform the three most widely used expert-based interpretation systems. Although the EuResist dataset included many obsolete therapies because of its long observation period, the same results

Dataset	Drug combinations, no.	AUC	
		Standard cross-validation	Therapy-fold cross-validation
Stanford-California			
With ZDV+3TC+IDV	876	0.901	0.870
Without ZDV+3TC+IDV	875	0.898	0.895
Stanford-CaliforniaBT			
With ZDV+3TC+IDV	323	0.838	0.812
Without ZDV+3TC+IDV	322	0.812	0.813
EuResistDB	712	0.847	0.844

Table 4.6: Area under the receiver operating characteristic curve (AUC) values obtained by standard cross-validation and therapy-fold cross-validation. The rows of the table correspond to the five datasets, with both Stanford-California datasets studied with and without the ZDV+3TC+IDV drug combination, in light of the overrepresentation of this regimen caused by the ACTG 320 data. (Stanford-CaliforniaBT is the "balanced therapies" subset.) Note that computation of the AUC requires positive and negative samples, but because of the nature of therapy-fold cross-validation, it could not be ensured that positive and negative samples were present in every subset of the cross-validation. Thus, it was not possible to compute fold-wise AUC values or their standard deviations and statistical significance.

were confirmed when therapies started before 1 January 2001 were removed. The g2p-THEO system was more accurate than ANRS, Stanford HIVdb, and Rega by 16.2% - 19.8% in the detection of therapeutic success (20% FPR). However, all of the systems were comparable in detecting treatment failure at a 10% FNR. This finding suggests that expert-based systems are better suited to detect the failure of therapy than to detect success, probably because their original purpose was to detect resistance to individual drugs. However, whether a treatment will most likely be successful is exactly the response a user expects from a decision support tool. Current expert-based approaches are indeed evolving into clinically oriented tools aimed at building effective combination regimens. Computing a regimen GSS by simple summation of the individual drug scores derived by expert-based systems fails by definition to weight both different drug potencies and drug interaction effects. However, such an unweighted GSS is still commonly used (Maggiolo et al., 2007; Cozzi-Lepri et al., 2007) in the absence of any agreed-upon standard for a weighted GSS. Notably, the latest Rega algorithm has introduced arbitrary drug weights in an attempt to account for the expected increased potency of ritonavir-boosted PIs and a lack of intermediate NNRTI activity.

The superior performance of g2p-THEO may have derived from two factors. First, the calculated genetic barrier provides useful information by estimating the probability of viral evolution under drug pressure (Altmann et al., 2007a). Second, the training process assigns weights to all drugs. Hence, during the decision making process, drugs are not treated equally. As shown by Altmann et al. (2007b, 2009b), this can significantly improve the performance of genotype interpretation tools. On the other hand, g2p-THEO is currently

	Stanford-California6			EuResistDB6, AUC					
Regimen	Regimen-specific model, AUC	Full model, AUC	p	Regimen-specific model	Full model	g2p-THEO	Stanford HIVdb	ANRS	Rega
ZDV+3TC+IDV	0.965 (0.005)	0.928 (0.002)	<0.001	0.596 (0.038)	0.570	0.610	0.656	0.692	0.755
D4T+3TC+SQV/r	0.747 (0.035)	0.740 (0.022)	0.256	0.420 (0.012)	0.815	0.444	0.407	0.463	0.426
ddI+d4T+EFV	0.988 (0.006)	0.950 (0.002)	<0.001	0.925 (0.003)	0.879	0.914	0.803	0.798	0.777
d4T+3TC+EFV	0.924 (0.028)	0.876 (0.015)	0.003	0.727 (0.016)	0.771	0.769	0.693	0.724	0.708
ddI+d4T+NFV	0.974 (0.005)	0.957 (0.005)	<0.001	0.821 (0.008)	0.709	0.739	0.665	0.671	0.684
d4T+3TC+NFV	0.946 (0.010)	0.915 (0.010)	<0.001	0.829 (0.016)	0.528	0.782	0.782	0.781	0.784
Mean	0.924	0.895		0.720	0.712	0.710	0.668	0.688	0.689
Full dataset				0.810 (0.005)	0.733	0.781	0.707	0.718	0.707

Table 4.7: Area under the receiver operating characteristic curve (AUC) values for cross-validation and training test set-up. Rows correspond to the six regimens in the two datasets. The columns for the Stanford-California6 data show the AUC values derived by 10 repetitions of five-fold cross-validation for the regimen-specific models and the full model using all available training data (SDs are in parentheses); p-values for the comparison of the performances of the two models were obtained by the Wilcoxon rank-sum test. The columns for the EuResistDB6 dataset show the AUC values for the regimen-specific models (standard deviations are in parentheses), with the full model using all Stanford-California data, the original geno2pheno-THEO (g2p-THEO) predictions, and the three expert-based approaches (Stanford HIVdb, ANRS, and Rega).

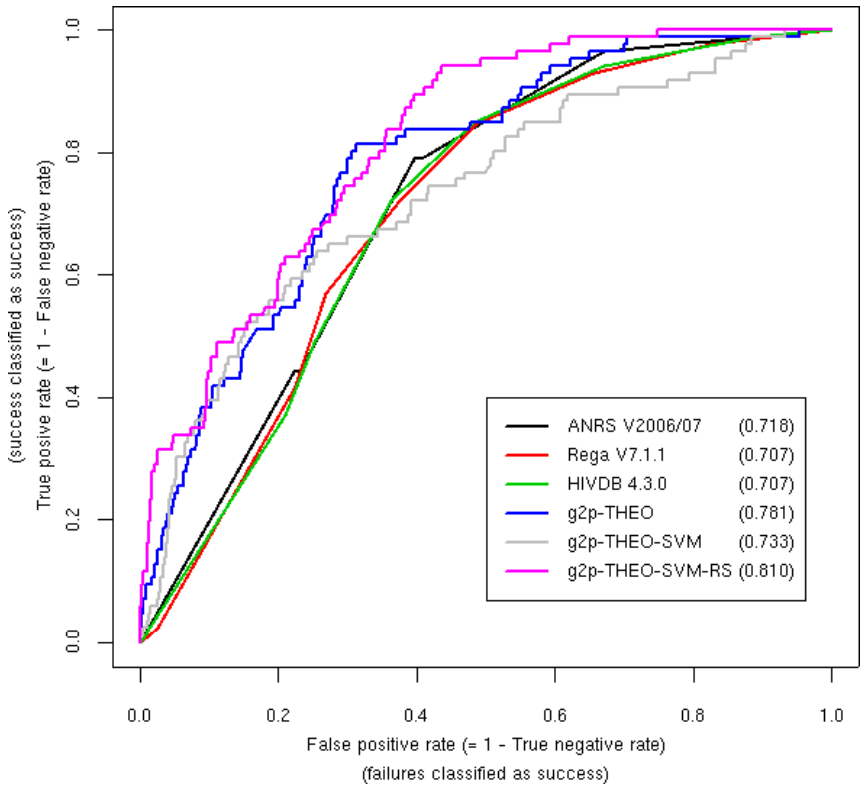


Figure 4.8: Receiver operating characteristic (ROC) curves for regimen-specific models applied to the EuResistDB6 dataset. Every method is represented by a ROC curve for a subset of the EuResistDB data comprising six different treatments. In addition to the methods depicted in Figure 4.7, ROC curves are shown for g2p-THEO-SVM (g2p-THEO in which logistic model trees were replaced by SVMs) and g2p-THEO-SVM-RS (g2p-THEO using regimen-specific SVMs). For each method, the area under the ROC curve is given parenthetically in the box.

limited to a set of well-established resistance mutations, which might explain its inability to improve the detection of failing regimens.

In the computation of a score for a drug combination, the robustness of the tool with respect to unseen drug combinations is an important issue. In both Stanford-California datasets, a slight decrease in the AUC was observed in predictions for unseen drug combinations. Overrepresentation of ZDV+3TC+IDV therapy due to inclusion of the ACTG 320 clinical trial in the datasets was identified as a possible confounder of this analysis, because no decrease in performance with unseen drug combinations was observed after removal of TCEs containing the ZDV+3TC+IDV combination. Thus, our stability analysis indicated that g2p-THEO returns reliable scores for unobserved drug combinations. The major reason for the preservation of performance is that the prediction is based on a linear model. Indeed, during the learning process of the linear model, contributions of every single covariate to the outcome are computed. This also holds for drugs in a regimen, because the impact has to be distributed among these drugs. In the end, observed and

unobserved drug combinations are both composed of observed compounds. However, this property is also a potential disadvantage of linear models. Specifically, if the prediction is based on a linear model, the synergistic effects between drugs or mutations cannot be represented unless a large number of covariates are introduced to explicitly model these effects. In contrast, in regimen-specific models all mutations are evaluated in the context of the same drug combination, thus rendering the explicit modeling of interactions unnecessary. In the Stanford-California6 dataset, the regimen-specific models for all drug combinations exhibited increased performance compared with the full model, even though the full model had access to many more training samples. This finding was confirmed in independent data. However, the benefit of regimen-specific models decreased when they were compared with the full statistical model for g2p-THEO. This can be explained by the fact that g2p-THEO applies LMTs that directly train multiple linear models on distinct subsets of the training data. Unfortunately, only a few outdated regimens gave rise to enough training data for regimen-specific models. However, a promising approach has been recently proposed (Bickel et al., 2008), one that pools data on "similar" regimens to overcome this limitation in generating regimen-specific models.

A major issue with any fully data-driven system is the inability to generate predictions for newly licensed compounds because of the delayed availability of sufficient training data. This drawback can be addressed only by multicenter efforts and cooperation between drug companies and regulatory bodies for immediate release of clinical trial data. However, because an optimized background regimen is recommended for the effective use of any new compound, interpretation systems are still relevant for choosing the backbone drugs, particularly in heavily experienced patients. Expert-based systems can complement data-driven systems for predicting the activity of novel drugs until sufficiently large genotype-response datasets are available. An attempt to combine expert scores for novel drugs with a data-driven approach will be presented in Section 5.8.

Data-driven systems need a large amount of data for training, so observational cohort data are often used. These provide a valuable source for assessing the impact that drug resistance has on the response to treatment but typically lack other relevant information, including adherence levels and pharmacokinetics data. Weighting for these factors is expected to help us develop better systems aimed at building effective regimens. It must also be noted that modern and future antiretroviral treatment strategies are expected to limit the development of drug resistance by providing increased potency and convenience, perhaps making treatment toxicity issues relatively more relevant than resistance over time. However, drug resistance and cross-resistance remain issues for a substantial proportion of patients harboring viral populations that display a complex mutational pattern because of multiple treatment failures. In addition, toxicity is also a major contributor to the selection of drug resistance through decreased adherence. Developed as a clinically oriented tool, g2p-THEO allows the user to exclude specific drugs for toxicity issues and provides a total pill count for each regimen. Although no data-driven system is meant to replace a comprehensive patient evaluation by an expert HIV specialist, validated tools such as g2p-THEO can provide an appropriate support to most care-givers of HIV-infected patients.

5 EuResist: Uniting Data for Fighting HIV

In the previous chapter we demonstrated that features derived from the viral genotype are helpful in predicting response to combination therapy. The predictive performance of statistical learning models, however, depends to a great deal on the amount of available training data. In the case of geno2pheno approximately 1000 genotype-phenotype pairs were sufficient for building well performing models for inferring phenotypic drug resistance against a singe drug from genotype. The predicted resistance information is based on information derived *in vitro*. Clinicians, however, are mainly interested in knowing whether a drug-cocktail will work *in vivo* or not. The fact that anti-HIV drugs can be used in hundreds of combinations calls for massive amounts of training data for systems such as g2p-THEO that directly predict response to combination therapy. The HIV Resistance Database Initiative (RDI) was the first international initiative aiming at the integration of multiple databases from different centers for collecting sufficient data to train such statistical models (Larder et al., 2002). Since its foundation in 2002, the RDI database grew from 3,500 to approximately 30,000 patients in 2007[1]. However, till now RDI failed to provide a freely accessible web tool that predicts response to combination treatments.

This chapter describes our work within the EU project: EuResist. EuResist, like RDI, aims at the collection of data and a the development of a free web accessible tool that predicts response to combination treatments. The current version of the database holds approximately records from 33,000 different patients. Unlike RDI, EuResist made the prediction system freely available already in 2007. This chapter begins with a background on the EuResist project (Section 5.1). In Sections 5.2 and 5.3 we introduce the individual EuResist prediction engines and the combination strategies, respectively. The final web service is described in Section 5.5. Sections 5.6 and 5.7 investigate how changes in the definition to treatment response affect the prediction performance. Furthermore, in Section 5.8 an approach for overcoming a serious limitation of all prediction engines aiming at inferring response to combination therapies, namely the update ability with respect to novel compounds, is studied. Finally, section 5.9 explores the possibility of predicting response to antiretroviral therapy based on the treatment history alone.

5.1 Project Background

The EuResist project aimed at developing a European integrated system for clinical management of antiretroviral drug resistance. The system should provide the clinicians with a prediction of response to antiretroviral treatment in HIV patients, thus helping the clinicians to choose the best drugs and drug combinations for any given HIV genetic variant. To this end, a massive European integrated dataset was created, linking some of the largest locally existing resistance databases.

[1] source: http://www.hivrdi.org

The project involved participants from eight institutions. Three institutions acted as data providers and consultants concerning virological knowledge and clinical aspects of HIV treatment, namely the University of Siena, the Karolinska Institute in Stockholm, and the University of Cologne. Four participating institutions were involved in modeling the prediction engines: IBM Israel, KFKI Research Institute for Particle and Nuclear Physics Hungarian Academy of Sciences, Informa S.r.l., and the Max Planck Institute for Informatics. A division of IBM Israel was also responsible for the integration of the multiple data sources into a unified database, and Informa S.r.l. was concerned with the overall project management and the development of the web front end for the prediction engine. The Kingston University dealt with matters of data quality of the integrated database.

Like other HIV resistance tools, the response models should be able to provide a prediction even when the genotype is the only available information. The data collected in the resistance databases, however, enable exploiting patient related data: for instance the patient's treatment history, the current immune status, and demographic data. To allow for both: the use of the genotype only and the exploitation of additional data, the modeling was restricted to genotype and treatment information only (minimal feature set) and opened for all available information (maximal feature set). In operation mode, the model chosen for the prediction depends on the information provided by the user.

A considerable effort within the project was the definition of the standard datum. Briefly, the standard datum is the equivalent to the TCEs in the previous chapter and defines which data from the database are considered for learning and what has to be learned. For instance, if one is interested in virological response to an anti-HIV treatment around three months after onset of the regimen (short-term response), then one has to extract treatments from the database where a viral load measurement is available around three months of treatment. The standard datum definition can also be a source of error or overfitting. For instance, in the work by Larder et al. (2007) the time between onset of the treatment and the viral load measure is used as a feature. In their study, one treatment with different viral load measurements gave rise to at most three training points for the same statistical model. The setup ensured that the patients in training and test set formed a distinct set. However, the high correlation of the cases, due to multiple samples from the same treatment, in the small test set of 50 instances leads very likely to an overestimation of model performance. Another important issue, which the standard datum definition has to address, concerns the baseline measurements, i.e. what is the maximal allowed time span between measurement of a biomarker and start of the therapy so that the value can be considered a baseline measurement. For example, a genotype obtained about one year before the treatment start is clearly too outdated to serve as a baseline genotype. On the other hand, a genotype that was generated a few weeks before treatment start is clearly suitable. In general, the stricter a standard datum definition, the fewer data are available for the statistical learning step, thus the definition has to find a good balance between strictness and resulting training data.

The standard datum definition applied within the EuResist project considers values that were obtained at most three months before start of the regimen as baseline values – only if there was no intermediate treatment of at least two weeks length between the measurement and start of the treatment to be considered. With this definition, EuResist follows the recommendations of The Forum for Collaborative HIV Research. Furthermore,

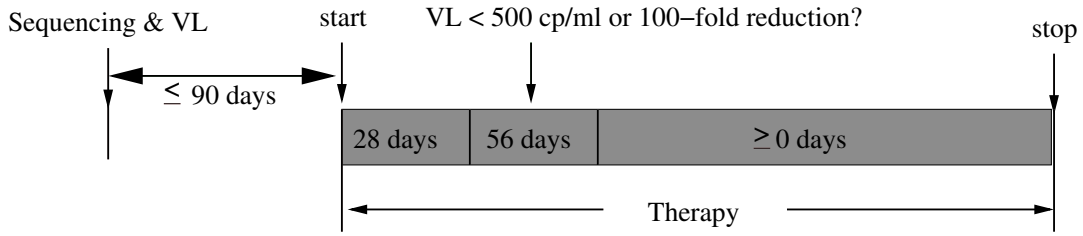


Figure 5.1: Standard datum. A treatment change is considered a valid training/testing instance if the baseline measures (viral genotype and VL) were obtained at most 90 days before start of the new treatment and given that there was no other treatment lasting longer than 14 days between time point of measurement and start of treatment. A treatment is considered successful if the VL at 8 weeks (±4) is below the limit of detection (here: 500 copies of viral RNA per ml blood serum) or constitutes a 100-fold reduction compared to the baseline value.

the EuResist standard datum focuses on initial response around eight (four to 12) weeks of treatment; one treatment can only give rise to only one training instance since only the closest measurement to the eight weeks time point is considered. Thus, the definition facilitated a regression approach, i.e. predicting the change in viral load between the baseline measurement and the follow-up measurement. In addition, like in the development of g2p-THEO, a classification approach was investigated within the project. To this end, treatment response was dichotomized into success and failure based on the reduction of viral load. Precisely, if the viral load could be reduced below the limit of detection, i.e. 500 copies of viral RNA per ml blood serum, then the treatment was considered a success. Some patients, though, start with extremely high viral load values and a reduction below the limit of detection is hard to achieve within the short time frame. Thus, alternatively, a success can be achieved by a 100-fold reduction of the viral load compared to the baseline value. Figure 5.1 depicts a schematic overview of the standard datum definition. Modifications of this standard datum definition that focus on sustained response (VL at 24±8 weeks) are studied later in this chapter.

Figure 5.2 depicts the growth of the EuResist Integrated database (EIDB) over the period of the project. The amount of patients and HIV sequences almost doubled from initially about 17,000 to now 34,000 entries. Likewise, the number of stored treatments nearly tripled from 35,000 to 98,000. The observed increase of data is not only a result of the increased size of the initial three databases, but also a direct consequence of the addition of new databases. In November 2007 the Luxembourg database joined the EuResist integrated database, and in October 2008 three additional databases started their collaboration with EuResist: IRSICAIXA Foundation (Spain), Catholic University of Leuven (Belgium), and Univeristy of Brescia (Italy). More databases are expected to join the EuResist effort. It strikes that the number of usable standard datum instances is by far smaller than the number of sequences in the database. This discrepancy is a result of the strict requirements posed by the standard datum definition. For instance a sequence has to be available at most 90 days prior to treatment start, thus not all available sequences in the database gave rise to a usable learning instance. However, the collection effort increased

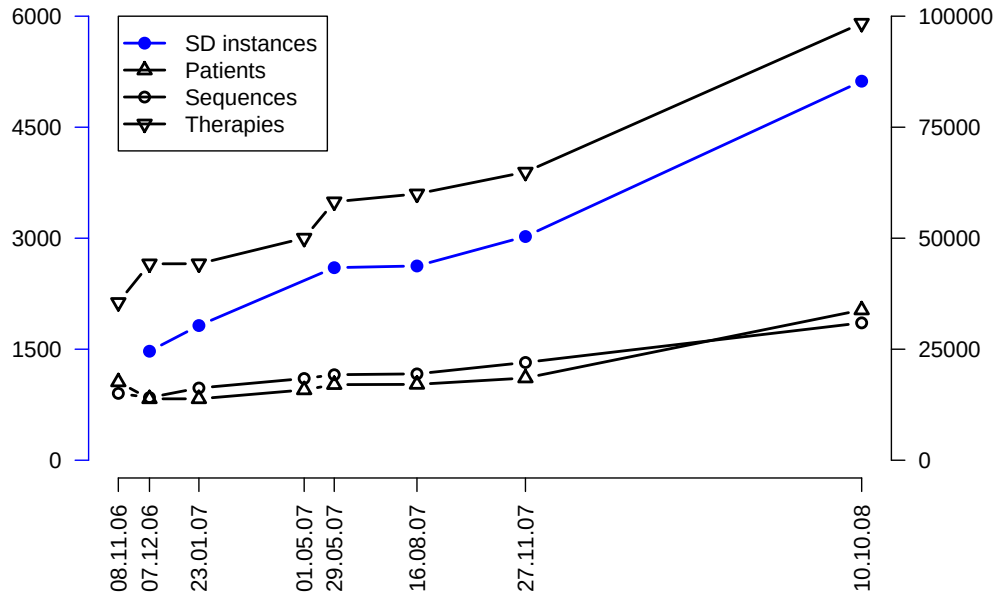


Figure 5.2: Growth of the EuResist integrated database (EIDB). The graph depicts the increase in patients, viral sequences, and recorded treatments in the database of the time of the project (scale on the right). The blue line (scale on the left) corresponds to the increase of standard datum (SD) instances available for training and testing the prediction engine.

the available training instances from about 1500 to over 5000. Since not all experiments have been carried out with the most recent release of the EuResist integrated database, we provide for every computer experiment the release date of the database that was used, and thus allowing to put the results in the correct context.

5.2 EuResist Prediction Engines

Within the EuResist project four institutions developed statistical models for predicting the response to combination antiretroviral therapy. Three of the prediction engines were selected for the final system to be implemented as a web service. Our contribution, the Evolutionary (EV) engine, was based on the g2p-THEO software described in Chapter 4. The two other contributions the Generative Discriminative (GD) engine and the Mixed Effects (ME) engine originate from the Machine Learning Group of the IBM Research Laboratory in Haifa, Israel and from the Department of Computer Science and Automation of the University of Roma TRE located in Rome, Italy, respectively. In the following sections we provide a brief summary for the GD and ME engines. Furthermore, we specify the differences between the original g2p-THEO software and our EV engine.

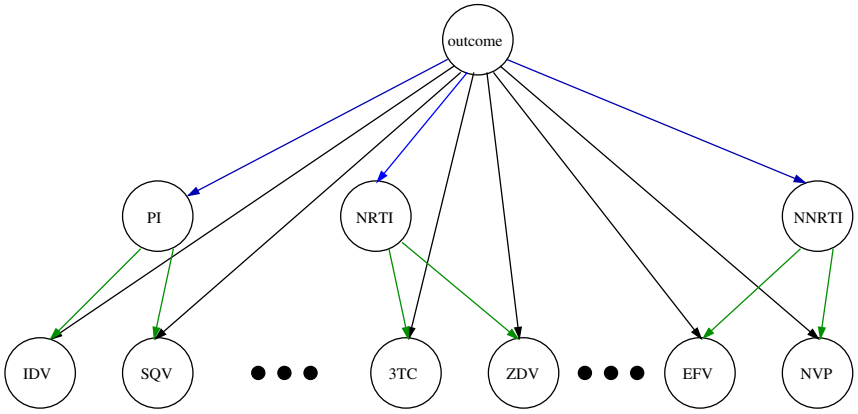


Figure 5.3: The Bayesian network used in the GD engine. Variables indicating use of individual drugs are connected to the outcome (black arrows) and to indicators for their drugs class or in the case of the maximal feature set: historic use of a drug from the same class (green arrows). The variables in the middle layer representing the (historic) use of a drug class are also connected to the outcome variable (blue arrows).

5.2.1 Generative Discriminative Engine

The Generative Discriminative (GD) engine applies generative models to derive additional features for the classification using logistic regression. When inspecting Figure 5.2 it strikes that only a small percentage of therapies in the database (Table 5.1) have an associated genotype and are therefore suitable for training a classifier, which is supposed to receive sequence information. However, a much larger fraction of the therapies can be labeled as success or failure on the basis of the baseline and follow-up VL measures alone, since for the labeling no viral genotype is required. The GD engine thus trains a Bayesian network on about 20,000 therapies (with and without associated genotype). The network is organized in three layers and uses an indicator for the outcome of the therapy, indicators for individual drugs, and indicators for drug classes (Figure 5.3). This generative model is used to compute a probability of therapy success on the basis of the drug combination alone. This probability is used as an additional feature for the classification by a logistic regression model: the discriminative step of the approach. Furthermore, indicators for individual drugs and single mutations are input for the logistic regression.

Indicators representing a drug class are replaced with a count of the number of previously used drugs from that class when working with the maximal feature set. In this way information about past treatments is incorporated. In addition to features from the minimal set, the maximal feature set comprises indicators for mutations in previously observed genotypes, the number of past treatment lines, and the VL measure at baseline. Correlation between single mutations and the outcome of the therapy was used to select relevant mutations for the model. A detailed description on the network's setup and the selected mutations can be found in Rosen-Zvi et al. (2008).

5.2.2 Mixed Effects Engine

The Mixed Effects (ME) engine explores the benefit of including second- and third-order variable interactions. Basically, since modern regimens combine multiple drugs, binary indicators representing usage of two or three specific drugs in the same regimen are introduced. Further indicators represent the occurrence of two specific mutations in the viral genome for modeling interaction effects between them. Moreover, interactions between specific single drugs and single mutations or pairs of drugs and single mutations are represented by additional covariates. In addition to the terms modeling the mixed effects, other clinical measures (VL and $CD4^+$ T cell counts), demographic information (risk factor, country of infection, risk, sex, ...), covariates based on previous treatments (indicators for previous use of drugs and drug classes), and the predicted viral subtype are used as additional covariates.

The large number of features (due to the mixed effects) requires a strong effort in feature selection. Thus, multiple feature selection methods were used for generating candidate feature sets. Filters and embedded methods, i.e. methods that are intrinsically tied to a statistical learning method, were applied sequentially: (i) univariable filters, such as χ^2 with rank-sum test and correlation-based feature selection (Hall, 1999), were applied to reduce the set of candidate features; (ii) embedded multivariate methods, such as ridge shrinkage (Le Cessie and Van Houwelingen, 1992) and the Akaike information criterion (AIC) selection (Akaike, 1974) were used to eliminate correlated features and to assess the significance of features with respect to the outcome in multivariate analysis. In multiple 10-fold cross-validation runs on the training data the performance of the resulting feature sets were compared with a t-statistic (adjusted for sample overlap and multiple testing). The approach leading to the best feature set was applied on all training samples to generate the final model. Unlike the GD engine, the ME engine employs one logistic regression model that only uses the maximal feature set. Missing variables, e.g. when working with the minimal feature set, are replaced by the mean (or mode) of that variable in the training data.

5.2.3 Evolutionary Engine

As stated before, one major obstacle in HIV-1 treatment is the development of resistance mutations. The Evolutionary (EV) engine uses derived evolutionary features to model the virus's expected escape path from drug therapy. The representation of viral evolution is based on mutagenetic trees. Unlike in g2p-THEO the mixture of mutagenetic trees comprises only two components, the noise component and one tree that was estimated from the data. A further difference to the original approach concerns the way of how mutational patterns that define complete drug resistance against an individual compound are identified. This information is crucial for computing the genetic barrier to drug resistance. The original approach, which is briefly described in Section 4.1.2, uses the available genotype-phenotype pairs. Since there are only about 1,000 such pairs, it is very likely that mutational patterns that can occur are not observed within the small sample. Here, we used geno2pheno for predicting the drug resistance phenotypes for approximately 16,000 viruses stored in the EIDB. Based on this much larger sample the procedure of identifying resistant patterns was carried out as initially explained. The modified approach resulted in

genetic barrier values that were more predictive for response to treatment than the values obtained with the original protocol (increase of 0.02 and 0.027 in AUC and correlation, respectively, in a 10-fold cross-validation on the training set). The genetic barrier to drug resistance was provided together with other features, like indicators for individual drugs in the treatment and indicators for IAS mutations in the genotype as in the prototype g2p-THEO. In the maximal feature set, indicators for previous use of a drug and the baseline VL measure extend this list. As a general improvement over the original g2p-THEO prototype, interactions up to second order between indicator variables were considered as well. This was achieved by introducing further indicator variables that explicitly model these interactions.

The resulting large number of covariates required the implementation of a feature selection step as part of the training process of the EV engine. The chosen feature selection approach was based on SVMs with a linear kernel. The approach works in three steps: (i) optimization of the misclassification cost parameter C in a 10-fold cross-validation setting to maximize the area under the ROC curve (AUC); (ii) generation of 25 different SVMs by five repetitions of five-fold cross-validation using the optimized cost parameter; (iii) computation of the z-score for every feature. All features with a mean z-score larger than 2 were selected for the final model based on logistic regression.

5.3 Combining Classifiers

In order to provide the end-user of the prediction system with a single outcome for a request, instead of multiple results generated by the different engines, the three classifications have to be combined to one final decision. In principle, there are two approaches for combining classifiers, namely classifier fusion and classifier selection. In classifier fusion, complete information on the feature space is provided to every individual system and all outputs from the systems have to be combined; in classifier selection, every system is an expert in a specific domain of the feature space and the local expert alone decides for the output of the ensemble. However, the individual classifiers described above were designed to be global experts, thus only classifier fusion methods were explored.

Methods for classifier fusion can operate on class labels or continuous values (e.g. support, posterior probability) provided by every classifier. The methods range from simple non-trainable combiners like the majority vote, to very sophisticated methods that require an additional training step. In order to find the best combination method we compared several approaches ranging from simple methods to more sophisticated ones. All results were compared to a combination that has access to an oracle telling which classifier is correct. Intuitively, the predictive performance of this oracle represents the upper bound on the performance that can be achieved by combining the classifiers. The following subsections briefly introduce the combination methods considered.

Non-trainable Combiners

As mentioned above, there are a number of simple methods to combine outputs from multiple classifiers. The most intuitive one is a simple majority vote, whereby every individual classifier computes a class label (in this case success or failure) and the label

that receives the most votes is the output of the ensemble. One can also combine the posterior probability of observing a successful treatment as computed by the logistic regression. This continuous measure can be combined using further simple functions: *mean* returns the mean probability of success by the three classifiers (Kittler et al., 1998); *min* yields the minimal probability of success (a pessimistic measure); *max* results in the maximal predicted probability of success (an optimistic measure); *median* returns the median probability.

Meta-classifiers

The use of meta-classifiers is a more sophisticated method of classifier combination, which uses the individual classifiers' outputs as input for a second classification step. This allows for weighting the output of the individual classifiers. In this work we applied quadratic discriminant analysis (QDA), logistic regression, decision trees, and naïve Bayes (operating on class labels) as meta-classifiers.

Decision Templates and Dempster-Shafer

The decision template combiner was introduced by Kuncheva et al. (2001). The main idea is to remember the most typical output of the individual classifiers for each class, termed decision template. Given the predictions for a new instance by all classifiers, the class with the closest (according to some distance measure) decision template is the output of the ensemble.

Let $\vec{x}$ be an instance, then $DP_{\vec{x}}$ is the associated decision profile. The decision profile for an instance contains the support (e.g. the posterior probability) by every classifier for every class. Thus, $DP_{\vec{x}}$ is an $I \times J$-matrix, where I and J correspond to the number of classifiers and classes, respectively. The decision template combiner is trained by computing the decision templates DT for every class. The DT for the class j is simply the mean of all decision profiles for instances $\vec{x}$ belonging that class. Hence,

$$DT_j = \frac{1}{N_j} \sum_{\vec{x} \in \omega_j} DP_{\vec{x}},\ \forall j \in \{1, \ldots, J\},$$

where N_j is the number of elements in class ω_j. For a new sample, the corresponding decision profile is computed and compared with the decision templates for all classes using a suitable distance measure. The class with the closest decision template is the output of the ensemble. Thus, the decision template combiner is a nearest-mean classifier that operates on decision space rather than on feature space. We used the squared Euclidean distance to compute the support for every class:

$$\mu_j = 1 - \frac{1}{JI} \sum_{j'=1}^{J} \sum_{i=1}^{I} [DT_j(j', j) - DP_{\vec{x}}(j', i)]^2,$$

where $DT_j(j', i)$ is the (j', i)-th entry in DT_j. Decision templates were reported to outperform other combiners, e.g. Kuncheva et al. (2001) and Kuncheva (2002).

Decision templates can also be used to compute a combination that is motivated by the evidence combination of the Dempster-Shafer theory. Instead of computing the similarity

between a decision template and the decision profile, a more complex computation is carried out as described in detail in Rogova (1994). We refer to these two methods as *Decision Templates* and *Dempster-Shafer*, respectively.

Clusters in Decision Space

Regions in decision space where the classifiers disagree on the outcome are of particular interest in classifier combination. Therefore, we propose the following method that finds clusters in decision space and learns separate logistic regression models for every cluster for fusing the individual predictions. Let s_i be the posterior probability of observing a successful treatment predicted by classifier i. Then we express the (dis)agreement between two classifiers by computing:

$$\alpha_{ij} = \begin{cases} 0 & \text{classifiers } i \text{ and } j \text{ agree on the label,} \\ s_i - s_j & \text{else,} \end{cases}$$

for all i and j where $i < j$. Thus, in case of disagreement between two classifiers, the computed value expresses the magnitude of disagreement. These agreements are computed for all instances of the training set and used as input to a k-medoid clustering (Rousseeuw and Kaufman, 1990). For all resulting k clusters, an individual logistic regression is trained on all instances associated with the cluster using the s_i as input. The idea is that in clusters where e.g. classifier 1 and 2 agree, and classifier 3 tends to predict lower success probabilities the logistic regression can either increase or decrease the influence of classifier 3, depending on how often predictions by that classifier are correct or incorrect, respectively.

When a new instance has to be classified then first the agreement between the classifiers is computed for locating the closest cluster. In a second step the logistic regression associated with that cluster is used to calculate the output of the ensemble. The number of clusters k, the only parameter of this method, is optimized in a 10-fold cross-validation. The approach is motivated by the behavior knowledge space (BKS) method (Huang and Suen, 1995), which uses a look-up table to generate the output of the ensemble. However, the BKS method is known to easily overtrain, and does not work with continuous predictions.

Local Accuracy-based Weighting

Woods et al. (1997) propose a method that uses one k-nearest-neighbor (knn) classifier for every individual classifier to assess the local accuracy of that classifier given the input features. The final output is then solely given by the most reliable classifier of the ensemble. Since the three classifiers in this setting are trained to be global experts, we applied the proposed method to compute the reliability estimate for each classifier given the features of an instance. In contrast to the method proposed by Woods et al. (1997), the output is a weighted mean based on these reliability estimates.

In order to use a knn classifier as a reliability estimator the labels from the original instances are replaced by indicators of whether the classifier in question was correct on that instance or not. With the replaced labels and the originally used features the knn classifier reports the fraction of correctly classified samples in the neighborhood of the query instance. This fraction can be used as a local reliability estimator. The output by

the ensemble is then defined by a weighted mean:

$$\bar{s} = \frac{\sum_i r_i s_i}{\sum_i r_i}$$

where s_i and r_i are the posterior probability of observing a successful treatment and the local accuracy for classifier i, respectively. For simplicity, only Euclidean distance was used in the knn classifier, the number of neighbors k was optimized in a 10-fold cross-validation setting.

Combining Classifiers on the Feature Level

As described in Section 5.2, every individual classifier uses a different feature set, specifically, different derived features, but the same statistical learning method. Thus, a further combination strategy is the use of all features selected for the individual classifier as input to a single logistic regression rather than computing a consensus of the individual classifiers' predictions.

5.4 Predictive Performance of the EuResist Prediction Engine

5.4.1 Data

About 3,000 instances in the EIDB (version from 2007/11/27) met the requirements of the standard datum definition and can therefore be used as learning instances. From this complete set, 10% of the data were randomly set aside and used as an independent test set (Table 5.1). The split of the training data in 10 equally sized folds was fixed, allowing for 10-fold cross-validation of the individual classifiers. The same 10 folds were used for a 10-fold cross-validation of the combination approaches. Classification performance was measured in terms of accuracy (i.e. the fraction of correctly classified examples) and the area under the receiver operating characteristics curve (AUC). Briefly, the AUC is a value between 0.0 and 1.0 and corresponds to the probability that a randomly selected positive example receives a higher score than a randomly selected negative example (Fawcett, 2006). Thus, a higher AUC corresponds to a better performance.

5.4.2 Results

Results for the individual classifiers using the minimal and maximal feature set are summarized in Table 5.2. The use of the extended feature set significantly improved the performance of the GD and EV engine with respect to the AUC ($p \approx 0.002$ for both using a paired Wilcoxon test). With respect to accuracy only the improvement observed by the EV engine reached statistical significance ($p = 0.007$). Remarkably, replacement of all missing additional features in the case of the ME engine when working with the minimal feature set did not result in a significant loss in performance ($p = 0.313$ and $p = 0.312$ with respect to AUC and accuracy, respectively).

Correlation among classifiers

The performances of the individual classifiers were very similar. Pearson's correlation coefficient (r) indicated that the predicted probability of success for the training instances

	Patients	Sequences	VL measurements	Therapies	Successes	Failures
EIDB	18,467	22,006	240,795	64,864	-	-
Labeled Therapies	8,233	3,492	40,498	20,249	13,935	6,314
Training Set	2,389	2,722	5,444	2,722	1,822	900
Test Set	297	301	602	301	202	99

Table 5.1: Summary of the EuResist Integrated Database (version from 2007/11/27) and training and test set. The table displays the number of patients, sequences, VL measurements, and therapies for the complete EuResist Integrated Database (EIDB) and the set of therapies that could be labeled with the standard datum definition. 469 of the sequences associated with all labeled therapies belong to historic genotypes and are not directly associated with a therapy change. Moreover, detailed information on training set and test set (comprising labeled therapies with an associated sequence each) is given.

Engine	minimal feature set				maximal feature set			
	AUC		Accuracy		AUC		Accuracy	
	Train	Test	Train	Test	Train	Test	Train	Test
GD	0.747 (0.027)	0.744	0.745 (0.024)	0.724	0.768 (0.025)	0.760	0.752 (0.028)	0.757
ME	0.758 (0.019)	0.745	0.748 (0.031)	0.757	0.762 (0.021)	0.742	0.754 (0.030)	0.757
EV	0.766 (0.030)	0.768	0.754 (0.031)	0.748	0.789 (0.023)	0.804	0.780 (0.032)	0.751

Table 5.2: Results for the individual classifiers on training set and test set. The table displays the performance, measured in AUC and Accuracy, achieved by the individual classifiers on the training set (using 10-fold cross-validation; standard deviation in brackets) and the test set using different feature sets.

using the minimal (maximal) feature set were highly correlated (i.e. close to 1.0): GD-ME 0.812 (0.868); GD-EV 0.797 (0.786); ME-EV 0.774 (0.768). In fact, the three classifiers agreed on the same label in 80.4% (81.7%) of the cases using the minimal (maximal) feature set. Notably, agreement of the three classifiers on the wrong label occurred more frequently in instances labeled as failure than in instances labeled as success (39% vs. 4% and 37% vs. 4% using the minimal and maximal feature set, respectively; both $p < 2.2 \times 10^{-16}$ with Fisher's exact test).

This behavior led to further investigation of the instances labeled as failures in the EIDB. Indeed, 145 of 350 failing instances, which were predicted to be a success by all three engines, achieve a VL below 500 copies per ml once during the course of the therapy. However, this reduction was not achieved during the time interval that was used in the applied definition of therapy success. Among the remaining 550 failing cases this occurred only 100 times. Using a Fisher's exact test this difference was highly significant ($p = 4.8 \times 10^{-14}$). These results were qualitatively the same when using the maximal feature set.

Results of combination methods

Table 5.3 summarizes the results achieved by combining the individual classifiers, and Figure 5.4 depicts the improvement in AUC on training and test set of the combination meth-

Method	minimal feature set				maximal feature set			
	AUC		Accuracy		AUC		Accuracy	
	Train	Test	Train	Test	Train	Test	Train	Test
SB	0.766 (0.030)	0.768	0.754 (0.031)	0.748	0.789 (0.023)	0.804	0.780 (0.032)	0.751
Oracle	0.914 (0.015)	0.911	0.842 (0.025)	0.844	0.917 (0.013)	0.920	0.850 (0.022)	0.860
Min	0.771 (0.020)	0.765	0.746 (0.027)	0.761	0.792 (0.021)	0.793	0.760 (0.030)	0.764
Max	0.760 (0.023)	0.765	0.742 (0.030)	0.731	0.779 (0.021)	0.779	0.757 (0.030)	0.741
Median	0.773 (0.020)	0.766	0.759 (0.027)	0.766	0.789 (0.029)	0.786	0.768 (0.029)	0.761
Mean	0.777 (0.020)	0.772	0.760 (0.024)	0.744	0.794 (0.019)	0.793	0.780 (0.028)	0.781
Majority	0.683 (0.023)	0.660	0.759 (0.027)	0.738	0.697 (0.027)	0.683	0.768 (0.029)	0.761
QDA	0.771 (0.020)	0.763	0.755 (0.031)	0.738	0.790 (0.022)	0.794	0.769 (0.027)	0.764
LR	0.778 (0.021)	0.774	0.762 (0.028)	0.744	0.798 (0.020)	0.805	0.781 (0.030)	0.771
DTrees	0.718 (0.044)	0.741	0.748 (0.032)	0.757	0.722 (0.033)	0.678	0.777 (0.032)	0.757
NB	0.732 (0.027)	0.740	0.759 (0.027)	0.738	0.752 (0.028)	0.753	0.768 (0.029)	0.761
DT	0.777 (0.021)	0.774	0.755 (0.027)	0.754	0.796 (0.019)	0.797	0.766 (0.026)	0.767
D-S	0.777 (0.021)	0.772	0.755 (0.024)	0.754	0.796 (0.019)	0.796	0.767 (0.026)	0.764
Cluster	0.775 (0.019)	0.773	0.758 (0.029)	0.741	0.797 (0.018)	0.800	0.783 (0.028)	0.784
LA	0.777 (0.020)	0.771	0.761 (0.025)	0.741	0.795 (0.019)	0.791	0.781 (0.029)	0.777
Feature	0.750 (0.026)	0.747	0.745 (0.029)	0.751	0.786 (0.021)	0.779	0.780 (0.029)	0.767

Table 5.3: Results for the combined classifiers on training and test set. The table summarizes the results achieved by the different combination approaches on the training set (10-fold cross-validation; standard deviation in brackets) and the test set. The reference methods are Single Best (SB) and Oracle, the non-trainable combiners are named according to their function, the meta-classifiers according to the statistical learning methods, logistic regression, decision trees, and Naïve Bayes are abbreviated by LR, DTrees, and NB, respectively. Decision Templates (DT), Dampster-Shafer (D-S), Clustering (Cluster) and Local Accuracy (LA) are the methods described in detail in Section 5.3. Feature is the combination on the feature level.

ods compared to the single best and single worst classifier, respectively. Most combination methods improved performance significantly over the single worst classifier. However, only the oracle could establish a significant improvement over the single best classifier. Overall, performances of the combination approaches were quite similar. Of note, the pessimistic min combiner yielded better performance than the optimistic max combiner. Among the non-trainable approaches tested, the mean combiner yielded the best performance. The logistic regression was the best performing meta-classifier. In fact, the logistic regression can be regarded as a weighted mean, with the weights depending on the individual classifier's accuracy, and the correlation between classifiers. Moreover, using all features of the individual classifiers as input to a single logistic regression did not improve over the single best approach.

Figure 5.5 shows the learning curves for the three individual classifiers, the mean combiner, and the combiner on the feature level. The curves depict the mean AUC (after 10 repetitions) on the test set achieved with varying sizes of the training set (25, 50, 100, 200, 400, 800, 1600, 2722). In every repetition the training samples were randomly selected from the complete set of training instances. The mean combiner appeared to learn faster and significantly outperformed the single best engine with a training set size of 200 samples

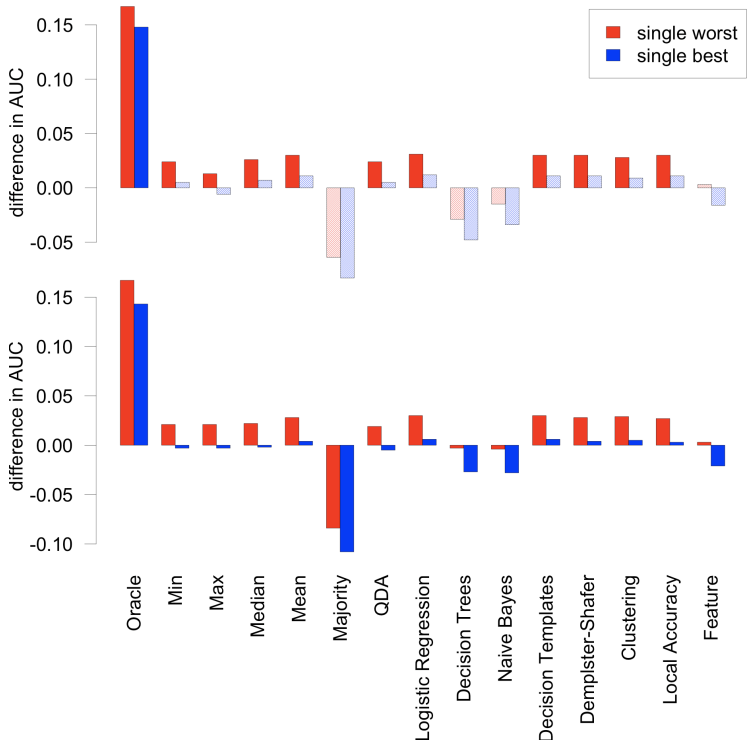


Figure 5.4: Improvement in AUC of combination methods compared to the single best and single worst classifiers. The figure displays the improvement in AUC of all combination methods over the single best (blue bars) and single worst (red bars) classifiers on the training set (upper panel) and the test set (lower panel). Significance of the improvement on the training set was computed with a one-sided paired Wilcoxon test. Solidly colored bars indicate significant improvements (at a 0.05 p-value threshold), as opposed to lightly shaded bars for insignificant improvements. On the test set no p-values could be computed.

($p \approx 0.010$ with a paired one-sided Wilcoxon test). The improvement remained significant up to a training set size of 1,600 samples ($p \approx 0.002$). The combination on feature level was significantly worse ($p \approx 0.002$) than the worst single approach for all training set sizes (except for complete set).

Finally, Figure 5.6 depicts the ROC curves of the three individual engines and the combined engine (mean) using both feature sets on the test set. The results are compared to a Stanford HIVdb based GSS prediction (for details see e.g. Section 4.3). The Stanford HIVdb service did not support the rarely used drug ddC. Consequently, the four instances in the test set containing ddC were excluded. Moreover, Stanford HIVdb did not provide prediction results for two sequences. Thus, the reported AUC values (in parentheses in the figure legend) differ slightly from the values reported in Tables 5.2 and 5.3 as only 295 of the 301 test instances were used for generating the figure.

Impact of Ambiguous Failures

In order to further study the impact of ambiguous failures (i.e. instances labeled as failure but achieving a VL below 500 cp per ml once during the course of treatment) on the

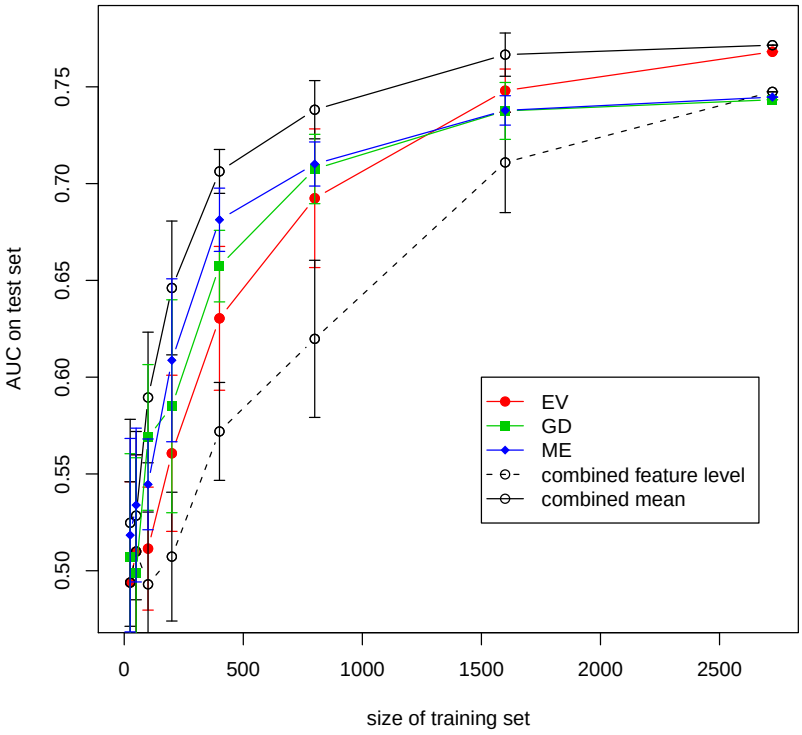


Figure 5.5: Learning curves for the individual classifiers, the mean combiner, and the combination on feature level. The figure shows the development of the mean AUC on the test set depending on the amount of available training data for the individual classifiers, the mean combiner, and the combination on the feature level using the minimal feature set. Error bars indicate the standard deviation on 10 repetitions.

performance of the individual classifiers and the combination by mean or on the feature level, they were removed from the training set, the test set, or both sets. After removal the classifiers were retrained and tested on the resulting new training and test set, respectively. The results in Table 5.4 suggest that training with the ambiguous failures does not impact the classification performance (columns "none" vs. "only train", and columns "only test" vs. "both"). However, the ambiguous cases have great impact on the assessed performance. Removal of these cases increases the resulting AUC by 0.05.

However, there might still be an influence of these ambiguous failures on the performance of the trainable combination methods. For verification we removed these cases whenever performance measures were computed (also in 10-fold cross-validation) and trained a selection of the combination methods on the complete training data and on the cleaned (i.e. without ambiguous failures) training data. The results in Table 5.5 suggest that the trainable combination methods were not biased by the ambiguous failures.

A possibility for circumventing the need for dichotomization of virological response is the prediction of change in VL between the baseline value and the measurement taken

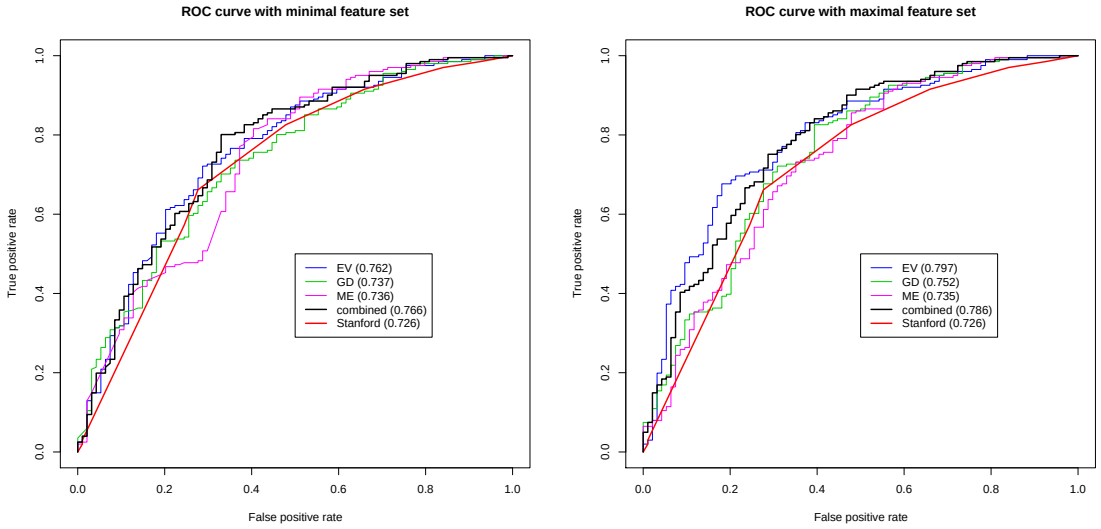


Figure 5.6: ROC curves for the prediction engines on the test data. The left (right) plot depicts ROC curves for the individual prediction engines and the combination using the mean combiner when working with the minimal (maximal) feature set. Results are compared to the performance of the Stanford HIVdb based genotypic susceptibility score.

at the follow-up time point. Logistic regression was replaced by linear regression in the individual classifiers for predicting the change in VL. Using the maximal feature set the GD (ME, EV) engine achieved a correlation (r) of 0.658±0.023 (0.664±0.023, 0.679±0.020) on the training set (Rosen-Zvi et al., 2008). The mean combiner yielded a correlation of 0.691±0.019. However, the oracle achieved $r = 0.834 \pm 0.012$. Although small, the difference between EV and the mean prediction reached statistical significance ($p \approx 0.005$) using a one-sided paired Wilcoxon test. Results on the test set were qualitatively the same: GD (ME, EV) reached a correlation of 0.657 (0.642, 0.678) and the mean combiner (oracle) reached 0.681 (0.814).

5.4.3 Discussion

The performance of the methods considered for combining the individual classifiers improved only little over the single best method on both sets of available features. It turns out that the simple non-trainable methods perform quite well, especially the mean combiner. This phenomenon has been previously discussed in literature (Kuncheva et al., 2001; Liu and Yuan, 2001). Here we focused on finding the best combination strategy for a particular task. The advantage of the mean combiner is that it does not require an additional training step (and therefore no additional data), although it ranges among the best methods studied. Moreover, this combination strategy is easy to explain to end-users of the prediction system.

The learning curves in Figure 5.5 show that the mean combiner learns faster (gives more reliable predictions with fewer training data) than the individual prediction systems. Moreover, the curves show that the combined performance is not dominated by the single

Engine	minimal feature set *ambiguous* instances removed from				maximal feature set *ambiguous* instances removed from			
	none	only Train	only Test	both	none	only Train	only Test	both
GD	0.744	0.738	0.784	0.786	0.760	0.747	0.808	0.806
ME	0.745	0.739	0.770	0.771	0.742	0.757	0.808	0.810
EV	0.768	0.776	0.811	0.824	0.804	0.812	0.846	0.855
Mean	0.772	0.767	0.812	0.814	0.793	0.791	0.849	0.849
Feature	0.747	0.754	0.797	0.808	0.779	0.787	0.832	0.842

Table 5.4: Results on the (un)cleaned test set when individual classifiers are trained on the (un)cleaned training set. The table summarizes the results, measured in AUC for the individual classifiers, the mean combiner, and the combination of feature level when retrained on the (un)cleaned training set and tested on the (un)cleaned test set. Cleaned refers to the removal of ambiguous failing instances.

best approach as the results on the full training set might suggest. Furthermore, the learning curve for the combination on the feature level indicates that more training data is needed to achieve full performance. In general, combining the three individual approaches leads to a reduction of the standard deviation for almost all combination methods. This suggests a more robust behavior of the combined system.

In the cases of failing regimens, all three classifiers very frequently agree upon the wrong label, precisely in 350 of 900 (39%) failing regimens in the training data using the minimal feature set. There are two possible scenarios why the VL drop below 500 copies per ml did not take place during the observed time interval despite the concordant prediction of success by all the three engines:

1. Resistance against one or more antiretroviral agents is not visible in the available baseline genotype but stored in the viral population and rapidly selected, which would lead to an initial decrease in VL shortly after therapy switch, and a subsequent rapid increase before the target time frame (Figure 5.7 (b) lower right).

2. The patient/virus is heavily pretreated and therefore takes longer to respond to the changed regimen, or the patient is not completely adherent to the regimen, both cases lead to a delayed reduction in VL after the observed time frame (Figure 5.7 (b) lower left).

Figure 5.7 (a) shows the distribution of predicted success provided by the mean combiner using the minimal feature set. There is a clear peak around 0.8 for instances labeled as success whereas the predictions for the failing cases seem to be uniformly distributed. Interestingly, the distribution of the failing cases with a VL below 500 copies per ml resembles more the distribution for success than for failure.

The approach to predicting the change in VL exhibited moderate performance. In general, the task of predicting change in VL is harder, since many host factors, which are not available to the prediction engines, contribute to the effective change in individual patients. However, guidelines for treating HIV patients recommend a complete suppression of the virus below the limit of detection (Hammer et al., 2008). Thus, dichotomizing the outcome

Method	minimal feature set		maximal feature set	
	Train	Test	Train	Test
removed from test only				
Single Best	0.809 (0.021)	0.811	0.839 (0.017)	0.847
Oracle	0.935 (0.012)	0.936	0.945 (0.014)	0.950
Min	0.817 (0.019)	0.807	0.847 (0.022)	0.848
Max	0.807 (0.024)	0.810	0.832 (0.018)	0.824
Median	0.820 (0.020)	0.810	0.844 (0.021)	0.835
Mean	0.823 (0.019)	0.816	0.850 (0.019)	0.847
Logistic Regression	0.824 (0.019)	0.816	0.852 (0.017)	0.856
Decision Templates	0.823 (0.018)	0.818	0.851 (0.019)	0.850
Clustering	0.822 (0.020)	0.808	0.852 (0.017)	0.850
Local Accuracy	0.823 (0.019)	0.813	0.850 (0.019)	0.844
removed from train and test				
Logistic Regression	0.825 (0.019)	0.816	0.852 (0.017)	0.856
Decision Templates	0.823 (0.018)	0.818	0.851 (0.019)	0.850
Clustering	0.822 (0.021)	0.796	0.852 (0.017)	0.844
Local Accuracy	0.823 (0.019)	0.813	0.850 (0.019)	0.843

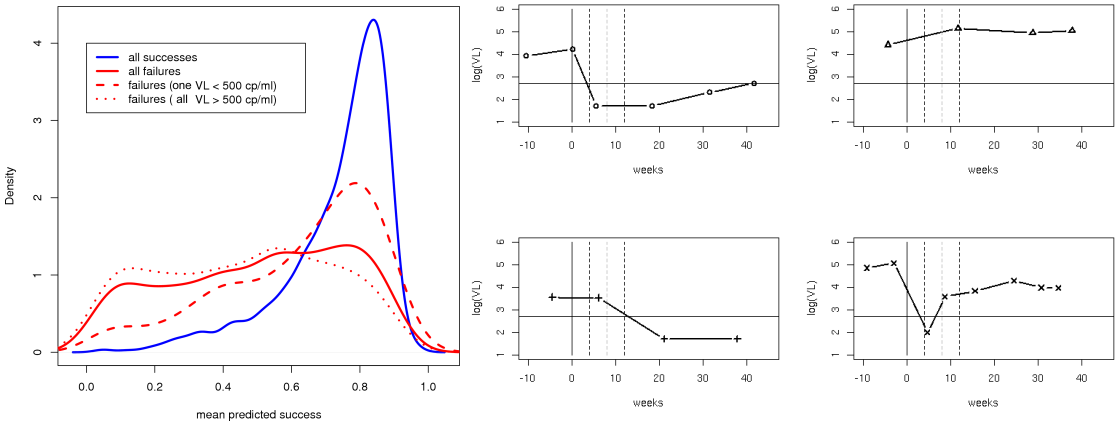
Table 5.5: AUC for the combined engines on training set and test set with the ambiguous cases removed from test set and training set or test set only. The table displays the results, measured in AUC, on training set (10-fold cross-validation; standard deviation in brackets) and test set for a selection of combination approaches when trained on the (un)cleaned training set. For computation of the AUC the ambiguous cases were always removed.

and instead solving the classification task is an adequate solution, since classifiers can be used for computing the probability of achieving complete suppression.

5.4.4 Conclusion

The use of the maximal feature set consistently outperformed that of the minimal feature set in the combined system. Among the studied combination approaches the logistic regression performed best, although not significantly better than the mean of the individual classifications. The mean is a simple and effective combination method for this scenario. Variations in the size of the training set showed that a system combining the individual classifiers by the mean achieves better performance with fewer training samples than the individual classifiers themselves or a logistic regression using all the features of the individual classifiers. This and the consistent reduction of the standard deviation of the performance measures lead to the conclusion that the mean combiner is more robust than the individual classifiers, although the performance is not always significantly improved. Moreover, the mean is a combination strategy that is easily explainable to the end-users of the system.

In this study we discovered ambiguous failures. These therapies are classified as failure but have a VL measurement below 500 copies per ml. Although these instances did significantly influence neither the learning of the individual classifiers nor the learning of the combination method, they lead to an underestimation of the performance. This suggests



(a) Distribution of predicted success probabilities

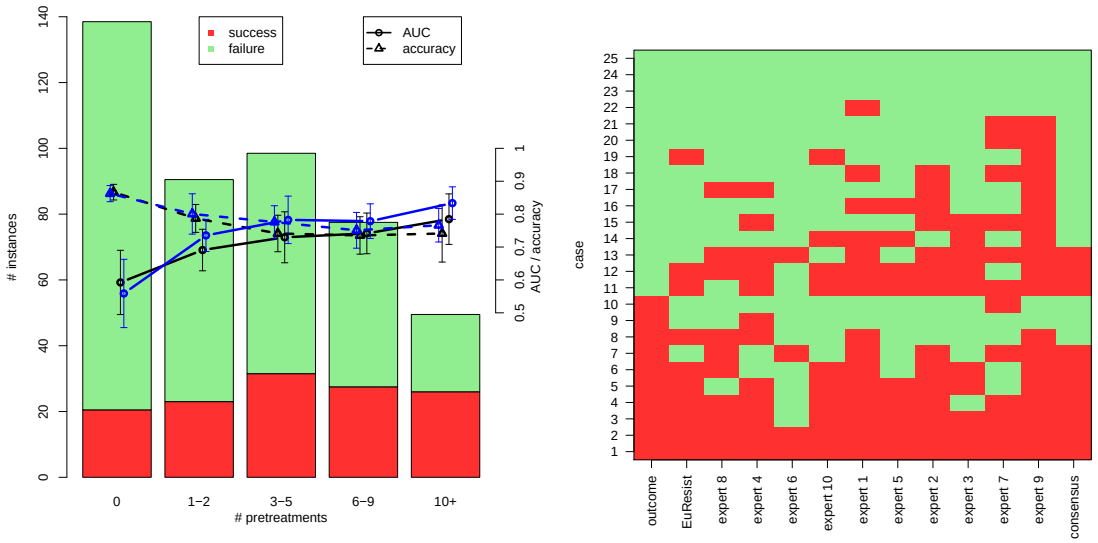
(b) Different viral load trajectories

Figure 5.7: (a) Distribution of the predicted success for all successful therapies (blue solid), all failing therapies (red solid), failing therapies with at least one VL measure below 500 during the regimen (red dashed), and failing therapies with all VL measures above 500 (red dotted) of the mean combiner using the minimal feature set. (b) Different viral load trajectories. The top row depicts a clear success (left) and a clear failure (right). The bottom row depicts trajectories that are labeled as failures because the VL closest to eight weeks of treatment exceeds the limit of detection (horizontal line). Those cases are nevertheless often predicted as success.

that clinically relevant adjustments of the definition of success and failure can result in increased accuracy of the combined engine.

The training and test data comprise patients with very differing level of pre-treatment, i.e. from therapy-naïve patients to patients receiving their 15th antiretroviral regimen. Hence, it is of interest how the engine performs for those patients. Figure 5.8 a) depicts the performance of the EV engine for different groups of patients measured by 10-fold cross-validation on the training set (EuResist release from 2008/10/10). The accuracy is the highest for naïve patients, and interestingly, the AUC is the lowest for this group. The low AUC can be explained by the fact that all patients in that group receive high predicted success probabilities, and a substantial fraction of the few failing regimens might be indeed ambiguous failures. Nevertheless, the AUC rapidly increases and reaches 0.84 for the group with the highest level of pre-treatment. The accuracy, on the other hand, decreases from the initial 0.82 and stabilizes at 0.75. Again, usage of the maximal feature set shows a better performance in all groups (except for naïve patients).

An older version of the EuResist engine (trained on release from 2007/08/29) was compared to the opinion of 12 international HIV treatment experts (who were allowed to use any available interpretation algorithm) on 25 cases from the test set. Only ten experts handed in all their predictions and were therefore considered. The result of this Engine vs. Experts (EVE) study is shown in Figure 5.8 b). It turned out that the prediction engine performed as accurately as the best expert (expert 8). Moreover, the EuResist prediction



(a) Group-wise Performance

(b) Result of the EVE Study

Figure 5.8: Performance of the EV engine with respect to pre-treatment level (a). Bars denote the mean number of samples in the group. Green (red) denotes the mean number of successful (failing) therapies in the group. Performance is measured in AUC (solid line) and accuracy (dashed line) using the minimal (black) and maximal (blue) feature set. Result of the EVE study (b). Every row corresponds to one case, and the columns denote the true outcome of the treatment, the EuResist prediction, predictions by the 10 experts, and the consensus of the expert predictions. Green relates to (predicted) success and red to (predicted) failure.

engine provides only six wrong predictions, and interestingly, the decision profile of the prediction engine is very similar to the consensus decision of all ten experts, which makes also six mistakes.

5.5 The EuResist Web Service

The prediction engines developed by the three institutes are hosted on servers within these institutes. Every institute is running a SOAP (once defined as: simple object access protocol) server. A SOAP protocol defines which data and in which format may be exchanged between a client and a server. In essence, SOAP is the successor of XML-RPC. The specifications are usually written down in a WSDL (web service description language) file. Modern programming languages support the construction of SOAP clients and servers on the basis of the WSDL files.

Each engine has its own endpoint that conveniently provides the required WSDL file upon request (i.e. by adding "?wsdl" to the endpoint of the engine): EV engine[2], GD engine[3], and ME engine[4]. Thus, once the endpoint of an engine is known, virtually everyone can build a SOAP client to send valid queries to the server. The structure of the SOAP services selected in the EuResist project allows every institution to implement the service in a preferred programming language. Moreover, the use of SOAP services allows for a variety of tools: web sites, stand-alone programs, or integration into existing programs for management of clinical data. If preferred, then only a subset of the engines can be queried as well. All the different clients accessing the array of prediction engines can be implemented without any adaptation of the underlying SOAP servers.

Indeed, the EuResist prediction engine was integrated into InfCare HIV (http://www.infcare.se), a Swedish management tool for HIV data.

5.5.1 Implementation of the EV Engine

The SOAP server of the EV engine is written in PHP. Functions that are specific to the EV engine – especially the computation of the genetic barrier to drug resistance – are implement in C++ and made available to PHP via a PHP extension. The original computation of the genetic barrier to drug resistance was rather slow, mainly due to the requirement of computing all transition probabilities in the mixture of mutagenetic trees. The speed up was achieved by precomputing all possible values of the genetic barrier offline. This was possible since the mixture of mutagenetic trees perceive the virus only as a binary pattern of length l, with l being the number of predefined events (Section 4.1.2). Thus, there is only a finite set of possible values the genetic barrier can take. This workaround facilitates fast response times of the web service.

5.5.2 EuResist web tool

The most visible portal for accessing the EuResist prediction engine is the dedicated web site (http://engine.euresist.org), which was created by one of the EuResist partners.

[2] http://euresist.bioinf.mpi-inf.mpg.de/prediction/?wsdl
[3] http://srv-peres.haifa.il.ibm.com:9080/EuResistIbmWeb/services/EuResistIBMEngine?wsdl
[4] http://www.informadoc.net/euresistibmserviceinterfaces.asmx?wsdl

The web site is responsible for collecting the data that is sent to the array of prediction engines. Figure 5.9 displays the data input page of the web site. Mandatory information for a successful request is the HIV sequence comprising protease and reverse transcriptase. If no further information is provided, then the sequence is aligned to a consensus sequence; the list of mutations along with a predefined list of putative treatments is sent to the three individual engines. The user can also provide optional information that is available on the patient: the number of previous treatment lines, previously taken antiretroviral drugs, VL, $CD4^+$ T cell counts, and demographic data (gender, age, mode of HIV transmission). Like in g2p-THEO, the user can influence the list of putative treatments by excluding drugs from being part of any treatment. The three engines predict response to each of the treatments on the provided list and send their results back to the web site.

After receiving predictions for all putative treatments, the web site generates and displays a top 10 list of combination treatments (Figure 5.10). Unlike in g2p-THEO, the ranking is not only based on the mean predicted success but also on the range of the predictions. Precisely, the ranking criterion is:

$$(\text{mean\_of\_predictions}) \times (1 - \frac{\text{range\_of\_predictions}}{\alpha}),$$

with $\alpha = 1.3$. The user can then interactively reorder the top 10 list with respect to success rate or range of the predictions only. Moreover, for drugs that are currently not supported by the EuResist prediction engine, the web site queries Stanford's HIVdb and provides the result. The prediction results are followed by a summary of the data that was provided to the engines and an analysis of the mutations, e.g. whether they are known resistance mutations or not. At the top right of the results page the user can click on "Request detail". The resulting page (Figure 5.11) states which engines were contacted and whether the request was successful.

5.6 Towards Prediction of Sustained Response

Currently, as many other tools, the EuResist prediction engine focuses on predicting initial virological response, i.e. reduction of VL within 4-12 weeks of treatment. However, clinicians are also interested in the response to the selected treatment beyond this short period. Thus, performance of the EuResist prediction engine of inferring sustained response has to be properly assessed. For this analysis, sustained response is measured at 16-32 weeks of treatment. Sustained response data were extracted from the EuResist integrated database comprising data from Italy, Sweden, Germany, and Luxembourg (i.e. version from 2007/11/27).

5.6.1 Material and Methods

Data

The data for this study were extracted according to the standard datum definition presented in Section 5.1. For the comparison we used the originally employed definition that focuses on initial response without changes, i.e. response measured at 8 ($\pm$4) weeks of treatment, and a modified version with sustained response measured at 24 ($\pm$8) weeks of

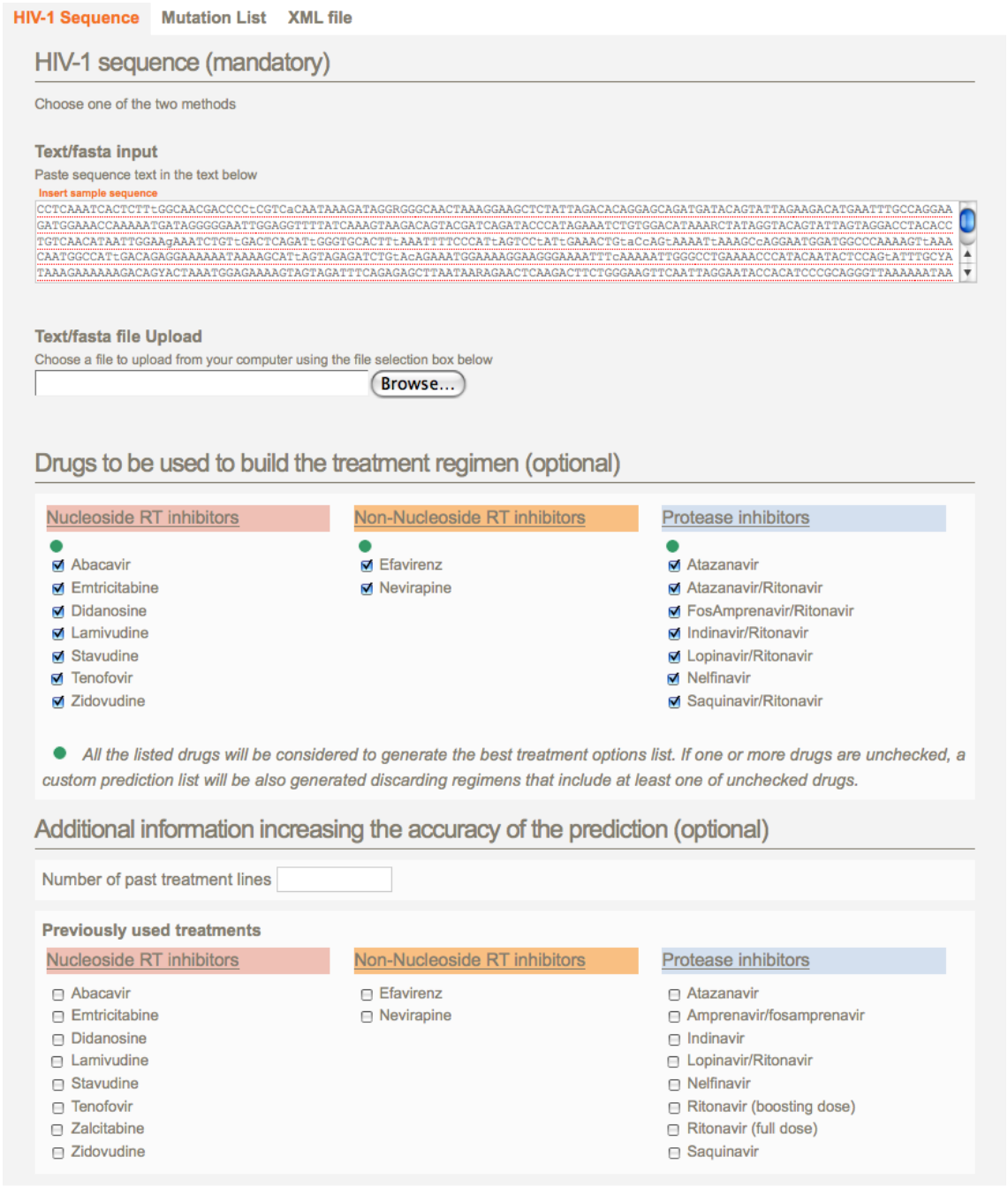
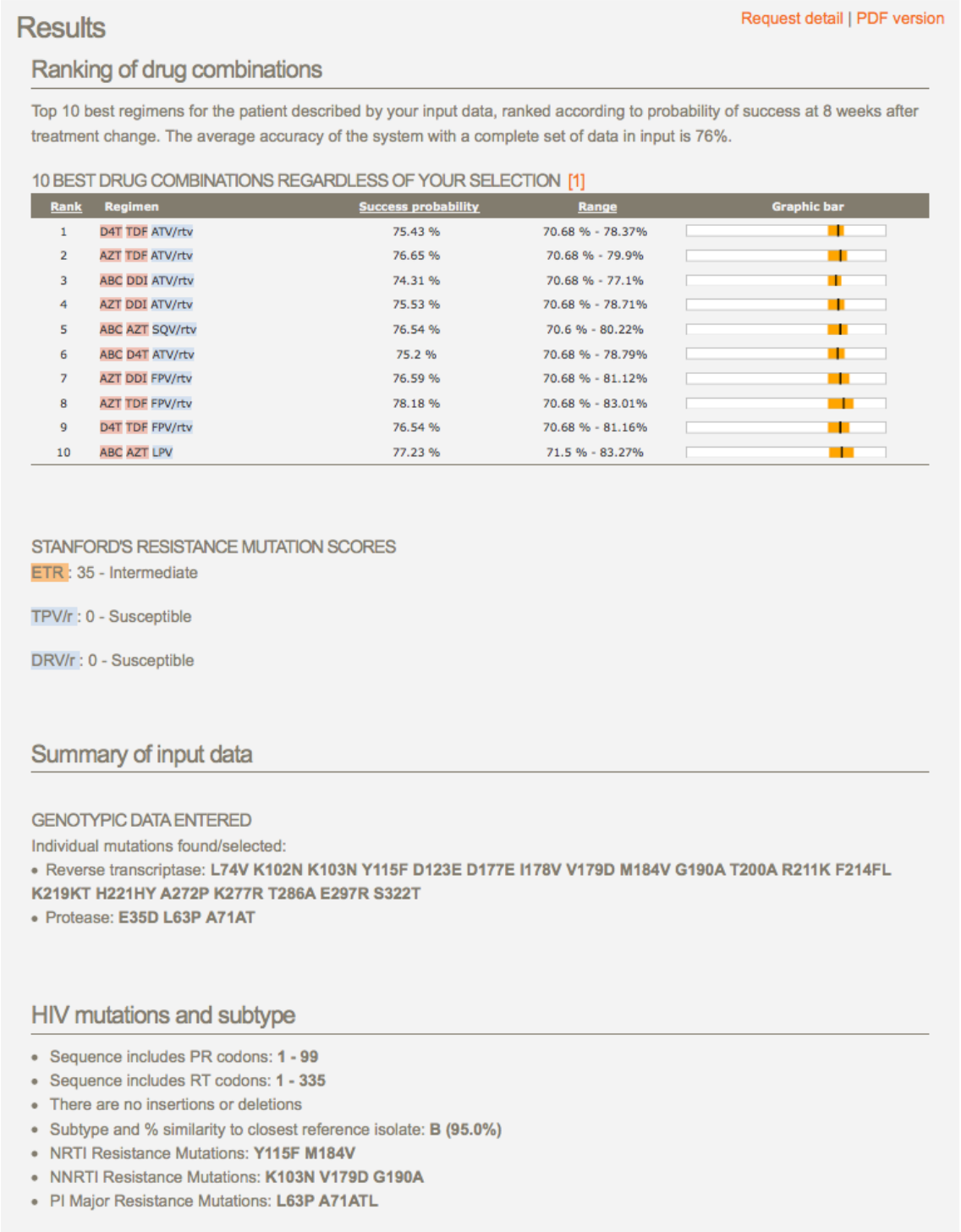


Figure 5.9: Data input page of the EuResist prediction engine. The form contains fields for pasting a HIV sequence comprising protease and reverse transcriptase. Alternatively, the sequence can be uploaded as fasta or mutation can be selected from a drop-down menu. In the field below one can select the drugs that should be considered when ranking different regimens. The remaining fields ask for additional data, e.g. number past treatment lines and previously used drugs.



Results

Request detail | PDF version

Ranking of drug combinations

Top 10 best regimens for the patient described by your input data, ranked according to probability of success at 8 weeks after treatment change. The average accuracy of the system with a complete set of data in input is 76%.

10 BEST DRUG COMBINATIONS REGARDLESS OF YOUR SELECTION [1]

Rank	Regimen	Success probability	Range	Graphic bar
1	D4T TDF ATV/rtv	75.43 %	70.68 % - 78.37%	
2	AZT TDF ATV/rtv	76.65 %	70.68 % - 79.9%	
3	ABC DDI ATV/rtv	74.31 %	70.68 % - 77.1%	
4	AZT DDI ATV/rtv	75.53 %	70.68 % - 78.71%	
5	ABC AZT SQV/rtv	76.54 %	70.6 % - 80.22%	
6	ABC D4T ATV/rtv	75.2 %	70.68 % - 78.79%	
7	AZT DDI FPV/rtv	76.59 %	70.68 % - 81.12%	
8	AZT TDF FPV/rtv	78.18 %	70.68 % - 83.01%	
9	D4T TDF FPV/rtv	76.54 %	70.68 % - 81.16%	
10	ABC AZT LPV	77.23 %	71.5 % - 83.27%	

STANFORD'S RESISTANCE MUTATION SCORES

ETR : 35 - Intermediate

TPV/r : 0 - Susceptible

DRV/r : 0 - Susceptible

Summary of input data

GENOTYPIC DATA ENTERED

Individual mutations found/selected:

- Reverse transcriptase: **L74V K102N K103N Y115F D123E D177E I178V V179D M184V G190A T200A R211K F214FL K219KT H221HY A272P K277R T286A E297R S322T**
- Protease: **E35D L63P A71AT**

HIV mutations and subtype

- Sequence includes PR codons: **1 - 99**
- Sequence includes RT codons: **1 - 335**
- There are no insertions or deletions
- Subtype and % similarity to closest reference isolate: **B (95.0%)**
- NRTI Resistance Mutations: **Y115F M184V**
- NNRTI Resistance Mutations: **K103N V179D G190A**
- PI Major Resistance Mutations: **L63P A71ATL**

Figure 5.10: Results page of the EuResist prediction engine.

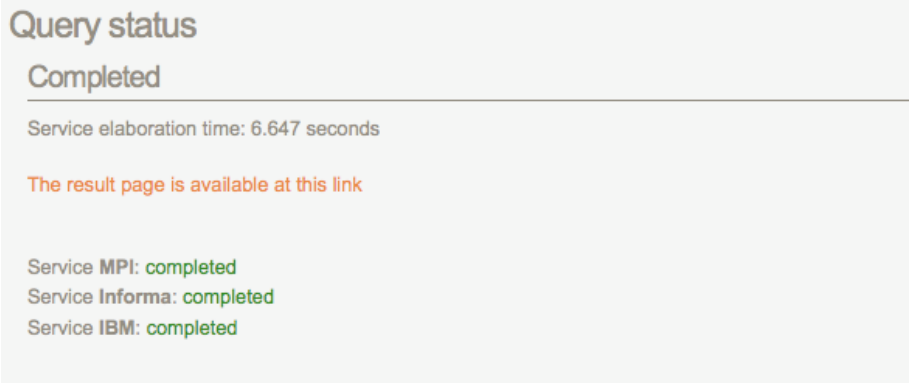


Figure 5.11: Details of the request. The page summarizes how long did the generation of the result took and which SOAP servers were queried.

treatment. In both cases a success was defined as a drop of viral load below 500 copies per ml or, alternatively, a 100-fold reduction compared to the baseline value. 3,023 treatments met these requirements for initial response and 2,722 (301) were used for training (testing) the EuResist prediction engine as described in the previous section (short-term dataset). A comparable number of 2,778 treatments met the requirements for sustained response (long-term dataset). For avoiding over optimistic estimation of the prediction performance, the instances in the long-term dataset were grouped depending on whether they were also present in the short-term dataset. Precisely, dataset A comprised 1,837 treatments of the short-term training set for which a classification with respect to long-term response was available (1,540 of those share the same label). Dataset B comprised all 742 treatments for which only a long-term label was available. And dataset C comprised 199 treatments from the short-term test set for which also a long-term label was available (167 of them have the same label). Treatments in dataset A and B contributed to the training set for a long-term prediction model, while dataset C was used as an independent test set. Changes of the label from success at initial response to failure at sustained response were with 142 (17) as frequent as changes from failure to success with 155 (15) in dataset A (C).

Comparative Analysis

The EuResist prediction engine was used to predict the response to antiretroviral therapy for all treatments in datasets B and C. In order to rate the achieved prediction performance correctly, the results were compared to a long-term prediction engine. This engine was trained from scratch (i.e. feature selection and statistical learning) for this task solely on data from datasets A and B. The resulting model was used to predict the sustained response for all treatments in dataset C. Predictions by this model for treatments in dataset B were obtained in a cross-validation like procedure that will be explained in the following paragraph.

The training procedure for the sustained prediction model was as follows. Training data for the long-term model comprised datasets A and B. Precisely, dataset B was split into 10 equally sized subsets, and dataset A plus nine of the subsets from dataset B were used as training set. The remaining subsets from dataset B operated as a test set. This procedure

was repeated 10 times. Hence, every subset of dataset B was used as a test set once. Therefore, results can be compared to the predictions given by the short-term model for instances in dataset B. The final long-term model was trained using the complete data from datasets A and B. And response to treatments in dataset C was predicted and results compared to the short-term predictions of dataset C. Due to unavailability of predictions of the GD engine for dataset B, only the ME and EV engines were used for the comparison on dataset B. Results on dataset C include the predictions made by all three engines.

As described in Section 5.2, logistic regression was the only statistical learning method for all engines. Again, the three predictions provided by the individual engines were combined using the mean. Performance was measured in AUC (area under the ROC curve). Statistical significance was computed using a Wilcoxon test for testing whether the two areas under the ROC curve are different.

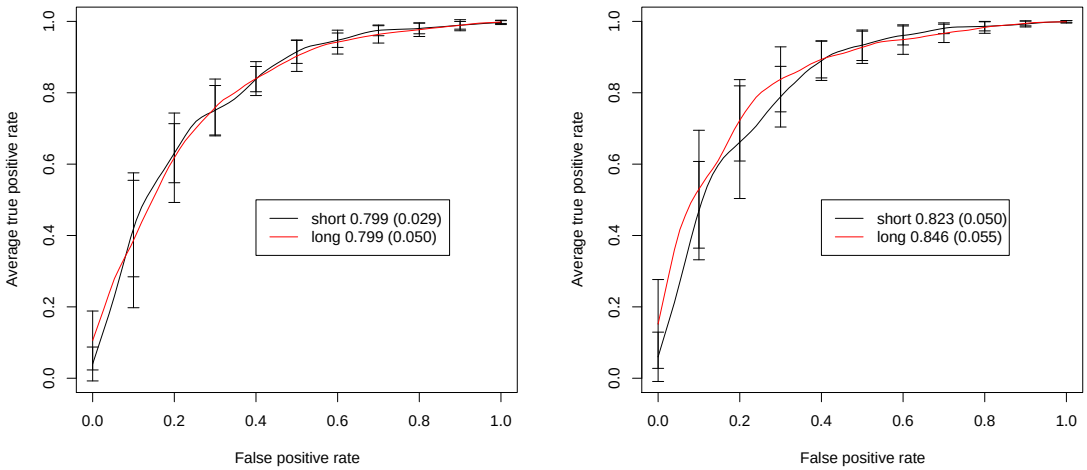
All experiments were carried out using both, the minimal feature set and the maximal feature set. Results on dataset C were also compared to a GSS prediction based on the Stanford HIVdb (for details see e.g. Section 4.3). In addition to the sustained outcome of the treatment, the initial response for the instances in dataset C was available. Thus, on dataset C performance of the all models is assessed with respect to both time points.

5.6.2 Results

Figures 5.12 (a) and (b) show the ROC curves on dataset B. Briefly, the short-term model achieved an AUC on dataset B of 0.799±0.029 (0.823±0.05) compared to 0.799±0.05 (0.846±0.055) achieved by the long-term model using the minimal (maximal) feature set. In both cases the difference was not significant (p=0.99). When using the maximal feature set, however, the performance difference between the two models was more pronounced in favor of the long-term prediction engine. Results were qualitatively the same on the independent test set C. The short-term model achieved an AUC of 0.757 (0.799) compared to 0.77 (0.81) achieved by the long-term model using the minimal (maximal) feature set. Figures 5.13 (a) and (b) show the corresponding ROC curves. Here the long-term model is better when predicting sustained response and the short-term model is better when studying initial response. The performance of HIVdb was better for predicting sustained response than initial response. In general HIVdb performed worse than either of the prediction models.

5.6.3 Discussion

In this analysis we showed that the EuResist prediction engine predicted long-term response as well as short-term response. This fact has been demonstrated previously for rules-based prediction systems (Zazzi et al., 2009). Moreover, the EuResist prediction engine performs as well as a system trained for predicting long-term response. A likely explanation is that in many cases the short-term response and long-term response share the same label (≈ 82% in dataset A and C). Interestingly, the performance for all tested systems increased when studying the sustained response.



(a) CV performance on B (min)

(b) CV Performance on B (max)

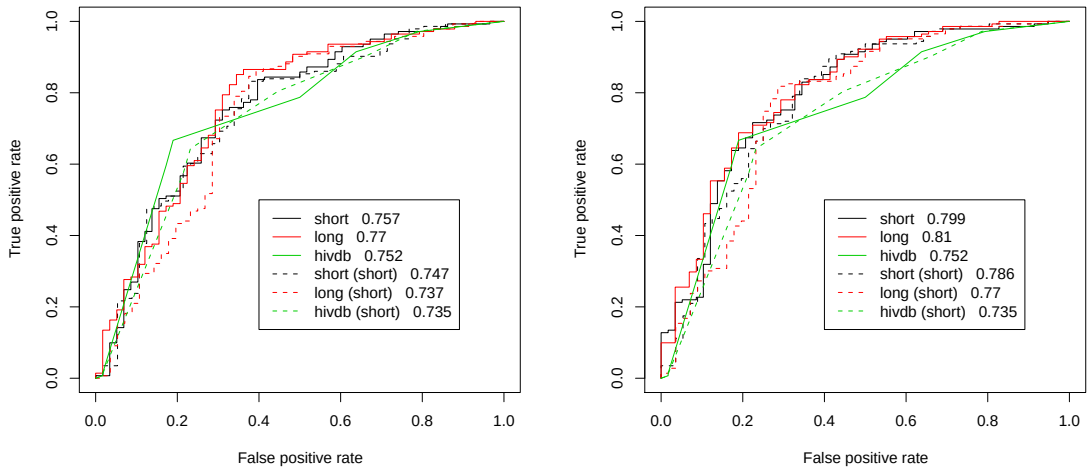
Figure 5.12: ROC curves for dataset B. Curves were generated using the minimal (a) and maximal (b) feature set, respectively, with whiskers indicating the standard deviation, for the initial response model (short) and the sustained response model (long).

5.7 Effect of Modified TCE Definitions

The results in Section 5.3 indicated that the standard datum definition is suitable but has serious drawbacks, the main one being that treatments are labeled as failures albeit the fact that the VL is sufficiently reduced during course of the treatment. As discussed, the factors leading to a delayed response, e.g. suboptimal adherence, usually cannot be inferred from the viral genotype. Luckily, the suboptimal labeling does not have a great impact on the learned statistical models but only on assessed performance leading to an underestimation of the usefulness of the tool in clinical practice (Section 5.3). Likewise, factors that lead to an initial response but a virological failure only a few weeks later may not be found in the viral genotype as well, for example, decreased adherence of the patient (after improved health conditions due to initial response or due to the appearance of side-effects), archived drug resistance mutations that are not visible in the baseline genotype, rapid development of drug resistance.

An improved standard datum definition could consult multiple VL measurements before labeling a treatment as success or failure. The problem with relying on multiple measurements at specific time points is the expected decrease of training samples. For instance, only about 2000 of the 3000 standard datum instances used in the previous section have also a VL available in the four months window around 24 weeks of treatment.

Here we introduce two modified standard datum definitions that use multiple time points of VL measurement for dichotomizing virological response to success and failure.



(a) CV performance on C (min)

(b) CV Performance on C (max)

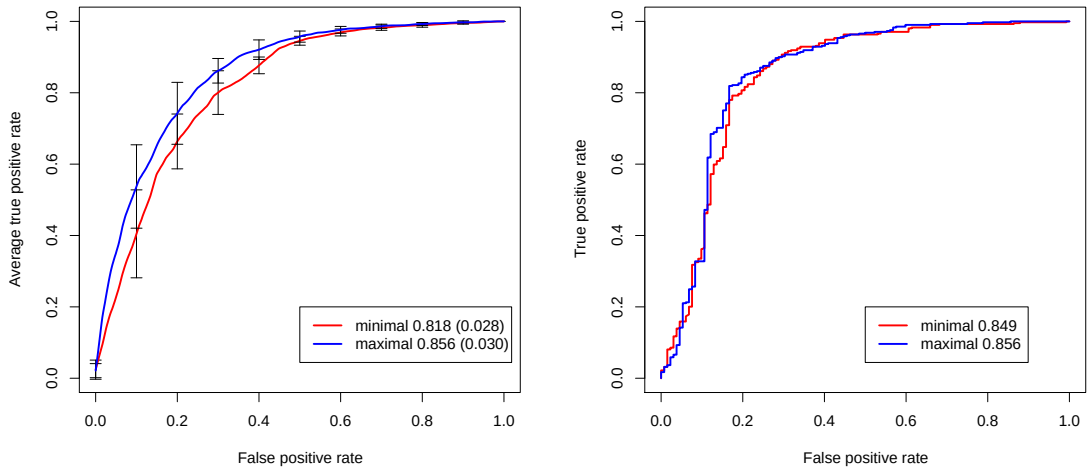
Figure 5.13: ROC curves for the independent test set C. Here both outcomes, initial response (dashed lines) and sustained response (solid lines) are studied for the short-term model (short) and the sustained response model (long). The performance is compared to the performance of Stanford HIVdb (hivdb).

5.7.1 Material and Methods

Alternative Standard Datum Definition

The basic idea behind the first modified standard datum definition is to extract from the database instances with initial and instances with sustained response separately. If there are instances that receive a label for both time points, then those with discordant labels are removed from the final dataset. For this study we use the most recent version of the EuResist integrated database (2008/10/10). The database gives rise to 5001 and 4456 instances for initial and sustained response, respectively. 3504 instances received a label for both time points; 2959 of those instances received the same label. This initial data is augmented by instances that only received a label for initial response (n=1497) or only for sustained response (n=952). The resulting dataset comprises 5408 treatment switches of which 10% are randomly assigned to a test set (n=541). The success rate was 0.747 and 0.756 in the training and test set, respectively. This definition is termed TEP (for: two end points).

The second modified standard datum definition examines the area under the $\log_{10}$(VL) curve after treatment start as basis for the labeling. Here, all available VL measurements between treatment start and at 24 ($\pm$8) weeks are used to compute the area under the $\log_{10}$(VL) curve. By definition, at least two VL measurements are required, the time point of the first measurement is set to treatment start, and the time point of the last measurement is set to 24 weeks. The minimum expected area under the VL curve is $168 \times \log_{10}(50) \approx 285$, corresponding to the fact that all measurements are at the level of detection, i.e. 50 copies of viral RNA per ml blood. However, we usually apply a threshold



(a) Performance on training data

(b) Performance on test data

Figure 5.14: Performance of the EV engine when trained and tested using the TEP definition. The engine was trained with the minimal and maximal feature set. Performance in 10-fold cross-validation (a) and on an independent test set (b).

of 500 copies per ml as the database features also values using an older, less sensitive VL assay. Applying the 500 copies threshold leads to a cutoff for the area under the VL curve of approximately 450. As the choice of the cutoff is not straight forward, we apply an array of cutoffs: 400, 450, 500, 600, and 700, corresponding to an average VL of approximately 240, 477, 946, 3727, and 14677, respectively. The dataset comprises a subset of instance from the sustained response dataset. Precisely, those instances with more than one VL measurement during the observation period (n=4059). An independent test set is created by randomly selecting 10% of the instances (n=406). This definition is termed AVL (for: area under VL).

Prediction Engine

Due to practical reasons only the EV engine is trained and tested on the datasets. For training the engine we use exactly the same features and feature selection model as described earlier in Section 5.2. For the AVL labeling, feature selection was carried out only once with a cutoff of 500. Performance is again assessed using the area under the receiver operating characteristics curve.

5.7.2 Results

Using the TEP definition, the EV engine achieves an AUC of 0.818±0.028 (0.856±0.03) and 0.849 (0.856) in the 10-fold cross-validation and on the independent test set, respectively, using the minimal (maximal) feature set. Figure 5.14 depicts the corresponding ROC curves.

Table 5.6 lists the AUCs achieved on training and test data using the AVL labeling. The

		400	450	500	600	700
success rate	train	0.627	0.728	0.785	0.858	0.915
	test	0.611	0.695	0.749	0.825	0.911
minimal	train	0.712 (0.019)	0.753 (0.028)	0.788 (0.035)	0.810 (0.032)	0.782 (0.040)
	test	0.724	0.742	0.772	0.840	0.829
maximal	train	0.791 (0.023)	0.822 (0.024)	0.860 (0.021)	0.867 (0.020)	0.867 (0.036)
	test	0.794	0.797	0.823	0.878	0.894

Table 5.6: Performance using the area under the $\log_{10}$(VL) curve for labeling. The columns correspond to different thresholds used for dichotomizing the area into success and failure. The first two rows state the fraction of successful treatments in the dataset (train and test). The following rows show the performance of the minimal and maximal feature set measured in AUC on the training and test data.

values range from 0.712±0.019 to 0.782±0.04 in the 10-fold cross-validation when restricted to the minimal feature set. Application of the maximal feature set leads to an increase of around 0.07 in AUC. Moreover, the higher the selected threshold the higher the achieved AUC.

5.7.3 Discussion

The observed performance with the TEP definition is close to the one achieved after removing the ambiguous failures from the dataset in Section 5.3. This suggests that the labeling using the modified definition is clearer. A confounding factor is obviously the increase in data compared to the previous analysis. However, when the modified definition was applied to an older version of the EuResist database (2007/11/27) a similar boost in performance was observed (data not shown). Moreover, using a training set of comparable size originating from the most recent release, with labels according to initial response only, shows AUC values around 0.75 (see e.g. results in Section 5.8). Thus, the observed boost in performance mainly originates from the change in the standard datum definition.

One could argue that removing the cases with discordant labels at the two time points would only leave the "easy to predict" cases in the training and test data. In general, cases are considered to be simple if the viral genotype shows few mutations, i.e. the patient is not very treatment experienced. However, the number of recorded treatments prior to the treatment switch, which is a measure of treatment experience, did not differ significantly between the time points for initial and sustained response ($p = 0.6529$ computed with a Kolmogorov-Smirnov test). Neither was the difference between initial response and the combined response significant ($p = 0.99$). Thus, a differing level of treatment experience as a source for improved prediction performance can be excluded.

In general, results obtained with the AVL definition are inferior to the ones obtained using the TEP definition, probably owing to the fact that the AVL definition does not exclude instances that represent unclear cases (e.g. slow decrease in VL). And indeed, if the instances with discordant labels in initial and sustained response are excluded from the dataset the AUC reaches 0.824±0.032 and 0.837 on training and test set, respectively, with the minimal feature set and a cutoff of 500. However, the increase in AUC with increasing

threshold used for the labeling indicates that the task of sorting out clear failures is not so hard to solve. The cutoffs 600 and 700 represent treatments with an average VL of 3727 and 14677, respectively.

As long as there are known factors that significantly influence the response to treatment and these factors cannot be included in the statistical model (for any reason) one should select a labeling that removes (or at least reduces) the impact of these factors. One such factor is the adherence of the patient to the prescribed treatment. Also the RDI tried to exclude cases that are likely to be related to poor adherence. Precisely, if at baseline the patient had a VL of less than 1000 copies per ml and the VL increased after the treatment change by more than 300-fold (i.e. 2.5 $\log_{10}$ VL) then the TCE was excluded (Larder et al., 2007). The TEP definition provided a labeling involving instances that are less likely corrupted by incomplete adherence. However, inspection of the VL trajectories of a patient's past treatments could give a clue about his or her general adherence and provide a useful covariate for predicting response to combination therapy.

There are many ways to modify the original standard datum definition, e.g. by changing the outcome time points or considering more than two outcome time points. When modifying the definition of response one has to consider not to pose too many restrictions as one would drastically reduce the amount of available data. Moreover, and most importantly, the definition of treatment response must be in concordance with clinical practice: only in this case one can provide a tool that can be used in clinical routine.

5.8 Integrating Novel Drugs

The work described in this section has been presented at the 7th European HIV Drug Resistance Workshop 2009, Stockholm, Sweden (Altmann et al., 2009c).

The beauty of purely data-driven approaches to predict either drug resistance of the virus to single drugs or the virological response to combination therapy is that they rely solely on data for constructing a model that solves the task. These approaches, however, require a sufficient amount of data to solve the task with a certain reliability. Hence, the beauty of the approach is also its major weakness, as data for newly licensed drugs are rarely available in sufficient amounts. In the case of geno2pheno, phenotypic resistance data are required. The data, however, demands the availability of the drug to the laboratory, and this usually only the case after official release of the drug on the market. Moreover, once the drug is available, the experimental data are still cost- and labor-intensive to obtain. In the case of prediction engines that infer response to combinations of antiretroviral agents, the databases collecting the training data have to be updated with treatments comprising the novel compounds. Unfortunately, this process can take years, e.g. in the latest release of the EuResist database (2008/10/10) there exist only 119 standard datum instances featuring three novel drugs (the oldest was approved in June 2005). Owing to the problems of update ability, the data-driven approaches are doomed to lag behind the new developments by the pharmaceutical companies (unless they will agree to share their data on clinical trials to facilitate updating such tools). For geno2pheno we investigated the use of semi-supervised learning techniques for improving the prediction of HIV drug resistance in situations with only few training samples (see Perner et al. (2009) and Section 3.3). Unfortunately, a benefit of semi-supervised learning could not be confirmed for every drug

class. More precisely, some model assumptions of the employed semi-supervised learning methods could even be harmful for the classification performance of some drugs.

A solution to the update problem of prediction engines like g2p-THEO and EuResist that is based on semi-supervised learning strategies is unlikely to be successful, simply owing to the use of drug combinations instead of single drugs as in the case of geno2pheno. Here we investigate a strategy that aims at incorporating expert-based prediction systems, which generally are rapidly updated following the release of a new drug, into the data driven approach. The approach borrows ideas from language processing where *discounting* is the process of replacing the original counts of events with modified counts in order to redistribute the probability mass from the frequently observed events to rare or even unseen events. For example, absolute discounting removes a constant number from the counts for all events (Ney et al., 1994).

5.8.1 Materials and Methods

Standard Encodings

From the latest release of the EuResist integrated database (2008/10/10) we extracted 4,538 standard datum instances (original definition as in Figure 5.1). These therapy switches were used as training data for the EV engine. Response to antiretroviral treatment was dichotomized to success and failure based on the 8 ($\pm$4) week follow-up viral load (VL).

The baseline model used the minimal feature set, i.e. indicators for drugs in the regimen, mutations in the baseline genotype, and the genetic barrier to drug resistance. The extended model used the maximal feature set, i.e. additionally baseline VL, indicators for previous use of a drug, and indicators modeling previous exposure to NRTIs, NNRTIs, and PIs. Both encodings are termed standard encodings.

ndGSS Encodings

The alternative EV engine incorporates rules-based ratings for novel drugs as additional features. We overcome the fact that only very few real novel drugs (considered by today's standards) are available in the training data by treating all drugs in a regimen that were approved by the FDA less than four years prior to treatment start as novel drugs; simply because these drugs were real novel drugs at the time the treatment started. The genotypic susceptibility score (GSS) for drugs that are considered novel was computed using Stanford's HIVdb (version 5.0.0; HIVdb5) and used as an additional covariate termed ndGSS (for: novel drug GSS). This step constitutes the discounting-like part of the modeling strategy, since indicators and genetic barrier for drugs contributing to ndGSS were set to 0, i.e. with respect to the standard encoding the novel drugs were not existing in the treatment. Table 5.7 lists the year of FDA approval for the considered drugs and the year until when those drugs were considered novel. Moreover, the table lists in how many instances the drugs were used and how many instances were affected by the discounting. In order to be able to distinguish between the absence of novel drugs and all the virus being resistant to all novel drugs (in both cases the ndGSS is 0), the ndGSS was set to -1 if all novel drugs were classified as resistant by HIVdb5. Moreover, as novel compounds can be expected

	ZDV	ddI	d4T	3TC	ABC	TDF	NVP	EFV
Year FDA approval	1987	1991	1994	1995	1998	2001	1996	1998
Year until *novel*	1991	1995	1998	1999	2001	2003	2000	2002
Drug used in instances	1355	1085	900	3123	752	1852	395	823
# instances changed	0	0	60	133	153	413	75	243

	IDV	SQV	NFV	LPV	APV	ATV
Year FDA approval	1996	1995	1997	2000	1999	2003
Year until *novel*	1999	1999	2000	2002	2003	2004
Drug used in instances	237	245	359	1726	314	423
# instances changed	66	54	128	394	69	110

Table 5.7: Year of approval of the drugs by the FDA and year until the drugs were considered novel. The last two rows list the number of standard datum instances in which the drugs were used, and the number of instances that were affected by the discounting-like step.

to be more powerful drugs, we introduce a weighting factor (wf) for the ndGSS feature. Precisely, the ndGSS feature is multiplied by the weighting factor during the prediction process.

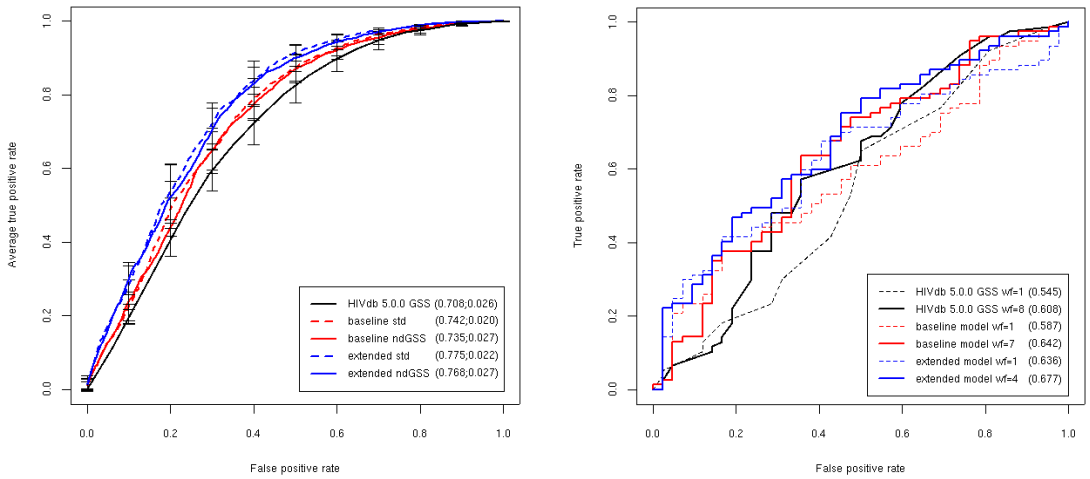
Performance Assessment

Performances of standard and the ndGSS encodings were compared in a 10-fold cross-validation on the training data. In this setting we seek to investigate whether the changes applied to the standard encoding have a serious impact on the performance with respect to the established compounds. To this end, the folds making up the training data for the ndGSS encoding were used as is to train the classifier, the test fold, however, comprises the standard encoding. Thus both approaches are tested on the same encoding (i.e. the standard encoding).

An additional 119 standard datum instances containing real new drugs (DRV n=59, TPV n=52, ETR n=4, DRV+ETR n=4) were extracted from the EuResist database for independently assessing performance of the ndGSS encoding. The ndGSS for these new drugs was computed using HIVdb5 as well. Performances of the ndGSS encodings were compared with the performance achieved by a HIVdb5-based GSS for the complete regimen. The weighting factor was optimized by five-fold cross-validation on these 119 instances. Classification performance was assessed using the area under the ROC curve (AUC).

5.8.2 Results

In the cross-validation setting the baseline (extended) model using the standard and the ndGSS encoding achieved an AUC of 0.742±0.020 (0.775±0.022) and 0.735±0.027 (0.768±0.027), respectively. For comparison, HIVdb5 reached an AUC of 0.708±0.026. On the 119 instances including real novel drugs HIVdb5 achieved an AUC of 0.545 (almost random) while the baseline (extended) model using the ndGSS encoding yielded an AUC



(a) Performance on training data

(b) Performance on instances with real new drugs

Figure 5.15: ROC curves on training data (a) and test data containing novel drugs (b). Whiskers indicate the standard deviation.

of 0.587 (0.636). Using the optimized weighting factor for ndGSS elevated the AUC to 0.608 (wf=8), 0.642 (wf=7), and 0.677 (wf=4) for HIVdb5, baseline, and extended model, respectively.

5.8.3 Conclusions

The cross-validation analysis demonstrated that there was no statistical significance between the standard and the ndGSS encoding ($p > 0.13$ using a paired Wilcoxon test). This indicates that the precision of the system regarding established drugs is not compromised by the discounting-like approach. The baseline encoding significantly outperformed HIVdb5 and the extended encoding significantly outperformed the baseline encoding (both: $p = 0.02$).

Performance on the 119 test instances featuring real novel drugs was disastrous. However, even the reference approach (HIVdb5) yielded a performance close to random. Both, the baseline and the extended ndGSS models outperformed HIVdb5, i.e. they could predict response better than the GSS alone. Thus, the modeling approach can be regarded as a step in the right direction. Boosting the influence of the ndGSS feature elevated the AUC of all models by approximately 0.05. This effect is probably related to the increased potency of the novel drugs.

5.9 Therapy History: Replacement for the Genotype?

The vast majority of tools that predict response to antiretroviral therapy focus on the viral genotype as the major source of information. Indeed, obtaining the sequence of the viral drug targets is a standard approach for finding the best treatment for HIV patients in developed countries. However, despite dropping prices for genome sequencing, genotyping is not a standard practice in resource-limited settings, where even the standard VL assay

exceeds the budget for standard care. Unfortunately, the majority of people infected with HIV live in such resource-limited countries.

The viral drug targets are steadily responding to the ongoing treatment and each treatment leaves traces in the viral genotype. In fact, the genotype obtained by bulk sequencing may not provide all mutations that ever occurred in the viral population. Precisely, if mutations do not present a replicative advantage, resistance mutations may disappear from the currently predominant viral variant. For example, the predominant viral population in patients having a treatment break has usually the characteristic of the wild type virus. Unfortunately, the patient harbors previous viral variants in the form of proviral DNA in several infected tissues. This constitutes a memory of resistance mutations provoked by previous treatments. As a consequence, recycling of drugs leads to a rapid reselection of previously existing resistance mutations, which are not prevalent in the predominant viral variant. For avoiding such a short-term viral rebound, the treating clinician considers the patient's treatment history, i.e. the previously taken drugs, in addition to the viral genotype when selecting a new regimen. Treatment history has long been recognized as clinically relevant (Bratt et al., 1998). More recently it was shown that taking all available genotypes into account improves the prediction of treatment response in heavily pretreated patients (Zaccarelli et al., 2009). Thus, a prediction model focusing on treatment history could be an effective alternative to a genotype-based model in resource-limited settings.

This study aimed at investigating which representation of the treatment history is the most useful one for inferring response to combination antiretroviral therapy. And how the predictions of an history-only engine compare to state-of-the-art predictions based on the viral genotype. The research was part of Fabian Müller's Bachelor Thesis (Müller, 2008) and only the main results are summarized in the following sections.

5.9.1 Material and Methods

Data and Response Definition

For the experiments we used the 2007/08/16 release of the EuResist integrated database. The definition of successful virological response was modified compared to the standard datum definition for allowing to exploit the wealth of data in the database. A success was defined if at least one VL measurement during the treatment was below 500 copies per ml. If the lowest measurement in the database exceeded the value, then the treatment was labeled as a treatment failure. Moreover, since we are interested in the impact of treatment history, a baseline genotype was not required for inclusion of a treatment in the study. Of note, the response definition used here is yet another alternative to the originally defined standard datum definition.

The definition resulted in 35,149 treatments being labeled as success or failure, we refer to this dataset as HO (for: history only). Of those treatment changes, a set of 3,910 could be associated with a genotype that was obtained at most 90 days prior to treatment start. These instances were collected in a second dataset termed WG (for: with genotype). The HO dataset comprised 1689 distinct combinations of antiretroviral drugs.

Features of Treatment History

The current therapy, i.e. the therapy of which the virological response has to be inferred, was encoded using 17 binary variables each representing a single drug. This encoding was used in previous sections and is termed *no history*.

The indicators for the current therapy were augmented with features derived from the patient's treatment history. The simplest encoding used one indicator for previous exposure to a drug for each of the 17 compounds, hence it is denoted as *binary history*. A more elaborate encoding takes into account the time since last exposure to the drug. Here the history is not encoded as a binary variable, but as a continuous variable ranging between 0 (never exposed) to 1 (very recently exposed). Precisely, the variable was defined as

$$f(t,k) = \begin{cases} 1 - \frac{1}{4300^k} t^k & 0 \leq t \leq 4300 \\ 0 & \text{else} \end{cases},$$

where t is the time since last exposure in days (with a maximum of 4300) and k is a parameter controlling the decay of the influence of previously used drugs. A value of 1 for k equals linear decay, values larger (smaller) than 1 lead to a slower (faster) decay of the influence. Due to the nature of the encoding it is called *continuous history*. The motivation for this feature is that compounds that were taken a long time ago may have a reduced impact on the current regimen.

In addition to a simple concatenation of features for the current drugs and historic use of drugs, special interaction features were introduced aiming at modeling the interaction between previously used drugs and drugs in the current regimen. For example, if the current treatment does not include any protease inhibitor, the previous use of protease inhibitors is expected to have little impact on the treatment outcome. Like in the ME engine, second order features represent the previous use of a compound and the current use of another drug from the same class, i.e. 49 (7×7) and 100 (10×10) additional features model interaction between PIs and RTIs, respectively. These 149 variables extend the 17 binary variables for the current regimen. This approach was applied to the binary and continuous history features, thus, leading to the *2nd order binary* and *2nd order continuous* encodings.

Genotype Features

For correctly rating the performance of the history-only prediction within a state-of-the-art context, we compared it to a prediction based on genotype. Precisely, we applied the encoding used for g2p-THEO (49 binary variables representing mutations in protease and reverse transcriptase, 17 binary indicators for the drugs in the current regimen, and 17 variables for the genetic barrier to drug resistance) and trained a classifier on the WG dataset.

Statistical Learning

The features were tested with two statistical learning techniques: logistic regression and random forest. The number of trees in the forest was set to 100. The parameters of the history encoding, the cutoff in days for a compounds for being considered as part

of the history, and k, which controls the decay of the influence of previously used drugs, were optimized in a 10-fold cross-validation setting (separately for both learning methods). Moreover, all history encodings were compared in a 10-fold cross-validation setting on the HO dataset. Performance of the genotype-based model was assessed in a 10-fold cross-validation setting on the WG dataset. And for the comparison to the genotype-based model the history-only models was trained on only those instances of HO for which no associated genotype was available (i.e. HO without WG: 31,239), the remaining instances were used as a test set and split into 10 equally sized folds matching the folds of the cross-validation for the genotype-based model.

In addition to a performance comparison we also studied the best way for combining genotype and history predictions. Here we focused on three strategies. In strategy one we added the history features as additional covariates to the genotype encoding (*concatenation*). This approach, however, limits the data used for training to the WG dataset. Strategy two used the prediction computed by the history-only model as an additional covariate in the genotype encoding (*prediction feature*). This approach is akin to the one used in GD engine (Section 5.2), which adds the prediction of a Bayesian network to the list of features. The third strategy simply computed the mean of the history-only and the genotype-based prediction (*combined by mean*). In the latter two approaches the history model could be trained on all available data (i.e. HO without WG), and the genotype-based model only on WG.

Performance is assessed as the area under the ROC curve. Significance is assessed using a paired one-sided Wilcoxon Rank Sum test.

5.9.2 Results

The results suggested that previous use of a drug should be considered from the first day on. Linear increase of the threshold at which drugs are considered for the history encoding resulted in a linear decrease in AUC for both classifiers (data not shown; see Müller (2008) instead). Moreover, the best parameter k for modeling the influence of past treatments on the current regimen was 1 (i.e. linear decay) for both classifiers. It turned out that values of k that are smaller than 1 impair the performance of logistic regression. The random forest classifier was more robust with respect to the choice of k. For all further experiments, k was set to 1 and drugs were considered from the first day on.

Table 5.8 summarizes the performance of the different history encodings obtained on the HO dataset. Using the binary history significantly, outperformed the predictions based on the current treatment only for both classifiers ($p < 0.001$). For logistic regression the best encoding was the 2^{nd} order binary encoding that significantly outperformed the standard binary encoding ($p < 0.001$). The random forest classifier achieved the best results with the continuous encoding, which also significantly outperformed the binary one ($p < 0.02$). The use of second order features seriously impaired the performance of random forest, which performed significantly worse than their standard counterparts ($p < 0.001$). The random forest classifier performs in general better than the logistic regression. However, the initial difference of 0.03 in AUC (no history) could be reduced to 0.008 using the best performing history features for both approaches.

Table 5.9 shows the performance of the best history encoding and the genotype-based

	logistic regression	random forest
no history	0.670 (0.011)	0.700 (0.008)
binary	0.713 (0.008)	0.729 (0.007)
continuous	0.714 (0.009)	0.734 (0.008)
2^{nd} order binary	0.726 (0.008)	0.720 (0.007)
2^{nd} order continuous	0.707 (0.011)	0.727 (0.007)

Table 5.8: Mean 10-fold cross-validation performance in AUC on the HO dataset. Standard deviation is given in parenthesis.

	logistic regression	random forest
history-only	0.722 (0.027)	0.746 (0.023)
genotype-based	0.766 (0.025)	0.771 (0.022)
concatenation	0.778 (0.025)	0.793 (0.018)
prediction feature	0.781 (0.024)	0.787 (0.020)
combined by mean	0.780 (0.025)	0.794 (0.022)

Table 5.9: Mean 10-fold cross-validation performance on the dataset WG achieved by the best performing history-only models and the genotype based prediction and the three combination approaches. Standard deviation is given in parenthesis.

predictor on the WG dataset. The genotype based prediction clearly outperformed the history-only predictions for both classifiers ($p < 0.002$). Again, the random forest classifier performs better than the logistic regression. Using the genotype features, however, the difference is only marginal – an effect that we observed before with g2p-THEO (Chapter 4). All three ways for combining the history information and the genotype information significantly outperformed the genotype alone ($p = 0.032$ for the smallest improvement: logistic regression and concatenation). None of the improvements among the different ways of combination achieved statistical significance.

5.9.3 Discussion

The study showed that prediction of treatment response based on the current regimen and the genotype of the virus only can significantly be improved by including information on the patient's treatment history. Moreover, recoding the exposure to a previously used drugs from the first day on provided the best results. We explored further features derived from the patient's treatment history, among those were similarities of the current therapy to the preceding regimen, but none of the other encodings could improve the performance over the results presented here (Müller, 2008).

A possible confounder of the analysis was the fact that the treatment history was likely to be incomplete for a large number of patients. However, repeating the analysis with a dataset comprising only patients with their treatment history completely recorded in the database, which resulted in a dataset with 17,720 treatment changes, led to the same results. The AUC for all encodings of the history (including no history) were about 0.03 higher then the results in Table 5.8 (data not shown).

In our study, the history-based prediction is clearly inferior to the genotype-based prediction. Recently, other studies presented results where the difference between a genotype-based and a treatment history-based prediction were much less pronounced (Prosperi et al., 2009a; Revell et al., 2009). A potential cause for the discrepancy could be that the latter two studies aimed at predicting virological response at a fixed time point and additionally allowed the use of baseline VL as an additional covariate. In both studies, which applied random forests as a statistical learning method, baseline VL was ranked as the most important feature. However, due to a known bias of the random forest to rate continuous or categorical variables higher than variables with few categories – especially binary ones (Strobl et al., 2007), the impact of VL should be reassessed. The baseline VL might nevertheless be a good predictor for response to combination therapy, its application in resource-limited settings, however, remains doubtful; owing to the high cost of the viral load assay. Fiscus et al. (2006) investigate alternatives of the established VL assay that can be used in regions with limited resources where response to antiretroviral treatment is still typically monitored using the $CD4^+$ T cell counts.

In an ongoing work Saigo et al. (in preparation) use a tailored machine learning approach based on subsequence mining (Nowozin et al., 2007) that tries to exploit the order in which drugs were applied in the patient's history. Here the value of genotypic information over treatment history alone could be confirmed using the original standard datum as response definition. Moreover, interpretation of the model revealed that the information about initial response (success or failure) of treatments in the past plays an important role.

To conclude, the study demonstrated that response to treatment can be predicted with good performance based on treatment history alone. The best representation of the treatment history was different for the two evaluated statistical learning methods. Such history-based models might be an option for resource-limited settings and are currently subject of further investigations (Prosperi et al., 2009a; Revell et al., 2009), and (Saigo et al., in preparation). Such models will certainly be useful as soon as the multitude of antiretroviral drugs is also available in resource-limited settings. So far, the WHO addressed the inability to provide personalized HIV treatment to those patients by issuing simplified treatment protocols that cover first- and second-line treatment (Gilks et al., 2006).

It still remains a question whether the performances of the history-based models assessed on databases reflecting the pandemic in the developed countries can be transferred to the situation in resource-limited settings; with the majority of viruses originating from other subtypes than B. Moreover, the performance of the history-based models is assessed on treatment change decisions that are usually genotype-based, either on a baseline genotype or a historic genotype. Thus, the genotypic information is implicitly contained in the drug combination that the model is asked to assess and therefore may lead to an overestimated performance in the comparison to genotype based methods. Therefore, either prospective studies or retrospective analyses on treatment changes that were not based on genotypes are required to assess the usefulness of history-only-based prediction models.

6 Planning Sequences of HIV Therapies

In the previous chapters we demonstrated that predicting virological response to combination antiretroviral therapy is a feasible task. The resulting models, however, focus solely on the efficacy of a regimen, which is only one aspect of a putative treatment the clinician has to assess. The other side requiring consideration is the effect of the regimen on the virus (or the whole viral population); specifically, which mutations will the regimen provoke and how will these mutations affect the efficacy of future treatments. Patients infected with HIV cannot be cured currently since the virus integrates its DNA into the genome of the host cell, thus making a life-long therapy a necessity. For ensuring the best use of the currently available panel of antiretroviral drugs it is essential to apply them in an order that avoids or at least minimizes cross-resistance effects and exploits potential resensitization effects.

Jiang et al. (2003) introduced the notion of future drug options (FDO) based on resistance of the virus against single drugs and complete drug classes as assessed by a rules-based interpretation algorithm. The FDO measure was applied to data from the clinical trial ACTG 364 (Katzenstein et al., 2003) conducted by the AIDS clinical trial group (ACTG) and focused on assessing the impact of baseline phenotypic resistance on treatment length. The trail comprised mainly treatments with two NRTIs and the PI NFV and/or the NNRTI EFV. Difference in the FDO between the viral genotype before treatment start and at virological failure was studied for determining the most effective regimen in terms of resistance cost. When corrected with respect to the time to virological failure, EFV based treatments exhibit a higher resistance cost than NFV or NFV+EFV based regimens. The study clearly demonstrates the use of the FDO concept for devising general treatment guidelines. For using this approach to develop a personalized treatment plan, however, one has to compare the current predominant viral population in the patient to the makeup of that population at failure of the possible regimens.

Various approaches have been undertaken to stochastically simulate the evolution of the viral population during treatment, for examples see Deforche et al. (2008); Prosperi et al. (2009b); Perelson (2002); Ribeiro and Bonhoeffer (2000) and references therein. These approaches focus on modeling of the infection dynamics of HIV, beginning with a simple hunter-pray dynamics up to models that consider changes in the viral environment affecting its replicative success, for example the presence of antiretroviral drugs. The latter models qualify for approaching the problem of estimating the viral population at failure of a regimen. The computational time, however, typically required for this *in silico* evolution significantly exceeds the time available for interactive web services (e.g. about 3 min for a single drug combination (Prosperi et al., 2009b)). Moreover, these models are based on certain modeling assumptions and rely on parameters (e.g. rate at which new target cells are generated, death rate of infected and uninfected cells) that are hard to estimate from available biological data.

This chapter introduces an approach for computing the viral population at the time of treatment failure. The approach is based on finite state machines and is scalable, thus eventually affording web services with moderate waiting times. The strategy for calculating personalized future drug options comprises three steps: (i) selection of putative effective regimens with methods described earlier, (ii) *in silico* evolution of the virus according to a particular treatment, and (iii) assessment of the FDO at failure of that putative drug combination.

Sections 6.1 and 6.2 introduce the framework employed for simulating the viral evolution during treatment. Section 6.3 describes the applied mutation models along with their assumptions, and Section 6.4 presents a validation for the mutation models. Section 6.5 concludes the chapter with final remarks and a case study focusing on the four most prominent first-line regimens.

6.1 Weighted Finite State Transducers

Briefly, weighted finite state transducers are a family of finite state automata. They extend the concept of simple acceptors that comprise an input alphabet ($\mathcal{A}$), a set of states (Q) with directed edges between them (E), as well as initial (I) and final (F) states by an output alphabet ($\mathcal{B}$), costs for every edge and cost functions for initial (λ) and final (ρ) states. Formally, a finite state acceptor is quintuple:

$$\mathcal{F} = (\mathcal{A}, Q, I, F, E),$$

where $I, F \subseteq Q$ and $E \subseteq Q \times \mathcal{A} \cup \{\epsilon\} \times Q$, with ϵ beeing the "empty word". Basically, finite state acceptors verify whether words defined over the input alphabet are valid or not. Here, a word is a concatenation of elements of the alphabet. The verification process starts always at the beginning of the word and in one of the initial states of the automaton. From this initial state, the elements of the word are read one-by-one: if from the current state there exists an outgoing edge that is labeled with the corresponding element, then the acceptor proceeds to the target state of the arc and is ready to read the next element of the word. A word is considered valid, only if there exists at least one initial state such that reading the word from that initial state ends in one of the final states.

Analogously, a weighted finite state transducer is defined as:

$$\mathcal{T} = (\mathcal{A}, \mathcal{B}, Q, I, F, E, \lambda, \rho),$$

where $I, F \subseteq Q$, $E \subseteq Q \times \mathcal{A} \cup \{\epsilon\} \times \mathcal{B} \cup \{\epsilon\} \times \mathbb{K} \times Q$, $\lambda : I \to \mathbb{K}$ and $\rho : F \to \mathbb{K}$, with $\mathbb{K}$ being part of a *semiring* $(\mathbb{K}, \oplus, \otimes, \bar{0}, \bar{1})$. Thus, $(\mathbb{K}, \oplus, \bar{0})$ is a commutative monoid with identity element $\bar{0}$, $(\mathbb{K}, \otimes, \bar{1})$ is a monoid with identity element $\bar{1}$, $\otimes$ distributes of $\oplus$, and $\bar{0}$ is an annihilator for $\otimes$, i.e. $\forall a \in \mathbb{K} : a \otimes \bar{0} = \bar{0} \otimes a = \bar{0}$. In addition to acceptors, transducers emit elements of the output alphabet when traversing along an edge from one state to the another. Consequently, for every valid input word, an word over the output alphabet is generated. Moreover, in weighted finite state transducers, edges, initial states, and final states are equipped with scores (over $\mathbb{K}$). Hence, a valid input word and the emitted output word are associated with a total score.

Finite state transducers have been used in the field of speech recognition and language translation and provide a uniform way of representing knowledge for these tasks. In natural

language processing knowledge is typically modeled in the form of conditional probabilities. For example. the probability of observing a certain word w depends on the preceding words in the sentence w', and can therefore be represented as $P(w|w')$. In a finite state automata representation each possible sequence of words, the history, gives rise to a state; every possible succeeding word is represented by an edge leaving that state, and the probability of this word following the history corresponds to the edge weight. For practical reasons the history is limited to a fixed length, i.e. n-gram models model the probability of the n-th word given the $n-1$ preceding ones. Weighted finite state acceptors are sufficient for representing such *language models*. Transducers, however, are required to transform instances from one alphabet to instances from another alphabet, e.g. translating a French input sentence into English. The naïve approach, a word-to-word translation, maps one French input word to an English word. Unfortunately, the translation of a single word is often ambiguous, i.e. the word can have different translations which are highly dependent of the context. As a consequence, one has to model $P(e, f|e', f')$, where e is the translation for f and f' and e' are the history of the French and English words, respectively. Thus, every history of English and French words corresponds to a state in the transducer, every French word together with a possible translations gives rise to an edge leaving that state, and the probability of the French word following the history and the translation to a particular English word given that history constitutes the edge weight. Since the probabilities have to be estimated from training data, which are always scarce, the distribution $P(e, f|e', f')$ is decomposed into $P(e|f)$, $P(f|f')$, and $P(e|e')$ representing the translation model, the French language model, and the English language model, respectively. These models can then be learned independently from different datasets and represented as individual transducers and acceptors. The algorithm *compose* described below constructs an approximation to the $P(e, f|e', f')$ representing transducer from the three individual finite state machines.

For applications in natural language processing the obvious choice for a semiring for transducers is the *Probability semiring*: $(\mathbb{R}_+, +, \times, 0, 1)$. Due to computational reasons the *Log semiring* is preferred: $(\mathbb{R} \cup \{-\infty, +\infty\}, \oplus_{\log}, +, +\infty, 0)$, here all probabilities are represented by their negative natural logarithm and $\oplus_{\log}$ is defined as: $x \oplus_{\log} y = -\log(e^{-x} + e^{-y})$. According to these two semirings the probability of events along a path from an initial state to a final state are multiplied, and probabilities of alternative paths are summed up. However, in language processing tasks the total probability of a sequence of events if often dominated by the probability of the most likely path. Thus, instead of accumulating the probabilities of all alternative paths it is sufficient to compute the probability of the most likely path. This approximation is termed *Viterbi approximation* or *maximum approximation* and is implemented by the *tropical semiring*: $(\mathbb{R} \cup \{-\infty, +\infty\}, \min, +, +\infty, 0)$. Hence, many tasks can be solved efficiently by simply searching the shortest path from an initial to a final state in an automaton.

Due to the computational demands required in natural language processing domains, libraries and toolkits offering efficient implementations of algorithms based on transducers are available, e.g. Mohri et al. (2000); Hetherington (2004); Lombardy et al. (2004); Allauzen et al. (2007). For the experiments presented here the RWTH FSA toolkit (Kanthak and Ney, 2004) is used. The toolkit is open source and offers a range of algorithms, C++ and Python interfaces as well as an command line tool. Many algorithms based on trans-

ducers afford an on-demand implementation. Briefly, an on-demand algorithm computes the outgoing edges (and their weights) of a given state only as soon as the state is visited. For instance, if one wants to multiply all edge weights of a transducer with a scalar and afterwards process a word of the input alphabet, then the multiplication can be done for all edges explicitly before reading the word. Alternatively, in an on-demand implementation the multiplication can be carried out while the word is processed. This has the advantage that only edges of states that are actually required during a computation are updated. Kanthak and Ney (2004) give a definition of local algorithms, which is a prerequisite for the on-demand implementation. Consider an algorithm that generates a new transducer based on one or more transducers constituting its input (e.g. composition of transducers; see following section). If this algorithm is able to generate for the new transducer any arbitrary state and all its outgoing edges depending only on the information about the corresponding state(s) of the algorithm's input transducer(s), then it has the *local* property. For instance, multiplication of edge weights with a scalar value has the local property: for computing an arbitrary state in the resulting transducer, only information on that state in the input transducer is needed (weights of the outgoing edges), all other states are irrelevant.

The command line toolkit reads transducers in a binary and an XML format. The latter is mainly used for construction of new transducers by the user, while the former allows efficient storage and input-output operations. The following sections introduce some basic and important algorithms like the composition of two weighted finite state transducers and the extraction of the shortest path or the n-shortest paths of acceptors.

6.1.1 Composition

Weighted finite state transducer composition is the generalization of nondeterministic finite state automata intersection. The algorithm is used to construct large complex automata from small and simple ones. The intersection of nondeterministic finite state machines is defined as follows. Let $\mathcal{F}_1 = (\mathcal{A}, Q_1, I_1, F_1, E_1)$ and $\mathcal{F}_2 = (\mathcal{A}, Q_2, I_2, F_2, E_2)$ be two finite state acceptors, then the intersection automaton is defined as:

$$\mathcal{F}_1 \cap \mathcal{F}_2 = \mathcal{F} = (\mathcal{A}, Q_1 \times Q_2, I_1 \times I_2, F_1 \times F_2, E),$$

where $((q_1, q_2), a, (q_1', q_2')) \in E \Leftrightarrow (q_1, a, q_1') \in E_1 \wedge (q_2, a, q_2') \in E_2$. Basically, the new automaton comprises one state for every combination of states of the two input automata ($Q_1 \times Q_2$). Thus, each state in the resulting acceptor can be mapped to exactly one state in each of the input acceptors. Likewise, the new initial (final) states are defined by all combinations of initial (final) states of the input acceptors. Edges in the resulting acceptor are added based on pairs of edges of the input automata that have the same edge label: the source (target) of the new edge is the state representing the two source (target) states of the edges in the two input automata.

When moving from acceptors to weighted transducers one has to respect the edge weights as well as the input and output symbols. Thus, let $\mathcal{T}_1 = (\mathcal{A}, \mathcal{B}, Q_1, I_1, F_1, E_1, \lambda_1, \rho_1)$ and $\mathcal{T}_2 = (\mathcal{B}, \mathcal{C}, Q_2, I_2, F_2, E_2, \lambda_2, \rho_2)$ be two weighted transducers. The composition of the two transducers is defined as:

$$\mathcal{T}_1 \circ \mathcal{T}_2 = \mathcal{T} = (\mathcal{A}, \mathcal{C}, Q_1 \times Q_2, I_1 \times I_2, F_1 \times F_2, E, \lambda, \rho),$$

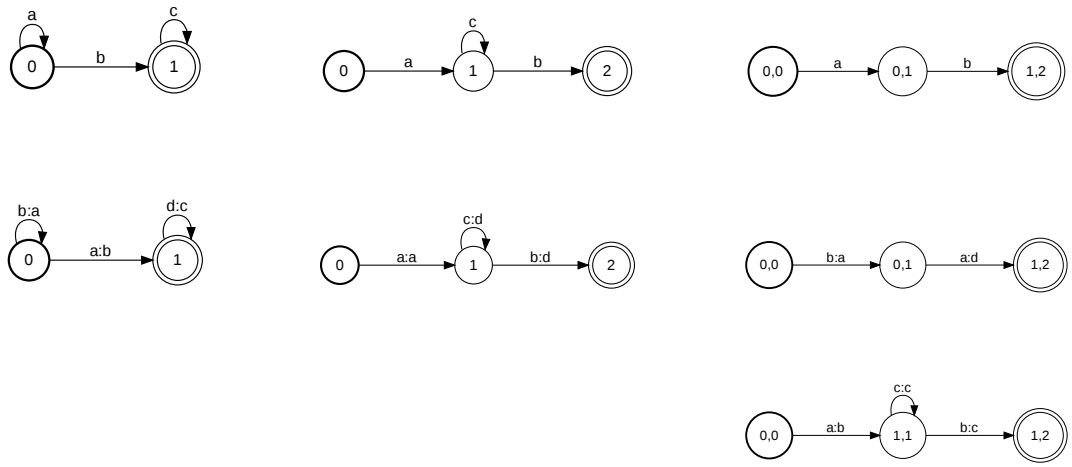


Figure 6.1: Example of acceptor intersection and non-commutative transducer composition. States are represented as circles, initial states and final states are marked using bold and double circle boundaries, respectively. Numbers inside the circles stand for the labels. Edge labels in acceptors comprise a single element of the input alphabet, while edge labels in transducers comprise an element of the input alphabet (left) and one element of the output alphabet (right) separated by a colon. The first row shows the two input acceptors and intersection result in the last column. The second row shows two input transducers and the result of the composition *left* $\circ$ *right* and *right* $\circ$ *left* in the last column of the second and third row, respectively.

where $((q_1, q_2), a, c, w_1 \otimes w_2, (q_1', q_2')) \in E \Leftrightarrow (q_1, a, b, w_1, q_1') \in E_1 \wedge (q_2, b, c, w_2, q_2') \in E_2$, $\lambda(q_1, q_2) \rightarrow w_1 \otimes w_2 \Leftrightarrow \lambda_1(q_1) = w_1 \wedge \lambda_2(q_2) = w_2$, and $\rho(q_1, q_2) \rightarrow w_1 \otimes w_2 \Leftrightarrow \rho_1(q_1) = w_1 \wedge \rho_2(q_2) = w_2$. The major difference (apart from updating the weights for edges, initial states, and final states) to the construction of the intersection acceptor is that edges in transducer composition are added based on pairs of edges of the two input transducers where the output element of one edge matches the input element of the other edge. Consequently, the composition of transducers is not commutative. An example of acceptor intersection (top row) and transducer composition (lower rows) is shown in Figure 6.1.

Algorithm 2 provides the composition algorithm based on Pereira and Riley (1997) in pseudocode. It constructs a transducer $\mathcal{T} = (\mathcal{A}, \mathcal{C}, Q, I, F, E, \lambda, \rho)$ from two input transducers $\mathcal{T}_1$ and $\mathcal{T}_2$. This algorithm, however, assumes ϵ-free input transducers and uses a queue S. The function $E[q]$ retrieves all edges leaving the state q and functions $i[e]$, $o[e]$, and $n[e]$ return the input label, output label, and target state of edge e, respectively.

A problem with ϵ-containing transducers can occur when an ϵ-output edge of $\mathcal{T}_1$ is matched to (a sequence of) ϵ-input edges of $\mathcal{T}_2$. In this case, the application of Algorithm 2 could generate redundant ϵ-paths in the resulting transducers that (depending on the semiring) could lead to incorrect path costs. Thus, all but one ϵ-path have to be removed from the composite transducers. Remarkably, this filtering can be carried out by another

Algorithm 2 Weighted-Composition($\mathcal{T}_1$, $\mathcal{T}_2$)

$Q \leftarrow I_1 \times I_2$
$S \leftarrow I_1 \times I_2$
while $S \neq \emptyset$ **do**
  $(q_1, q_2) \leftarrow$ Head(S); Dequeue(S)
  if $(q_1, q_2) \in I_1 \times I_2$ **then**
    $I \leftarrow I \cup \{(q_1, q_2)\}$
    $\lambda(q_1, q_2) \leftarrow \lambda_1(q_1) \otimes \lambda_2(q_2)$
  end if
  if $(q_1, q_2) \in F_1 \times F_2$ **then**
    $F \leftarrow F \cup \{(q_1, q_2)\}$
    $\rho(q_1, q_2) \leftarrow \rho_1(q_1) \otimes \rho_2(q_2)$
  end if
  for each $(e_1, e_2) \in E[q_1] \times E[q_2]$ such that $o[e_1] = i[e_2]$ **do**
    if $(n[e_1], n[e_2]) \notin Q$ **then**
      $Q \leftarrow Q \cup \{(n[e_1], n[e_2])\}$
      Enqueue($S, (n[e_1], n[e_2])$)
    end if
    $E \leftarrow E \cup \{((q_1, q_2), i[e_1], o[e_2], w[e_1] \otimes w[e_2], (n[e_1], n[e_2]))\}$
  end for
end while
return $\mathcal{T}$

transducer. To this end, the input transducers have to be preprocessed: output (input) ϵ-labels in transducer $\mathcal{T}_1$ ($\mathcal{T}_2$) are mapped to ϵ_2 (ϵ_1) resulting in $\tilde{\mathcal{T}}_1$ ($\tilde{\mathcal{T}}_2$). Using the filter transducer $\mathcal{FT}$, $\mathcal{T}_1 \circ \mathcal{T}_2$ can correctly be computed by $\tilde{\mathcal{T}}_1 \circ \mathcal{FT} \circ \tilde{\mathcal{T}}_2$ using Algorithm 2 (for details see e.g. Pereira and Riley (1997) and Mohri (2005)).

Returning to the language translation example from above, one would create two acceptors, one representing the English language model $\mathcal{E}_{lm}$ and one the French language model $\mathcal{F}_{lm}$. A simple translation transducer from French to English $\mathcal{FE}_t$ can be constructed by a transducer comprising one state and edges for each French input word and a possible English translation. The operation $\mathcal{F}_{lm} \circ \mathcal{FE}_t \circ \mathcal{E}_{lm}$ builds a transducer that translates French sentences to English $\mathcal{T}_{F \to E}$. For translation, a single French sentence is represented as a linear acceptor f with edge labels corresponding to French words, and $f \circ \mathcal{T}_{F \to E}$ generates all possible English translations of f. However, for translating f most parts of the translation transducer $\mathcal{T}_{F \to E}$ are not visited. Here the use of an *on-demand implementation* together with pruning (see following sections) results in improved computational performance.

A transducer example from the domain of Bioinformatics is given by the combination of translating a sequence of nucleotides into amino acids and aligning the result to a reference amino acid sequence. For this task two transducers are required. Transducer $\mathcal{T}_{Tr}$ translates a sequence of nucleotides into a sequence of amino acids, and a second transducer encodes the alignment $\mathcal{T}_{Al}$. $\mathcal{T}_{Tr}$ can be constructed so that all three reading frames are translated, i.e. there is the possibility to read at most two nucleotides without starting the translation

process (see Figure 6.2). The lower part of Figure 6.2 depicts a sketch of the alignment transducer, it supports affine gap costs (g for gap open and ge for gap extend) and one edge for each possible mismatch of amino acids, the cost of a mismatch m can, for instance, be related to the BLOSUM matrix. Let $S_{q,nt}$ and $S_{r,A}$ be linear acceptors representing the query nucleotide sequence and the reference amino acid sequence using the edge labels, respectively, then $S_{q,nt} \circ \mathcal{T}_{Tr} \circ \mathcal{T}_{Al} \circ S_{r,A}$ computes three possible translations of nucleotides to amino acids and the cost of all possible alignments to a reference.

6.1.2 Single Source Shortest Path

The previous section demonstrated how large transducers can be constructed from small ones. The information one wants to compute, however, is still concealed in these large directed graphs. In particular, the set of all possible translations of a French sentence into English is of less interest than the most likely translation. The same holds true for possible alignments. If the transducer is defined over a tropical semiring, however, then the desired information can be retrieved rather efficiently from the large automaton, since standard algorithms can be applied to solve the *single source shortest path (SSSP)* problem for computing the shortest path from an initial state to a final state. For example, Dijkstra's algorithm (Dijkstra, 1959) can be applied if all edge weights are non-negative, the *single source* limitation can be circumvented by a slight modification of the automaton: introduction of a new single initial state that links via ϵ-edges to all previous initial states. The same modification can be applied to the final states, thus only the cost of the shortest path between the single initial and the single final state is of interest. Theoretically, as soon as the execution of Dijkstra's algorithm extracts the final state from the queue, it can be stopped. For allowing also the more general case with negative edge weights (but non-negative circles) the Bellman-Ford algorithm (Bellman, 1958) can be used. Of note, if the semiring used by the transducer is defined over $\mathbb{R}_+$, then one can safely apply Dijkstra's algorithm.

The RWTH FSA toolkit implements Algorithm 3 that was introduced by Mohri (2002). The algorithm is termed *Generic Single Source Shortest Distance* because some well known and widely used algorithms are special cases of this algorithm (hence *generic*) and depending on the semiring the term *path* is not pertinent for what the algorithm computes (hence *distance*). Mohri showed that the algorithm is correct for any semiring and queuing discipline. Precisely, given a transducer that is defined over a tropical semiring the queuing discipline determines which algorithm is executed: if the queue used in Algorithm 3 follows a *first-in first-out* or *shortest-first* principle then the algorithm is equivalent to executing Bellman-Ford or Dijkstra, respectively.

Briefly, the algorithm maintains two properties for each state, $d[q]$ and $r[q]$ store the current distance estimate from the initial state i to state q and the total weight added to $d[q]$ since the last time q was extracted from the queue S, respectively. Like in Algorithm 2, $E[q]$ represents all edges leaving q, and $n[e]$ and $w[e]$ return the target state and the edge weight, respectively. In the tropical semiring, $\oplus$ is equal to *min*. Thus, intuitively, in this case the statement in Line 11 checks whether the current distance of $n[e]$ to i is shorter than the current distance R from i to q plus the edge weight from q to $n[e]$. If this is not the case, then $d[n[e]]$ and $r[n[e]]$ are updated. Furthermore, if $n[e]$ is not in the queue it is

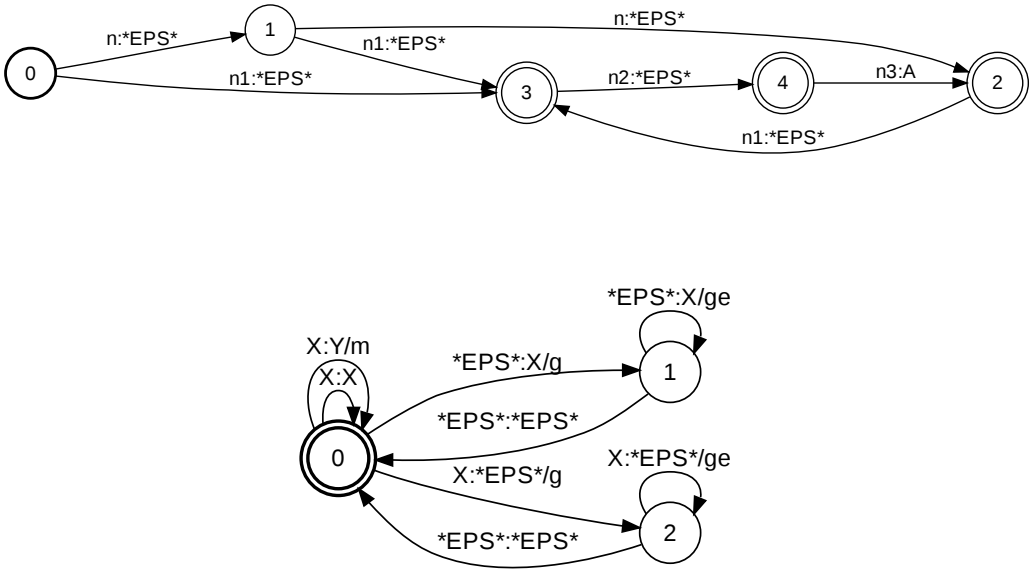


Figure 6.2: Example transducers for translating nucleotide sequences into amino acid sequences and alignment of two sequences. The notation for the transducers is the same as in Figure 6.1. Furthermore, *EPS* corresponds to ϵ (the "empty word") and the slash in the edge label separates the edge weight (on the right) from input and output elements (on the left). The top transducer $\mathcal{T}_{Tr}$ translates a sequence of nucleotides into amino acids, the first two states (0 and 1) allow to read any nucleotide (n) without starting the translation process, the states (3 and 4) represent the fact that the first (n1) and second (n2) nucleotide of a codon have been read, respectively, and the arc between state 4 and 2 emits the amino acid (A) encoded by the codon n1n2n3. The bottom transducer $\mathcal{T}_{Al}$ encodes an alignment with affine gap costs. State 0 has arcs for matches (X:X) and mismatches (X:Y) at cost m. Furthermore, states 1 and 2 encode deletions and insertions, respectively, with respect to the reference sequence.

Algorithm 3 Generic-Single-Source-Shortest-Distance($\mathcal{T}, i$)

```
for j ← 1 to |Q| do
  d[j] ← r[j] ← 0̄
end for
d[i] ← r[i] ← 1̄
S ← {i}
while S ≠ ∅ do
  q ← Head(S); Dequeue(S)
  R ← r[q]
  r[q] ← 0̄
  for each e ∈ E[q] do
    if d[n[e]] ≠ d[n[e]] ⊕ (R ⊗ w[e]) then
      d[n[e]] ← d[n[e]] ⊕ (R ⊗ w[e])
      r[n[e]] ← r[n[e]] ⊕ (R ⊗ w[e])
      if n[e] ∉ S then
        Enqueue(S, n[e])
      end if
    end if
  end for
end while
```

added to S so that its outbound edges can be updated later. For a detailed analysis of the running time, possible modifications of the algorithm and correctness see Mohri (2002).

Given the result of any single source shortest distance algorithm, the shortest path from the source to all other states can easily be extracted. In particular, the paths leading to final states. Therefore, the best translation of a French sentence into English can be found by computing $\mathsf{best}(f \circ \mathcal{T}_{F \to E})$. Analog to this, the best reading frame and alignment of a nucleotide sequence to an amino acid sequence is retrieved by executing $\mathsf{best}(S_{q,nt} \circ \mathcal{T}_{Tr} \circ \mathcal{T}_{Al} \circ S_{r,A})$.

6.1.3 N-Shortest-Strings

Due to imperfect models or violated model assumptions, the best hypothesis of an automaton, as represented by the shortest path, need not be the best solution with respect to a scoring function. If the model is sufficiently accurate, however, then the best (or a better) solution might be among the top n (i.e. most likely) hypotheses represented by the automaton. Moreover, these n-best lists can be post-processed to form a final solution.

In the case of a deterministic acceptor, the n-best paths in the automaton are equivalent to the n-best hypotheses or n-best strings. In nondeterministic acceptors, however, the n-best paths may contain duplicate hypotheses (with different costs). Thus, in order to obtain n distinct hypotheses, nondeterministic automata have to be made deterministic before computing the n-best paths.

The process of computing a deterministic weighted finite state automata from a nondeterministic one works analogously to the unweighted case via the classic subset construction algorithm (Aho et al., 1986). Unlike in the unweighted case, not all weighted automata

can be made deterministic. Fortunately, any acyclic weighted automaton is determinizable (Allauzen and Mohri, 2002). Furthermore, the determinization algorithm affords a local implementation (Mohri and Riley, 2002). This makes the algorithm extremely useful for extraction of n-shortest-strings.

Given an (non)deterministic acceptor $\mathcal{F} = (\mathcal{A}, Q, I, F, E)$, for which we can assume (without loss of generality) that $|I| = |F| = 1$, one has to carry out the following steps to compute the n-best strings. First, a potential function ϕ is computed, which provides for every state $q \in Q$ the shortest distance to the single final state. Obviously, ϕ can be computed by inverting all edges in the automaton and executing Algorithm 3 starting from the single final state. Then, the automaton $\mathcal{F}$ is determinized, resulting in the deterministic automaton $\mathcal{F}' = (\mathcal{A}, Q', I', F', E')$. Based on the potential function ϕ defined for $\mathcal{F}$, the potential function Φ for $\mathcal{F}'$ can be derived on demand (Mohri and Riley, 2002). This potential function for the states of the determinized automaton is the basis for the ordering of the elements in the priority queue S that is used in Algorithm 4 (based on Mohri and Riley (2002)). Precisely, the ordering of the queue is defined as: $(p, c) < (p', c') \Leftrightarrow (c + \Phi[p] < c' + \Phi[p'])$, where $p, p' \in Q'$, and c and c' are the costs from the initial state to the states q and q', respectively. Since Φ never underestimates the cost from the current state to a final state, one can view Φ also as the admissible heuristic that plays a central role in the A^* algorithm (Hart et al., 1968).

Algorithm 4 n-shortest-paths$(\mathcal{F}', n)$

```
    for p ← 1 to |Q'| do
      r[p] ← 0̄
    end for
    π[(i', 0)] ← NIL
 5: S ← {(i', 0)}
    while S ≠ ∅ do
      (p, c) ← Head(S); Dequeue(S)
      r[p] ← r[p] + 1
      if r[p] = n ∧ p ∈ F then
10:     Exit
      end if
      if r[p] ≤ n then
        for each e ∈ E[p] do
          c' ← c ⊗ w[e]
15:       π[(n[e], c')] ← (p, c)
          Enqueue(S, (n[e], c'))
        end for
      end if
    end while
```

The algorithm maintains for each state of the automaton, which was made deterministic, the attribute $r[p]$ that stores the number of times the state (p) was extracted from the priority queue S. This attribute is essential as it provides the stopping criterion for the algorithm: $r[p]$ is initialized with 0 for all states and as soon as the single finial state is

extracted from the queue n times (lines 9-11) the algorithm terminates. Paths are defined by storing a predecessor π for each pair (p, c) (Mohri and Riley, 2002).

The potential function ϕ can be exploited in further ways together with pruning. Briefly, when applying forward pruning the search algorithm does not follow edges that lead to a total path cost exceeding a certain threshold. Mohri and Riley (2001) introduced a weight pushing algorithm that requires the potential function in order to move path weights further to the beginning (or end) of a path. Having the costs of the complete path accumulated at the beginning of the path facilitates more efficient forward pruning, i.e. one can decide not to search along paths that promise a high total cost early in the search process.

6.2 Transducers in Therapy Planning

The strategy for assessing future drug options includes the computation of an estimate of the viral sequence at failure of the therapy. The baseline sequence can be represented as a linear acceptor with the edge labels corresponding to the sequence positions. For allowing an unambiguous identification, edges are labeled Tar_Pos_AA with Tar, Pos, and AA corresponding to the protein the sequence encodes (here $Tar \in \{\text{PRO, RT}\}$), the amino acid position, and the amino acid at that position, respectively.

Given a function $M_t(p, a, a')$ that provides the probability of mutating the amino acids a to a' at position p under the therapy t, one can easily construct a transducer representing this function. This is illustrated by the left transducer in Figure 6.3. The self loops at the initial state and the final state read single sequence positions and leave them unchanged (at no cost). All transitions from the initial state to the final state introduce exactly one mutation, e.g. at position p from a' to a' at the cost of $M_t(p, a, a')$. A composition of the linear sequence automaton with this mutation transducers introduces all possible mutations at all positions. Using the algorithms described above, one can retrieve the sequence with the most likely mutation or the n most likely sequences. Figure 6.4 displays a toy example. The base sequence comprises (the first) three amino acids of the protease (a). The mutation transducer (b) allows one possible mutation for each position (colored edges). The composition of the two automata results in a transducer that has one mutation for each path from the initial state to the final state (c). The most likely mutation can be extracted by finding the shortest path in that transducer (d), and the list of n-best mutations can be generated by extracting the n-best paths (e). The latter algorithm, however, is only defined for acceptors, thus the mutations (colored edges) cannot be inferred from the graph anymore. For introducing two mutations in the baseline sequence one simply has to compute the composition of the transducer in Figure 6.4 c) and the mutation transducer.

Transducers constructed in a different way can be used to assess the activity of a regimen with respect to a given genotype. The right transducer in Figure 6.3 illustrates the concept of these *rating transducers*. In this automaton the output labels are identical to the corresponding input label for all edges. Thus, a composition of an other transducers with the rating transducers leaves the edge labels unchanged. In the rating transducer, the edges are equipped with a weight derived from the relevant rating function $R_t(p, a)$. The rating function $R_t(p, a)$ provides a score for amino acid a at position p in the context of treatment t. For instance, $R_t(p, a)$ can express the resistance amino acid a at position p causes to treatment t. The initial state has a cost $R_t(0)$ which corresponds to the offset

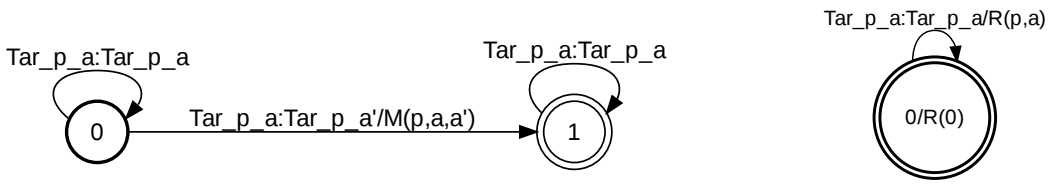


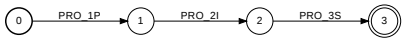
Figure 6.3: Concept of mutation and rating transducers. The notation of the transducers is the same as in previous figures. The mutation transducer (left) for each treatment has two states, connections between these states introduce mutations. The edge weight corresponds to the score of the mutation as represented by the function $M_t(p, a, a')$. The rating transducer (right) has one loop edge for every position and amino acid, the edge weight $R_t(p, a)$ can for example correspond to the weights for an amino acid at a certain position as derived from the linear SVM models, which predict *in vitro* drug resistance. The weight of the state $R_t(0)$ then corresponds to the offset of the linear model z_0 (see Eqn. 6.2).

of that rating function. After computing the composition of a linear acceptor representing a sequence (short: sequence acceptor) with such a rating transducer, each edge in the sequence acceptor receives a score. The sum of all scores along a path constitutes the activity of a treatment against the sequence represented by the path. Rating transducers can for instance be used to remove ambiguities from an input sequence. In our case, ambiguities are by parallel edges between subsequent states in the sequence acceptor. In the composition of the sequence acceptor and the rating transducer these edges are extended by a score corresponding to the amino acid at the alignment position. Thus, one can arrange that the shortest path in the transducer corresponds to the most resistant variant represented by the sequence.

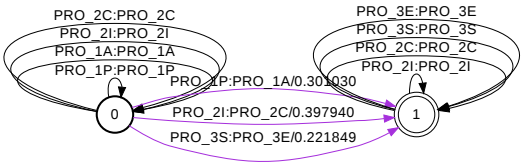
So far the transducers only represent mutation or rating instructions for a single treatment. However, it is trivial to extend them to allow mutation and rating instructions for different treatments within one automaton. This allows to store all models within a single large transducer instead of multiple small ones. Figure 6.5 extends the toy example with a second treatment in the mutation transducer. For selecting the correct mutation function, the first edge in the sequence acceptor is labeled with the corresponding treatment.

6.3 Mutation Models

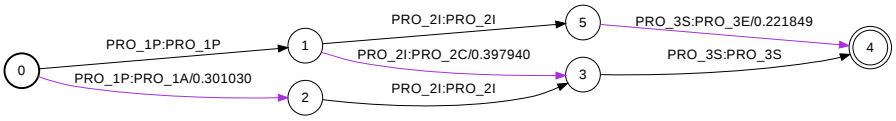
The major problem with building the mutation transducer is the estimation of correct mutation probabilities (or scores). The following two subsections present approaches that use either *in vitro* or *in vivo* data for estimating mutation probabilities. The proposed models all aim at assessing mutation probabilities with respect to individual drugs (exceptions are the RTI mutation probabilities estimated from *in vivo* data). This approach has a drawback since mutation models for individual drugs have to be combined in some way



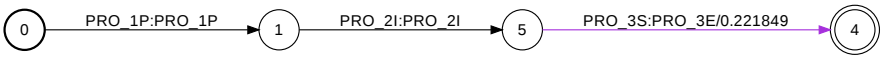
(a) sequence acceptor



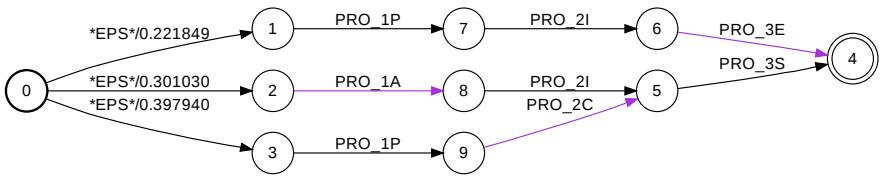
(b) mutation transducer



(c) composition

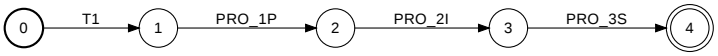


(d) best result

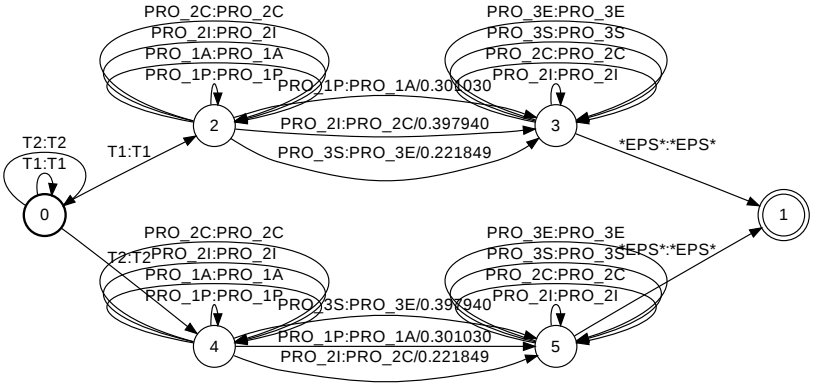


(e) 3-best result

Figure 6.4: Toy transducer example. HIV sequences are represented by linear acceptors (a), the mutation transducer for each treatment comprises two states, with the connecting edges encoding the mutations (b). The composition introduces every possible mutation into the sequence (c). The most likely mutation can be retrieved via finding the shortest path (d). Likewise the n most likely mutations can be extracted by listing the n-best paths.



(a) sequence acceptor



(b) mutation transducer

Figure 6.5: Toy transducer example for multiple treatments. The sequence information is preceded by an edge with a treatment label (a), the complete mutation transducer comprises two states for each treatment, the initial state connects to the pre-mutation state via an edge labeled with the treatment (b).

for building the M_t functions that model mutation probabilities for a specific treatment. On the other hand, having mutation models for individual drugs available, enables the computation of therapy mutation model for every possible combination of drugs. If not stated otherwise the therapy mutation models are constructed as geometric mean of the mutation models for individual drugs

$$M_t(p, a, a') = \prod_{d \in t} M_d(p, a, a')^{\frac{1}{|t|}},$$

where d are the drugs used in therapy t and $|t|$ corresponds to the number of drugs in the therapy. Since the transducers are defined over the tropical semiring this is equivalent to

$$\log M_t(p, a, a') = \frac{1}{|t|} \sum_{d \in t} \log M_d(p, a, a').$$

Of note, the framework also supports the use of a weighted geometric mean:

$$M_t(p, a, a') = \prod_{d \in t} M_d(p, a, a')^{w_{d,t}} \Leftrightarrow \log M_t(p, a, a') = \sum_{d \in t} w_{d,t} \log M_d(p, a, a') \tag{6.1}$$

with $\sum_{d \in t} w_{d,t} = 1$.

6.3.1 Mutation Probabilities Derived From GENO2PHENO

The first model is motivated by the work of Beerenwinkel et al. (2003b) and assumes that mutations, which confer the most dramatic change in resistance with respect to the current regimen, are selected first during therapy. Unlike the work by Beerenwinkel et al., the model used here does not make use of the "maximum assumption" between drugs of the same class (see Section 4.1.1), but it simply takes into account the change of resistance against all drugs when a mutation is introduced. The SVM models in GENO2PHENO use a linear kernel to predict the decadic logarithm of the resistance factor; this allows for rewriting the regression function as a linear model (see Section 3.1). In order to simplify further steps we rewrite the linear model $\vec{w} \cdot \vec{x} + b$ for predicting phenotypic drug resistance as:

$$\log\left(\mathrm{rf}(\mathsf{seq})\right) = z_0 + \sum_{p=1}^{N} \sum_{a} z_{p,a} * \delta(a, \mathsf{seq}(p)), \tag{6.2}$$

where z_0 is the offset (before b), $z_{p,a}$ is the weight for amino acid a at position $p \in \{1, \ldots, N\}$ (extracted from the corresponding position in $\vec{w}$), and $\delta(a, \mathsf{seq}(p))$ is 1 only if the query sequence presents amino acid a at position p (representing the feature vector $\vec{x}$). Because of this structure the weights $z_{p,a}$ can be used directly to construct the mutation function $\log M(p, a, a')$. Precisely, the mutation score is defined by the difference in predicted resistance between a sequence seq and a mutant at position p exchanging amino acid a with a' ($\mathsf{seq}_{p,a\to a'}$):

$$\begin{aligned}
\log M(p, a, a') &= \log\left(\mathrm{rf}(\mathsf{seq})\right) - \log\left(\mathrm{rf}(\mathsf{seq}_{p,a\to a'})\right) \\
&\overset{(6.2)}{=} z_0 + \sum_{p=1}^{N} \sum_{a} z_{p,a} * \delta(a, \mathsf{seq}(p)) \\
&\quad - \left(z_0 + \sum_{p=1}^{N} \sum_{a} z_{p,a} * \delta(a, \mathsf{seq}_{p,a\to a'}(p)) \right) \\
&= z_{p,a} - z_{p,a'}.
\end{aligned}$$

The use of the raw resistance scores, however, introduces a problem since different drugs show different ranges of resistance levels. Thus, for minimizing range-related problems, the models were scaled to exhibit similar resistance levels for each drug. The scaling works in two steps and focuses on the bimodal distribution of resistance factors (Figure 3.3): first mean μ and variance σ of the susceptible subpopulation were determined for each drug and the resistance factor was transformed to $\log(\mathrm{rf}') = \frac{\log(\mathrm{rf}) - \mu}{\sigma}$. Second, the values were scaled (i.e. divided by a scalar s) such that the mean of the resistant subpopulation μ_{resist} has the same value for all drugs. Note that the first transformation is similar to the z-scores used in GENO2PHENO (see Section 3.2). Using this scaling the score of a mutation is

$$\log M(p, a, a') = \frac{z_{p,a} - z_{p,a'}}{\sigma \cdot s} \tag{6.3}$$

and a combination of Eqn. 6.1 and Eqn. 6.3 yields for arbitrary therapies

$$\log M_t(p, a, a') = \sum_{d \in t} w_{d,t} \frac{z_{d,p,a} - z_{d,p,a'}}{\sigma_d s_d}$$

with d being a drug in treatment t, $w_{d,t}$ standing for the weight of the drug in the treatment ($\sum_{d \in t} w_{d,t} = 1$). All other variables receive and additional index d indicating that they are specific for an individual drug. This mutation model is termed g2p_{raw}.

An alternative model, also derived from GENO2PHENO, deduces $M(p, a, a')$ from the probability that a susceptible sequence becomes resistant after acquiring an additional mutation. Precisely, given a set of sequences, let S be the number of all susceptible sequences, $S_{p,a} \leq S$ be the number of all susceptible sequences with amino acid a at position p, and $R_{p,a \to a'} \leq S_{p,a}$ be the number of sequences that move to the resistant subpopulation after mutating a to a'. Hence

$$M(p, a, a') = \frac{R_{p,a \to a'}}{S_{p,a}} \tag{6.4}$$

expresses the probability that a sequence becomes resistant after mutating a to a'. Obviously, $0 \leq M(p, a, a') \leq 1$ holds and no scaling for the scores is required for achieving compatibility between different drugs. Here, we define the cutoff between susceptible and resistant viruses as the intersection of the Gaussians modeling the distribution of RF values for susceptible and resistant viruses (Figure 3.3). The probability of transcending the boundary from susceptible to resistant by accumulating one mutation clearly depends on the level of resistance the virus currently exhibits and the magnitude of resistance conferred by that additional mutation. Hence, the potential of a certain mutation is estimated from a dataset by averaging over all sequences. The resulting $M(p, a, a')$ scores are dominated by the magnitude of change in resistance conferred by the mutation from a to a' and are therefore highly correlated with the corresponding scores of g2p_{raw}. For instance, a mutation that increases phenotypic resistance by 10 fold can turn a larger fraction of susceptible viruses into resistance ones than a mutation that increases resistance only by 2 fold. Consequently the former mutation will have a larger $M(p, a, a')$ score. Approximately 30,000 HIV-1 *pol* sequences containing protease and reverse transcriptase from the EUResist database were used to estimate the probabilities. This model is termed $\text{g2p}_{\text{transition}}$. Figure 6.6 depicts scatter plots between $\log M(p, a, a')$ from g2p_{raw} (x-axis) and $\text{g2p}_{\text{transition}}$ (y-axis) for all drugs.

The scatter plots show clearly that there is a substantial correlation between the two mutation scores. However, the plots reveal also that for some drugs the assumption, that the most resistance conferring mutation will be selected first during treatment, may lead to the preferred accumulation of rather rare mutations (e.g. RT mutation Q151M for ddI and d4T). For smoothing the probabilities with respect to the chance of a mutation to actually emerge Eqn. 6.4 is modified to

$$M(p, a, a') = \frac{R_{p,a \to a'}}{S_{p,a}} * \frac{R_{p,a'}}{R},$$

with R and $R_{p,a}$ representing the number of all resistant sequences and the number of resistant sequences having amino acid a' at position p, respectively. Thus, $R_{p,a'}/R$ expresses the probability of observing a certain mutation in the resistant subpopulation. This mutation model is termed $\text{g2p}_{\text{mixed}}$. Figure 6.7 depicts scatter plots between $\log M(p, a, a')$ from g2p_{raw} (x-axis) and $\text{g2p}_{\text{mixed}}$ (y-axis) for all drugs.

For practical reasons the models were restricted to allowing only mutations from and to amino acids that actually occur in the training data of GENO2PHENO at a given position.

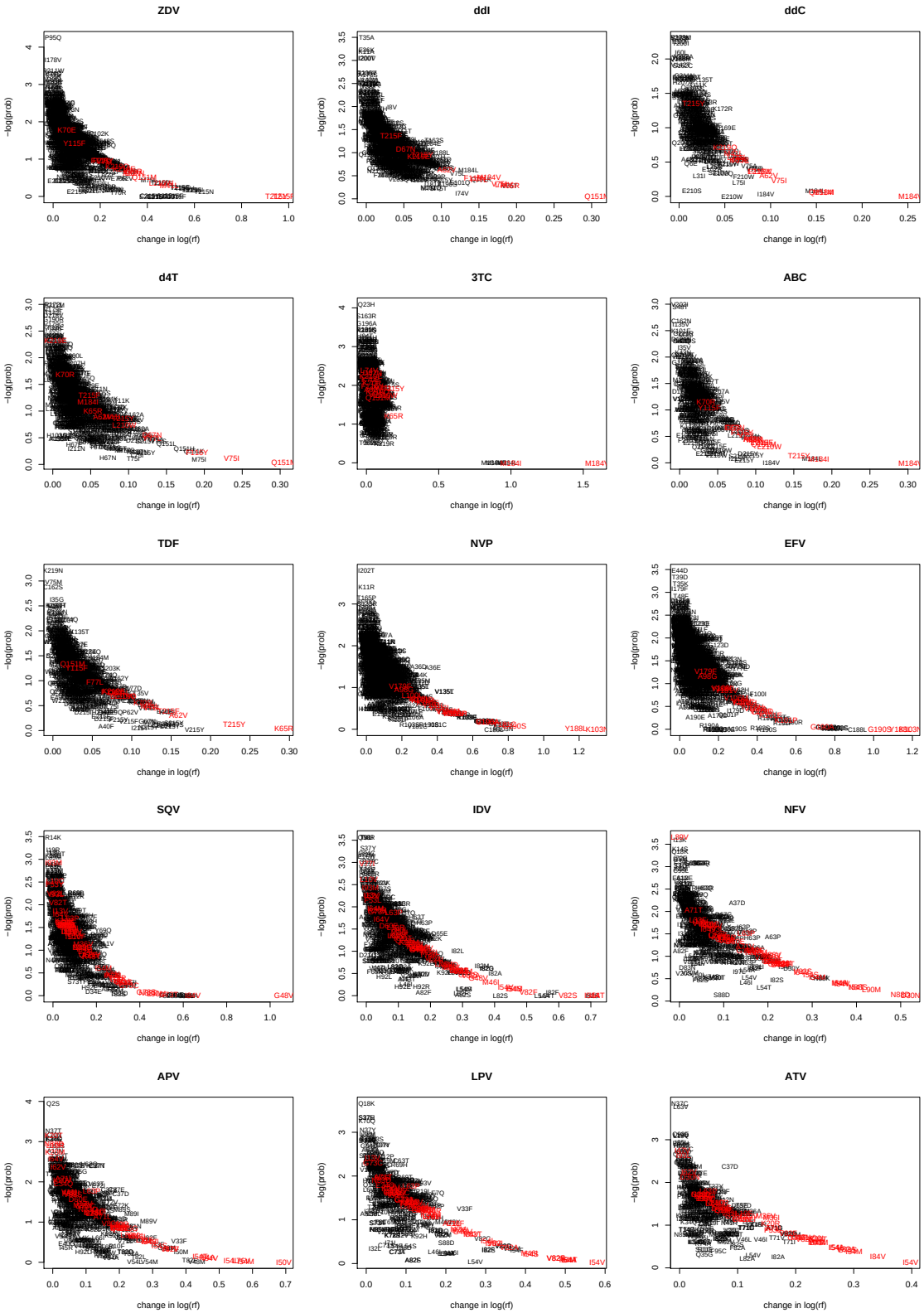


Figure 6.6: Scatter plots between mutation scores of $g2p_{raw}$ (x-axis) and $g2p_{transition}$ (y-axis) for different drugs. Mutations from the wild type amino acid to a mutation listed by the IAS list (Johnson et al., 2008) for the corresponding drug class are colored red.

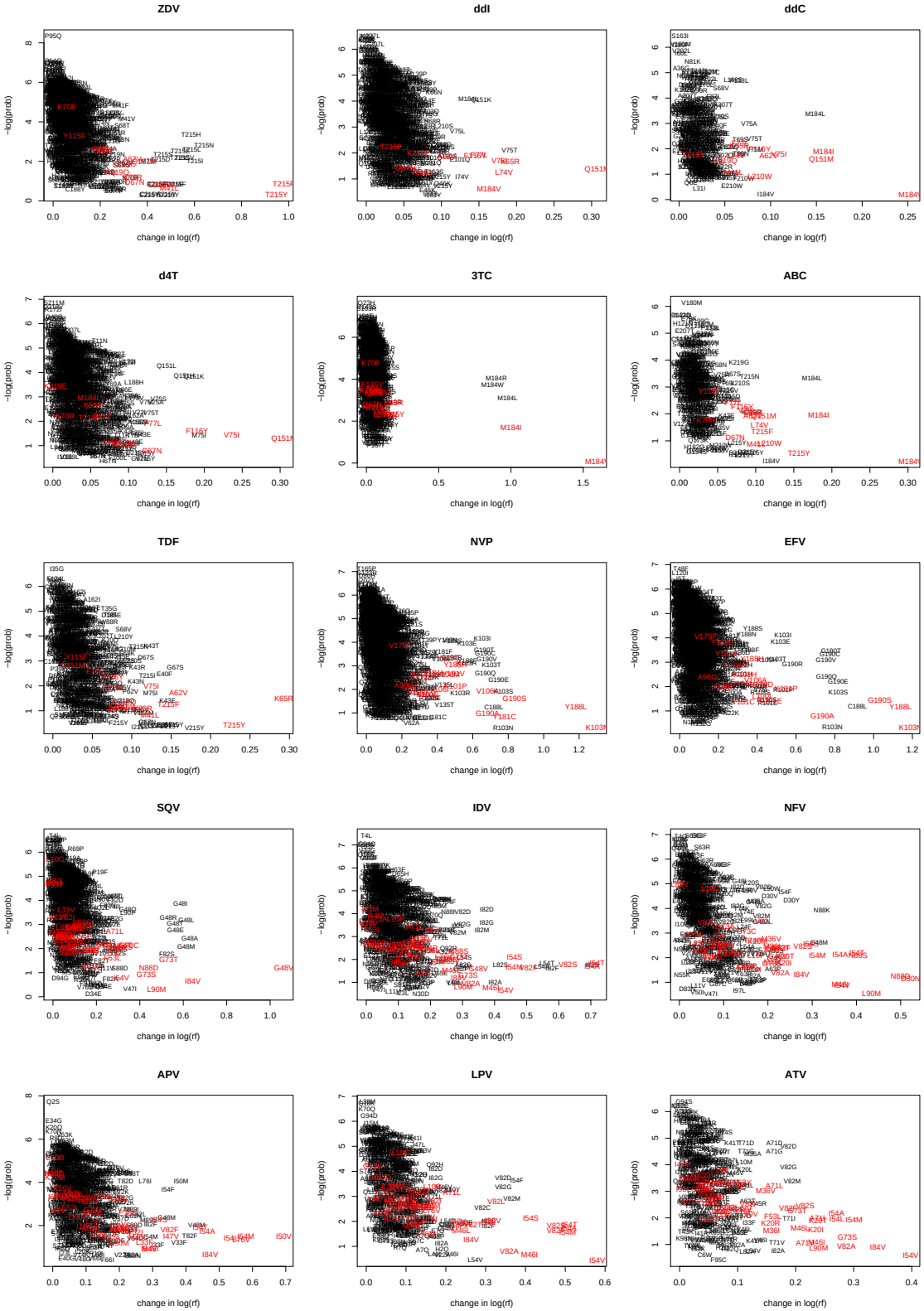


Figure 6.7: Scatter plots between mutation scores of $g2p_{raw}$ (x-axis) and $g2p_{mixed}$ (y-axis) for different drugs. Mutations from the wild type amino acid to a mutation listed by the IAS list (Johnson et al., 2008) for the corresponding drug class are colored red.

Moreover, the $g2p_{mixed}$ model excludes mutations that did not appear among resistant sequences from the EuResist database by definition, since $R_{p,a'} = 0 \Rightarrow M(p, a, a') = 0$.

6.3.2 Mutation Probabilities Derived From Treatment Data

The previously introduced mutation models assume (to a varying extent) that mutations conferring the largest increase of resistance as measured *in vitro* are selected first during treatment. Certainly, this assumption does not hold for real treatments. In fact, the order of mutations can also be expected to be influenced by restrictions imposed by the protein structure. For example, a mutation might result in an impossible side chain conformation and therefore requires a prior mutation in the vicinity that may relax the side chain placement. In the case of the HIV integrase, it was suggested that different pathways may be induced by alternative binding conformations of the inhibitor to the target protein (Loizidou et al., 2009). These circumstances are not respected by *in vitro* drug resistance. A perfect mutation model should consider all these aspects. However, as briefly mentioned in the previous section there are some rarely occurring mutations that confer a large change in resistance. Unfortunately, the mechanisms underlying these rare events are not yet understood and, in general, knowledge on how and why a certain mutation is preferred over another one is rather scarce.

Instead of integrating all these different (and probably incomplete) sources of knowledge, one can study the history of the virus for constructing a mutation model. The mutagenetic trees introduced in Chapter 4 rely on the fact that in a large database of viral sequences every step of a resistance pathway occurs sufficiently often. Briefly, if mutation X is rarely observed alone but rather frequently in the presence of mutation Y, one can assume that Y has to be acquired before X. The success of the mutagenetic trees demonstrates that relying on large databases for deriving evolutionary models is a promising alternative. Obviously, longitudinal sequence data from monotherapies constitute the ideal source of information for estimating *in vivo* mutation probabilities. Monotherapies, however, are considered insufficient regarding today's standard of care and are therefore not enriched in large databases. Nowadays standard therapies typically comprise multiple drugs from different drug classes. The majority of these treatments, however, uses only one protease inhibitor or one NNRTI, thus even modern therapies can be regarded as pseudo-monotherapies. Precisely, from the approximately 100,000 therapies stored in the EuResist database 99.6% and 95.6% of the NNRTI- and PI-containing regimens, respectively, apply only one representative of the corresponding class. In contrast, only 15.8% of the recorded treatments are pseudo NRTI monotherapies. NRTIs are given mostly in combination and consequently 73.5% of the NRTI regimens contain exactly two NRTIs. These statistics suggest that in general it should be possible to deduce *in vivo* mutation probabilities for individual drugs (PIs and NNRTIs) or at least for drug pairs (NRTIs) from large databases. The bottleneck of this estimation is the availability of sequence data, as at least one sequence is required before a treatment change (baseline genotype) and one during the treatment (follow-up genotype). Figure 6.8 depicts a schematic description of the training data requirements and a graph relating the time allowed between baseline genotype and treatment start to the size of the resulting training set.

Beyond a threshold of 180 days, the amount of available training sequences grows only

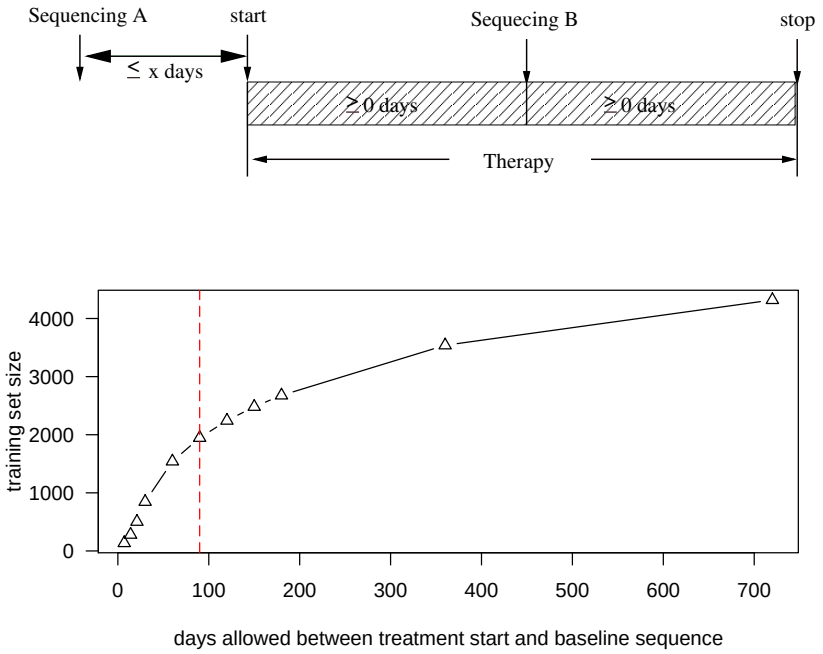


Figure 6.8: The top figure shows a schematic description of the training data requirements. The baseline sequence has to be obtained at most x before start of the treatment, while the follow-up sequence may be obtained at any time during the treatment. The lower graph depicts the relationship between x and the size of the dataset.

slowly with respect to the time frame allowed between obtaining the baseline sequence and treatment start. The downside of a large time frame is that the viral population is under continuous selective pressure by the old treatment and thus might accumulate additional mutations before start of the next treatment. From this point of view, a small time frame is preferred. Here, however, the downside is clearly the lack of available sequence data. For complying with the standard datum definition the cutoff of 90 days was selected and gave rise to approximately 1,900 sequence pairs. For exploiting the wealth of the EuResist database a cutoff of 720 days was investigated as well. These datasets provide for every treatment the applied drugs, the baseline sequence, the follow-up sequence, and the three time points. For estimating the *in vivo* mutation rates, the available information for a sequence pair is represented as an itemset. On this dataset, itemset mining is applied to discover interesting rules, e.g. rules in the form of "drug → mutation". The confidence of a rule serves as the mutation probability. The plain application of an itemset mining algorithm unfortunately retrieves a number of artifacts, for instance, rules stating that zidovudine causes NNRTI related mutations with a high confidence. These artifacts mainly result from the co-administration of several compounds and have to be removed in a postprocessing step. Furthermore, the approach assumes that RTIs do not influence mutation probabilities of positions in the protease and vice versa. This allows to treat the estimation of protease and RT mutations as two independent problems and consequently reduces the amount of artifacts. The following two sections describe the itemset mining

and subsequent filtering in detail. The resulting mutation models are termed vivo_{90} and vivo_{720}.

Mining Frequent Mutations

In frequent itemset mining instances are represented as sets containing a subset of N binary attributes termed *items*. Due to historic reasons, each instance is called *transaction* $t \subseteq \{i_1, i_2, \ldots, i_N\}$ and a collection of transactions is referred to as database $D = \{t_1, t_2, \ldots, t_M\}$. In our case, each drug applied in a regimen is represented by one item. Furthermore, every difference between the baseline sequence and the follow-up sequence is encoded as an item. For limiting the number of items to be considered in total, only the mutations in the baseline sequence with respect to the reference HXB2 were encoded as single items (instead of all positions of the baseline sequences). Ambiguities at positions were resolved, i.e. $a_1, \ldots, a_j$ at one position in the baseline sequence was encoded as j items. The same holds true for encoding the differences between baseline and follow-up sequence.

Given a large database, one wants to find sets of items that frequently occur together in transactions, i.e. item sets whose *support* exceeds a certain threshold. The support of an itemset X is defined as the number of times all items of X occur together in the same transaction in the database: $\text{supp}(X) = |\{m | X \subseteq t_m\}|$. Frequent itemsets can be used to derive rules $X_1 \Rightarrow X_2$, with $X_1, X_2 \subset X \wedge X_1 \cap X_2 = \emptyset$. The *confidence* of a rule is defined as

$$\text{conf}(X_1 \Rightarrow X_2) = \frac{\text{supp}(X)}{\text{supp}(X_1)}.$$

These rules are termed association rules and can be interpreted as $P(X|X_1)$ (Agrawal et al., 1993).

The best-known algorithm for finding frequent itemsets is the Apriori algorithm (Agrawal and Srikant, 1994). Apriori exploits the fact that

$$X_1 \subseteq X \Rightarrow \text{supp}(X_1) \geq \text{supp}(X). \tag{6.5}$$

In consequence, this means that if X_1 is not considered a frequent itemset, then a set containing X_1 can never be a frequent itemset, and also that if X is a frequent itemset, then all subsets of X are frequent itemsets, too. The algorithm limits the search space by first finding all frequent n-itemsets ($|X| = n$) and based on these defines candidates for frequent $n+1$-itemsets. Precisely, if X_1, X_2 are frequent n-itemsets and $X_1 \cap X_2 = n - 1$, then $X_1 \cup X_2$ is a candidate for a frequent $n + 1$-itemset. The non-frequent candidate sets are then discarded by checking if all subsets of size n are frequent. The candidate generation and validation, however, can be quite costly in terms of computation.

The *frequent pattern* (FP) tree mining algorithm avoids the candidate generation step (Han et al., 2004). The FP tree mining algorithm first represents the database D as an FP tree. In essence, an FP tree is a prefix tree with counts of occurrences of a prefix stored in the node representing that prefix. For efficient tree traversal the FP tree maintains an item header table pointing to an item's occurrences in the tree via a linked list. Figure 6.9 depicts an example comprising nine transactions and the resulting FP tree including its header table.

List of itemIDs	reordered list
{I1, I2, I5}	(I2, I1, I5)
{I2, I4}	(I2, I4)
{I2, I3}	(I2, I3)
{I1, I2, I4}	(I2, I1, I4)
{I1, I3}	(I1, I3)
{I2, I3}	(I2, I3)
{I1, I3}	(I1, I3)
{I1, I2, I3, I5}	(I2, I1, I3, I5)
{I1, I2, I3}	(I2, I1, I3)

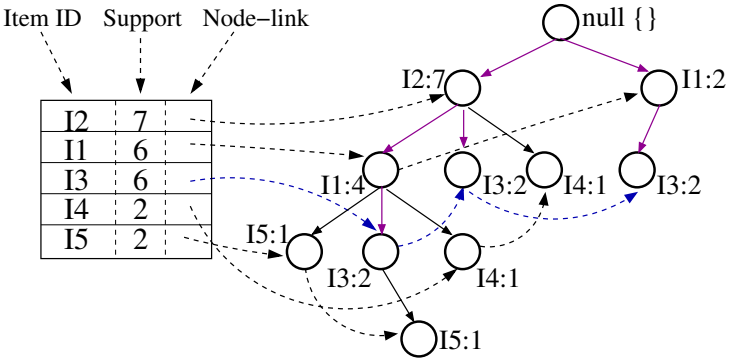


Figure 6.9: The left table lists a database comprising 9 transactions. In the right column the items I1,. . . ,I5 are ordered with respect to their frequency in the database. The FP tree represents exactly this database. The header table allows for an efficient retrieval of all occurrences of an item within the tree. This is needed to construct the conditional pattern base. The colored edges indicate the links one has to traverse for constructing the conditional pattern base for I3. Example adapted from Kamber and Han (2001).

For the FP tree construction the items within every transactions are first ordered in decreasing order of their support in the whole database. Of note, after the reordering the itemsets are no longer sets but item tuples. The root node of the tree is labeled with the empty set, then each transaction t_m is inserted into the tree in the following way: starting from the root the tree is traversed following the order imposed by the items in transaction t_m, if no branch exists then it is created and the counter at the target node is initialized with 0, after the last item of t_m the count on each visited node is increased by one. The task of finding frequent patterns in the database is now transformed to mine frequent patterns in the FP tree.

The FP tree is mined recursively. The process starts with constructing the *conditional pattern base* for each frequent 1-itemset. The frequent 1-itemsets can easily be determined from the header table. The conditional pattern base for an item i_n is simply given by the itemsets on all paths from nodes labeled with i_n to the root node with the support set to the count stored in the i_n node. For example, the conditional pattern base for I3 in Figure 6.9 is {(I2 I1: 2), (I2: 2), (I1: 2)} as indicated by the colored solid edges. The occurrences of I3 in the tree can easily be found by following the linked list (colored dashed edges). Another FP tree is constructed for the conditional pattern base and mined for frequent patterns. The end of the recursion is reached when either the FP tree for the conditional pattern base is empty, i.e. comprises only the root node, or the FP tree comprises only a single path. In the latter case, all combinations of nodes on that single path are (parts of) a frequent pattern. Algorithm 5 provides pseudo code for the recursive FPGROWTH function (adapted from Kamber and Han (2001)).

Due to the excessive computational requirements of Apriori, the FP tree mining was used to find frequent mutations caused by protease and reverse transcriptase inhibitors. A pattern was considered frequent in the case of protease mutations and reverse transcriptase mutations if it occurred 12 and 20 times, respectively. These thresholds were manually set

Algorithm 5 FPGROWTH($Tree, \alpha$)

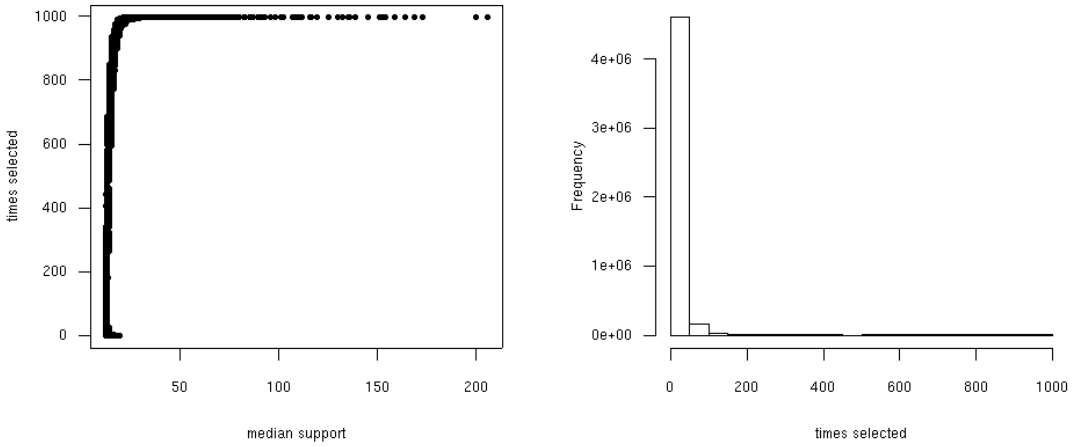
```
if Tree contains a single path P then
  for each combination (denoted as β) of the nodes in path P do
    generate pattern β ∪ α with supp(β ∪ α) = minimum support of any node in β
  end for
else
  for each α_i in the header of Tree do
    generate pattern β = α_i ∪ α with supp(β) = supp(α_i)
    construct β's conditional pattern base and β's conditional FP_tree: Tree_β
    if Tree_β ≠ ∅ then
      FPGROWTH(Tree_β, β)
    end if
  end for
end if
```

and represent a trade-off between robustness of mined rules and the number of distinct rules. For estimating the mutation probability and its variance 1,000 bootstrap samples of the original dataset were generated.

Postprocessing of Mutation Rules

For protease and reverse transcriptase, executing the FP tree mining algorithm results in 3323 and 2283 (61,498 and 24,968) frequent patterns on average per bootstrap replicate, respectively, using a cutoff of 90 (720) days. A large fraction of these frequent patterns is not of interest for our application. Combining results from the bootstrap replicates by taking the union of the discovered patterns and limiting the list of frequent patterns to itemsets containing at least a drug and one additional mutation (i.e. difference between baseline sequence and follow-up sequence) yields 171,993 and 21,124 (4,844,840 and 217,042) potentially interesting patterns for protease and RT, respectively, using the 90 (720) days cutoff. Figure 6.10 a) depicts the relationship between the median support (x-axis) and the number of bootstrap replicates, in which a pattern was considered frequent (y-axis) for the protease dataset using the 720 days cutoff. As expected, the majority of the patterns was selected only in very few bootstrap replicates (Figure 6.10 b)), e.g 105,032 and 4117 patterns were selected once and at least 200 times, respectively, on the protease 90 days dataset. However, at first, all frequent patterns were rewritten as rules. In general, a frequent n-itemset can be written as $\sum_{i=1}^{n-1} \binom{n}{i}$ different rules. In our application, however, there is only one meaningful rule per pattern. Precisely, an item of a pattern is either a compound, a baseline mutation, or an additional mutation, thus the only interesting rule has drugs and baseline mutations on the left side and additional mutations on the right side. This allows for interpreting the confidence of a rule as the probability of developing a certain mutation (or a list of mutations) given a drug and a set of baseline mutations.

Unfortunately, the application of an association rule mining algorithm on our kind of data generates a large number of artifact rules originating from the frequent co-administration of drugs. For example, since the year 2000 the activity of protease inhibitors is increased by administering a boosting dose of ritanovir (RTV) that occupies the cytochrome P450-



(a) support vs. bootstrap selection

(b) histogram bootstrap selection

Figure 6.10: Scatter plot between the median support and the times a pattern was selected in a bootstrap replicate (a). Histogram of the times a pattern was selected from the 1000 bootstrap replicates (b). Both figures are derived from the protease 720 days cutoff dataset.

mediated metabolism in the human liver (Kumar et al., 1996) and therefore allows improved plasma levels of drugs that undergo the same metabolism (typically protease inhibitors). As a consequence, today RTV (or RTVb for boosting dose) is part of every PI containing regimen and indeed, approximately 70% of PI containing treatments in the dataset list RTV as one of the drugs, thus many mined rules claim that RTV is the cause for certain mutations. This is most certainly not the case, however, and corresponding rules have to be filtered out. For the reverse transcriptase there exist similar cases, lamivudine (3TC) for example is used in approximately 60% of the NRTI containing therapies and a number of rules accordingly blame 3TC for causing a great variety of mutations. In fact, 3TC is only associated with mutations at positions 65 and 184 (Johnson et al., 2008). On the other hand, other RTIs are connected with the appearance of mutations at position 184 since due to the relation in Eqn. 6.5 all frequent itemsets containing 3TC imply a frequent itemset without 3TC. This example demonstrates that inferring which RTI causes which RT mutation is a challenge. The subsequent application of the mutation models, however, does not require the estimation of models for individual drugs. Especially the models for RTIs will have to be combined, while the PI models are independent to a greater or lesser extent. Therefore, solving the inference of causality for RTI models can be avoided by estimating models for combinations of RTIs directly. The challenges for a postprocessing filter differ between the drug targets. The following two paragraphs describe the filtering steps that were applied to remove most of the artifacts from the list of protease and reverse transcriptase mutation rules. In the following the left side and right side of a rule are denoted by X_l and X_r, respectively. The set of PIs and RTIs is denoted by C_{pro} and

C_{rt}, respectively, with $C = C_{\text{pro}} \cup C_{\text{rt}} \cup$ "empty″. The set of all baseline and additional mutations is denoted by E_{base} and E_{add}, respectively.

Protease Rules Filter: All rules containing RTV as the only PI are most likely artifacts and are therefore filtered: $|\text{RTV} \cap X_l| = 1 \wedge |C_{\text{pro}} \cap X_l| = 1$; the first part of the expression checks whether RTV is on the left side of the rule, while the second expression checks whether there is only one PI listed on the left side of the rule. Moreover, since not all regimens contain a PI, there are many transactions for the protease task that contain the item "empty″. These transactions represent pseudo-treatment breaks for the protease. Rules with "empty" as only compound on the left side are interesting, since they provide information on how resistance mutations disappear from the predominant viral population when there is no selective pressure on the protease. For this task, however, these rules constitute an artifact and have to be removed: $|\text{"empty}'' \cap X_l| = 1$. In the last step, all Patterns $(X_l \cup X_r)$ that were not selected a sufficient number of times in the bootstrap analysis were discarded.

Reverse Transcriptase Rules Filter: In a first step, all rules that only contain one drug are discarded ($|X_l \cap C| = 1$), this also includes the rare case of "empty" as the only listed compound. In contrast, all rules comprising three or more drugs are retained ($|X_l \cap C_{\text{rt}}| > 2$). Finally, all rules comprising exactly two RTIs have to undergo a further filtering step. Briefly, if two drugs are said to cause a certain mutation, then there is the chance that actually a third drug (not part of the rule) is the true cause. This case obviously occurs when three drugs are frequently administered together. For example, Atripla is available as a single pill and comprises three compounds, two NRTIs (FTC+TDF) and one NNRTI (EFV). As a consequence one obtains the artifact rule {FTC, TDF} ⇒ {K103N}. The mutation is clearly an NNRTI related mutation and therefore most likely caused by EFV. Given a rule $X_l \Rightarrow X_r$ containing exactly two drugs $\{c_1, c_2\} = X_l \cap C_{\text{rt}}$, then $E'_{\text{base}} = X_l - \{c_1, c_2\}$ is the set of baseline mutations. The filter checks for every RTI that is not part of the rule $x \in C_{\text{rt}} - \{c_1, c_2\}$ whether the rules $\{c_1, x\} \cup E'_{\text{base}} \Rightarrow X_r$ and $\{c_2, x\} \cup E'_{\text{base}} \Rightarrow X_r$ exist. If for any drug x both rules exist and both exhibit a significantly higher confidence than the original rule, then the original rule is discarded. For assessing whether the difference in confidence is significant a t-test based on mean and standard deviation of the confidence obtained from the bootstrap replicates can be applied. In accordance with the protease filter, all patterns that were not selected a sufficient number of times in the bootstrap analysis were discarded as well.

From Rules to Mutation Models

Applying the described filters with a bootstrap selection threshold of 20% for both sets results in 1188 and 454[1] (12,572 and 3861) rules for protease and RT, respectively, using the 90 (720) days cutoff. These rules have two major advantages over the mutation scores derived from *in vitro* measured resistance. Firstly, the rules reflect the mutation rates *in vivo*, and secondly, the presence of baseline mutations on the left side of the rules allows to model mutation rates in dependence of pre-existing mutations, rather than independently.

[1] 10% minimum selection threshold

An interesting example for rules obtained with the 720 day cutoff is given by the rather novel protease inhibitor tipranavir (TPV). The rule mining discovered three major TPV related protease mutations 33F, 82T, and 84V, with probability 0.27, 0.63, and 0.38, respectively. The mutation at position 82, however, does not occur from the wild type amino acid (Valine), but from a mutant (Alanine) that was probably selected during previous treatment with another PI. For this mutation there exist two rules in the list: {TPV} ⇒ {A82T} and {TPV, 82A} ⇒ {A82T} with confidence 0.28 and 0.63, respectively. For constructing the mutation model one has to take into account that A82T is not a mutation from wild type and therefore use the confidence of the rule listing the corresponding baseline mutation on the left side.

While in theory one can construct transducers that represent mutation models with an arbitrary number of baseline and additional mutations, their construction algorithm can be rather complex. For instance, in order to model the mutation probability in presence of baseline mutations, the mutation transducer must have a path that is different from the two states in Figure 6.3 and comprises at the position of the baseline mutation only the corresponding amino acid. Moreover, one has to distinguish between baseline mutations that occur before and after the follow-up mutation in the sequence (regarding the alignment position). For example, for extending the mutation transducer in Figure 6.4 with differences in the mutation probability of position two in the presence of mutations at position one or three, one has to introduce two new states (2 and 3). One new edge connects the initial state with state 2, its input and output label is "PRO_1A" and the weight is equal to the mutation probability of position 2 in presence of 1A. State 2 and the final state are connected via an edge with input label and output label "PRO_2I" and "PRO_2C", respectively. The same labeling is used for the edge between the initial state and state 3. Finally, state 3 is connected to the final state using an edge with input and output label "PRO_3E" and has the altered mutation probability as weight. Moreover, in theory, states 2 and 3 need loop edges for all alignment positions between the baseline and the follow-up mutation. As can be easily seen, the size of the transducer grows rapidly with respect to the number of baseline mutations that are represented. For practical reasons we restrict the transducer construction to rules that provide the probability of a single mutation, i.e. $|X_r| = 1$, and with at most one baseline mutation. The construction algorithm is simplified by the fact that for the vast majority of rules $\text{conf}(X_l \Rightarrow X_r) \leq \text{conf}(X_l \cup E'_{\text{base}} \Rightarrow X_r)$ holds (see Figure 6.11 a)). Thus, one simply has to add an alternative path featuring the elevated probability in addition to the normal mutation probability. In accordance with *in vitro* mutation models we define mutation models derived from treatment data as

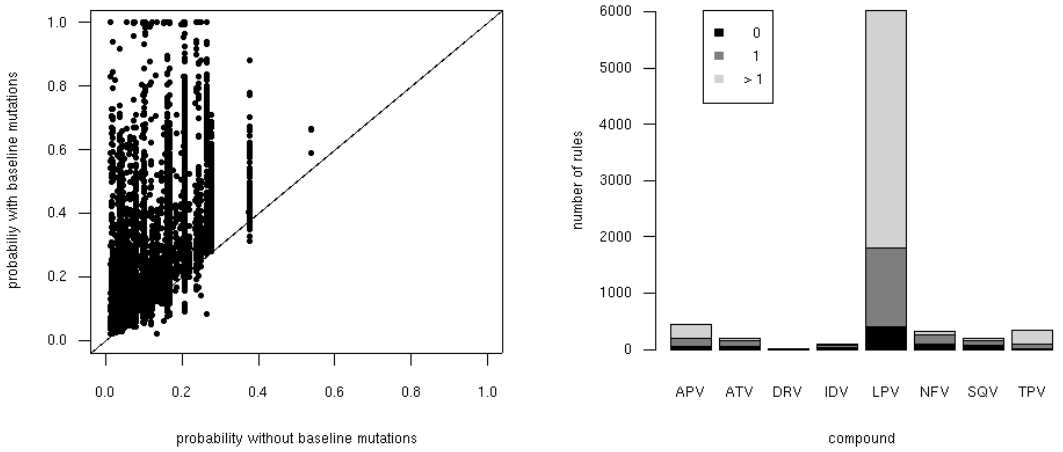
$$M_{d,m}(p, a, a') = \max\{\text{conf}(\{d, m\} \Rightarrow \{apa'\}), \text{conf}(\{d, m, pa\} \Rightarrow \{apa'\})\},$$

where m is a single baseline mutation, and d is either a single drug (like above) or a combination of drugs. Hence, for complete treatments we have

$$M_{t,m}(p, a, a') = \prod_{t' \subseteq t} M_{t',m}(p, a, a')^{-k},$$

with $k = |\{t' | t' \subseteq t \wedge M_{t'}(p, a, a') \neq 0\}|$.

Figure 6.11 b) depicts a the number of rules per PI derived from the 720 days cutoff dataset. The figure also clearly demonstrates the existence of a bias towards frequently



(a) effect of baseline mutations

(b) rules per protease inhibitor

Figure 6.11: Scatter plot between the probability for one mutation with and without the presence of baseline mutations given the same drug (a). Final number of rules for different protease inhibitors, different grey scales indicate the number of baseline mutations in those rules (b). Both figures were derived from the protease 720 days cutoff dataset.

used drugs that are obviously captured better by the model. For example, the majority of rules describe mutation development under LPV/r containing treatments. Moreover, the figure indicates that by restricting the mutation models to at most one baseline mutation the majority of discovered rules is exploited (except for LPV).

6.4 Validation

Finding a validation scenario for an FDO score is a challenging task. In contrast to the prediction of treatment response where one has the viral load as a direct measure, there is no useful covariate in the HIV treatment databases, which necessarily has to correlate with a good FDO. For instance, it is unlikely that a suboptimal choice regarding FDOs in the first treatment will significantly shorten the patient's life as so many other factors most likely have a more observable impact, e.g. adherence. Moreover, the group of patients that are followed from their first therapy till death is rather small and most likely did only have very limited alternatives for the first treatment.

Nonetheless, there are possibilities for (at least) validating the mutation models alone. The following two sections describe approaches for estimating the quality of the introduced mutation models.

6.4.1 Prediction of Response to the Next But One Treatment

In previous chapters we aimed at predicting the virological response to the patient's upcoming treatment. To this end we extracted retrospective treatment change episodes from a database collecting routine treatment data. Then, statistical learning was used to correlate information about the virus and the treatment to the patient's virological response to the antiretroviral combination treatment. The availability of a mutation models allowing to evolve the viral population during a treatment makes a modification of this initial setting possible. Knowing the genotype of a patient shortly before one treatment, one can estimate the additional mutations at end of that treatment, and predict response to the subsequent treatment based on this simulated genetic makeup of the viral population. This scenario allows to study the impact of parameters of the mutation models. First, one can study which mutation model provides the basis for the best predictions. Second, one can study whether a fixed or a variable number of newly introduced mutations provides better results. Third, one can assess the benefit of using the n-most likely results of the *in silico* evolution over the single best variant. The following sections provide the experimental setup, the results, and the discussion of this validation scenario.

Experimental Setup

Figure 6.12 illustrates the requirements for TCEs that are used for this validation. Briefly, each TCE comprises two treatment changes. A genotype has to be available at most 90 days prior to the start of treatment A. In analogy to the EuResist standard datum definition for assessing response to the treatment, a baseline viral load and follow-up viral load have to be available at most 90 days before and about eight weeks (4 to 12 weeks) after onset of treatment B, respectively. Application of this modified TCE definition to the EuResist integrated database (release 2008/10/10) yields 1952 instances. A treatment success is defined by suppressing the viral load below the limit of detection (400 copies of viral RNA per ml blood) or achieving at least a 100-fold reduction compared to the baseline value. Of note, the response to treatment A is of no interest, thus an instance is considered successful if treatment B is a successful treatment. Using this definition, 1395 instances are considered to be successful and 557 are considered treatment failures. The instances selected for this validation overlap with the instances used for training the EuResist prediction model. Precisely, 1428 TCEs and 350 TCEs of the EuResist training data correspond to treatment A and treatment B, respectively. Since we predict neither the response to treatment A nor use the real viral genotype obtained shortly before treatment B for our predictions, we can consider this dataset as an independent test set.

Each of the sequences in the dataset was represented by a linear acceptor. Ambiguities in the amino acid sequence were encoded as alternative edges between two states. In a first step only those amino acids in ambiguities conferring the most resistance against the current treatment were retained (using a rating transducer as in Figure 6.3). Then, novel mutations were introduced into the sequence by applying all five mutation models described above. The number of newly introduced mutations is a parameter. We follow three strategies for setting the number of mutations. The first strategy introduced a fixed number of mutations into each sequences and we vary this number from 0 to 10, with 0 representing the reference model that does not introduce any mutation. The second approach

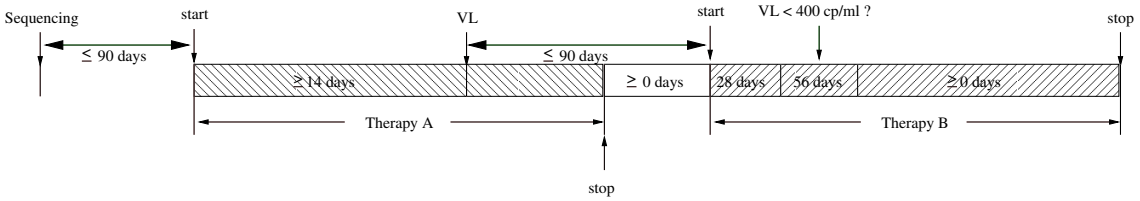


Figure 6.12: Modified TCE definition. Every TCE comprises two treatments. The sequence had to be obtained at most 90 days before treatment A. A baseline viral load measurement had to be available at most 90 days before treatment B and a follow-up viral load measure had to be available within 4 to 12 weeks after onset of the therapy. A TCE is successful if treatment B is successful, i.e. viral load is reduced 100-fold or below 400 copies of viral RNA per ml (limit of detection).

produces a number of mutations that depends on the composition of the antiretroviral therapy. More precisely, the number of mutations is simply based on the average *potency* of the regimen. Here, we use the instantaneous inhibitory potential (IIP) after 24 hours of the maximal concentration of the drug in the plasma (IIP_{24}) as defined by Shen et al. (2008). The IIP_{24} ranges from 0 for d4T to 8.4 for DRV/r. The number of mutations based on the average potency *pot* is computed by

$$n(pot) = \lceil \alpha \cdot \exp(1 - pot) \rceil, \text{ with } \alpha \in \{1, \ldots, 5\}. \tag{6.6}$$

If $n(pot)$ is less than 1 or exceeds 10 mutations, then $n(pot)$ is set to 1 or 10, respectively. The parameter α controls the increase of mutations with the decrease of average potency. The exponential function is used to realize a convex function.

The third strategy originates from the fact that an effective treatment gives the virus fewer opportunities to mutate. Thus, the number of mutations is determined by the predicted success probability of treatment A (p_A). Precisely, the number of mutations is calculated as

$$n'(p_A) = \lceil \alpha(\exp(1 - p_A) - 1) \rceil, \text{ with } \alpha \in \{1, \ldots, 5\}. \tag{6.7}$$

Again, the exponential function is applied for realizing a convex function.

Finally, we use a combination of potency and predicted success – the *effect* – for estimating the number of mutations. Precisely, we apply Eqn. 6.6 to $pot \cdot p_A$ instead of *pot* alone.

For models with varying number of mutations we use two different baselines approaches. The first approach generates a random number of mutations that is uniformly sampled from 1 to 10 (randomA). The second method generates a random number of mutations that follows the distribution of the p_A-based model. This is achieved by randomly shuffling the number of mutations computed with Eqn. 6.7 (randomB). For both random models the provided performance measure is the mean of 100 repetitions.

For each of these models we compute the list of 10,000 variants receiving the highest score according to the mutation model.

Our prediction model, which we contributed to the EuResist prediction engine (5.2), is used to predict the response of the 10,000 most likely outcomes of the simulated evolution

	$g2p_{raw}$	$g2p_{transition}$	$g2p_{mixed}$	$vivo_{90}$	$vivo_{720}$
baseline			0.757		
fixed number					
1	0.757	0.764	0.762	0.749	0.757
2	0.749	0.755	0.756	0.743	0.756
3	0.724	0.735	0.741	0.738	0.750
4	0.702	0.714	0.725	0.728	0.735
5	0.694	0.699	0.710	0.718	0.721
6	0.692	0.690	0.696	0.710	0.711
7	0.690	0.674	0.683	0.703	0.705
8	0.687	0.660	0.671	0.697	0.701
9	0.686	0.651	0.660	0.693	0.696
10	0.684	0.645	0.651	0.689	n/a
dependent on current treatment composition					
randomA	0.694	0.687	0.690	0.707	0.717
randomB	0.737	0.738	0.741	0.737	0.745
potency	0.759	0.768	0.768	0.757	0.764
success	0.766	0.767	0.768	0.755	0.765
effect	0.762	0.769	0.768	0.759	0.768

Table 6.1: Performance for different mutation models and mutation numbers measured in AUC.

to the drugs in treatment B. The predicted success for treatment B is the weighted mean of the top n of these 10,000 predictions

$$\frac{1}{\sum_{j=i}^{n} s_j} \sum_{i=1}^{n} s_i p_i, \tag{6.8}$$

where s_i and p_i are the mutation score and the predicted success probability of the ith variant, respectively. For $g2p_{raw}$ s_i corresponds to the increase in resistance, and for all other models s_i equals the real probability (not the negative decadic logarithm).

Results

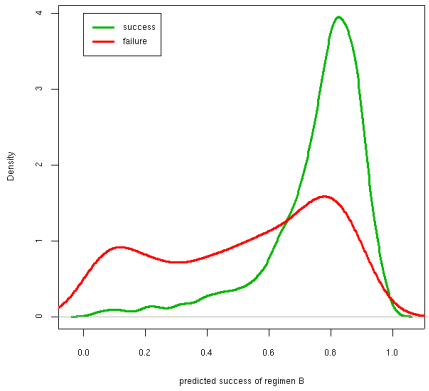
Table 6.1 lists the AUC values that were achieved in this setting with all five mutation models. None of the models manages to substantially improve over the baseline (no mutations) and only subtle differences are visible between the mutation models. A general trend is the decline of prediction performance with increased fixed number of mutations. Consequently, the introduction of only one mutations provides the highest AUCs.

The introduction of a variable number of mutations depending on the treatment composition of regimen A provides a slight improvement over the baseline (no mutations). Moreover, all three proposed ways of estimating the number of mutations improve substantially over models generating a random number of mutations. Here, the uniform sampling provides the worst results and is probably an underestimation of the baseline.

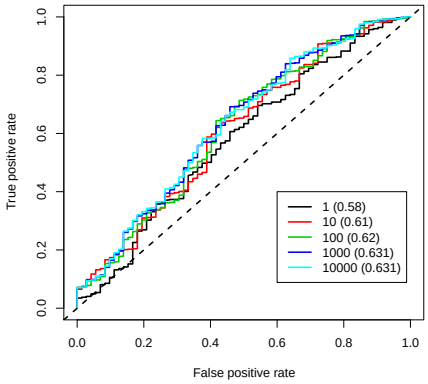
Still, none of the mutation models paired with an approach to estimate the number of mutations provides a substantial improvement over the baseline. One reason for this observation is doubtless the large fraction of treatments having a low success rate with respect to regimen B even before regimen A caused changes in the viral population. Figure 6.13 a) depicts the distribution of predicted success rate according to regimen B (p_B) at baseline for the viral sequence obtained before onset of regimen A. Interestingly, the distribution resembles the one that can be expected for p_A as seen in Figure 5.7 a). It becomes obvious that a large number of treatments failing regimen B had a bad perspective of succeeding even before regimen A was administered. Hence, the baseline of zero mutations is almost insurmountable and little improvement is visible by the mutation models. For a fair comparison of the mutation models, the distributions of the predicted success rate in the failing and succeeding groups should be identical for regimen B before onset of regimen A.

In order to have an ideal starting point for the validation of the mutation models, the distribution of p_B should be identical in failing and succeeding cases. Thus, for studying a distribution of p_B that is more similar in failing and succeeding cases, we applied a threshold on p_B and considered only cases exceeding that threshold. The cutoff was varied from 0.0 (all cases) to 0.9 in steps of 0.1. Moreover, ambiguous failures (i.e. treatments labeled as failure but achieving a VL reduction below 500 copies at least once during the treatment; see Section 5.3) were removed for obtaining a more robust quality assessment. Figure 6.13 c) displays the decline in AUC for increasing cutoffs of p_B for the baseline (no mutations), different mutation models with variable number of mutations based on p_A, and both random number of mutations models (using $\text{g2p}_{\text{mixed}}$). This setting reveals qualitative differences among the g2p-based mutation models and the mutation models derived from treatment data. All three g2p-based models perform the best and achieve an AUC of approximately 0.630 compared to the baseline of 0.547 at a p_B cutoff of 0.8. At this threshold, the vivo_{90} model is indistinguishable from the baseline with an AUC of 0.556, while the vivo_{720} model reaches a slightly better AUC of 0.602 and places itself between the baseline and the g2p-based models. The results also demonstrate the importance of linking the right number of mutations to the right treatment. The performance of randomB is very close to the baseline, while $\text{g2p}_{\text{mixed}}$ with p_A-based number of mutations demonstrates clearly superior prediction accuracy. Potency- or effect-based computation of the number of mutations produces similar curves (data not shown). Of note, only 10 cases were labeled as failure when the threshold of 0.9 was applied. Due to the low number of failing cases, results obtained using the highest cutoff cannot be considered reliable. For instance, the increase observed for both random models when moving from the 0.8 to the 0.9 cutoff are most likely an artifact.

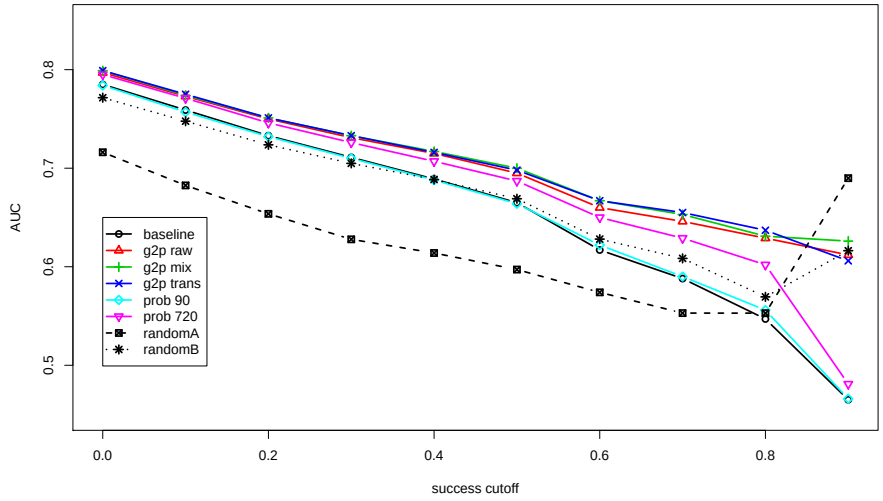
For investigating the impact of the number of most likely viral variants considered for the computation of the success for the next but one regimen, we chose the p_B cutoff of 0.8, the $\text{g2p}_{\text{mixed}}$ model, p_A-based number of mutations, and varied the number of considered *in silico* variants: $\{1, 10, 100, 1000, 10000\}$. Figure 6.13 b) depicts the corresponding ROC curves. Clearly, it is beneficial to consider more than the single most likely variant: the top 10 yield already an improvement of 0.03 in AUC. Nevertheless, no improvement was observed beyond the top 1000, which increased the AUC by 0.051 compared to the single most likely variant.



(a) Distribution of p_B



(b) Impact of varying top n variants



(c) Development of AUC

Figure 6.13: (a) Distribution of p_B before onset of regimen A in failing treatments (red) and successful treatments (green). (b) Change of AUC depending on the considered top n viral variants. AUC computed with g2p$_{mix}$ mutation model at cutoff 0.8 using. (c) Development of AUC after restricting the data in terms of p_B. Apart for the random models, the number of mutation was based on p_A.

Discussion

The results demonstrated that it is possible to infer response to the treatment following the next treatment on the basis of the current genotype. Unfortunately for our validation, the baseline (i.e. no additional mutations) provided already a very good prediction performance. Indeed, the performance was close to AUC values observed for predicting response to the next antiretroviral regimen (see results in Chapter 5). The high performance was the result of a large number of cases that had bad perspectives for succeeding with regimen B even before regimen A caused additional mutations. Consequently, the mutation models showed only little improvement.

A scenario that is more suitable for assessing the quality of the mutation models is the separation of failing and succeeding regimens that had the same (or at least similar) distribution of success rate p_B before the onset of regimen A. Here, mutations selected by regimen A are the cause for the later failure, and the mutation models should allow for a separation. The removal of all instances with p_B below a certain cutoff served as way for matching the distributions of p_B in successful and failing regimens.

The altered scenario revealed differences between mutation models: g2p-based models generally outperform the models estimated from treatment data. Moreover, the importance of assigning the right number of mutations to the treatments was made evident and we could demonstrate a clear improvement over the baseline (no mutations). Of note, the performance of the baseline can be seen as a measure of dissimilarity of the distribution of p_B in both groups (if the distributions were identical, then the baseline would achieve an AUC of 0.5).

6.4.2 Comparing Predicted Future Drug Options to Data From Real Cases

The second validation scenario aims at comparing predicted future drug options after a treatment to real future drug options observed at treatment failure in patients. In order to avoid problems with archived drug resistance mutations in patients, which had already received antiretroviral therapies, this scenario is restricted to treatment naïve patients. That restriction has another advantage, as it allows us to assume that the virus before the treatment is a wild type virus and therefore we only need to retrieve sequences of the patient at end of the first treatment serving for the computation of future drug options. Consequently, this maximizes the amount of available data for this analysis. The simplification bears also one risk: by assuming that a patient was infected with a wild type virus, we ignore the existence of conferred drug resistance mutations.

Validation Setup

The EuResist database was queried for HIV genotypes that were obtained during the first treatment the patient ever received. More precisely, the sequence comprising RT and protease had to be obtained 30 days after start earliest and 14 days before stop of the treatment latest, respectively. As argued earlier (Section 4.2), the ability to sequence the virus while the patient is on antiretroviral treatment indicates virological failure. In order to ensure that the patient indeed received the first treatment, it was required that all treatments, which the patient had ever received, were stored in the database. Table 6.2 lists

all antiretroviral treatments that met the requirements and occur at least three times. For all sequences fold-change in resistance against 17 drugs was computed with GENO2PHENO and the activity score for each drug was determined (Section 3.2). The FDO was simply defined as the sum of the activity for all 17 drugs, thus the FDO was in the range of $[0, 17]$. In addition to the FDO for all drugs, drug class specific FDOs were calculated.

Unfortunately, a number of sequences (about 5.9%) suggested complete resistance against a large number of drugs from all classes. Given that the sequences were obtained at the end of the first antiretroviral treatment, this was an unexpected result and most likely originates from faulty entries in the database or cases of transmitted drugs resistance. Hence, we excluded all cases that exhibit an FDO score of 5.0 or less from the analysis. Table 6.3 lists FDO and drug class wise FDO after removal of the cases.

Predictions for the FDOs were generated by constructing a linear acceptor representing the consensus B wild type sequence. Then, for every treatment listed in Table 6.3 a mutation model was used to introduce a fixed number of mutations (1 to 10), and the 10,000 variants receiving the highest mutation score were stored. For simplicity, the mutation model was restricted to $\text{g2p}_{\text{mixed}}$, which also ranked among the best models in the previous setting. Next, for all predicted variants the FDO was computed. The consensus FDO of the (at most) 10,000 variants was computed using a weighted mean (see Eqn. 6.8).

Finally, we correlated the predicted FDOs with the observed FDOs. Here, however, the number of mutations that have to be introduced into the wild type sequence is unknown. Thus, to get an upper estimate on the prediction performance we selected the number of mutations for each treatment that best fits the observed value (best fit). In addition, we used a crude heuristic for estimating the number of mutations (manual): the number of mutations to be introduced is 2, treatments containing boosted PIs introduce one mutation less, treatments with less than 3 drugs one mutation more, and so do treatments containing the combination d4T+ddI.

Results

Table 6.4 lists the obtained correlation for a varying number of top N most likely viral variants. For the best fit approach the correlation steadily increases with larger viral population and reaches up to 0.951 for the full list of 10,000 variants. Using the crude manual heuristic for inferring the number of mutations the correlation between predicted FDO and actual FDO decreases to approximately 0.570.

Figure 6.14 shows scatter plots of the real FDO against the predicted FDO using the 100 most likely variants. A surprising observation is that using the best fit approach a small number of mutations is sufficient to achieve a high correlation with the observed values. Precisely, the majority of treatments are best modeled when only one, two, or three mutations are introduced. In order to verify whether the drug classes are correctly modeled, we correlated the observed class specific FDO to the predicted class specific FDO. Here, we used the number of mutations determined with best fit approach on the FDO for all drugs. The correlations was very high for NRTIs and NNRTIs, reaching 0.809 and 0.796, respectively. The PIs, however, showed with 0.407 only an acceptable correlation. A confounder of the FDOs for PIs are treatments that did not contain any PIs. Consequently, removal of non-PI treatments elevated the correlation to 0.599. Figure 6.14 c) shows the

drug combination	count	all	NRTI	NNRTI	PI
ZDV	100	13.94 (3.39)	4.80 (2.21)	2.66 (0.69)	6.49 (1.70)
ddC+ZDV	44	14.13 (3.17)	4.83 (2.33)	2.82 (0.32)	6.49 (1.63)
3TC/FTC+ZDV	40	13.01 (2.49)	3.49 (2.13)	2.64 (0.77)	6.88 (0.34)
3TC/FTC+LPV+ZDV	38	15.55 (1.97)	6.17 (1.30)	2.88 (0.24)	6.50 (0.99)
3TC/FTC+NFV+ZDV	38	10.70 (3.99)	3.67 (1.63)	2.64 (0.69)	4.38 (2.60)
3TC/FTC+NVP+ZDV	34	13.11 (3.13)	4.60 (2.07)	1.63 (1.23)	6.88 (0.25)
3TC/FTC+IDV+ZDV	31	13.42 (3.39)	5.04 (1.80)	2.83 (0.27)	5.55 (2.27)
3TC/FTC+EFV+ZDV	24	12.17 (4.77)	4.62 (2.32)	1.48 (1.39)	6.07 (1.95)
3TC/FTC+LPV+TDF	23	14.15 (4.82)	5.94 (1.80)	2.38 (1.01)	5.83 (2.35)
3TC/FTC+ABC+ZDV	20	12.01 (3.13)	2.60 (2.60)	2.68 (0.77)	6.73 (0.72)
ddI+ZDV	20	12.74 (2.59)	3.23 (2.24)	2.55 (0.88)	6.97 (0.06)
3TC/FTC+SQV+ZDV	19	12.09 (3.61)	3.82 (1.90)	2.76 (0.67)	5.51 (2.36)
3TC/FTC+d4T+NFV	15	10.37 (4.89)	3.44 (2.34)	2.89 (0.18)	4.03 (2.87)
3TC/FTC+EFV+TDF	11	11.65 (5.42)	4.47 (2.44)	1.67 (1.44)	5.52 (2.18)
3TC/FTC+d4T	11	11.39 (1.67)	2.02 (1.24)	2.53 (0.86)	6.84 (0.34)
d4T+ddI+NFV	9	10.58 (4.31)	3.07 (2.29)	2.89 (0.07)	4.62 (2.80)
3TC/FTC+d4T+EFV	7	13.04 (2.10)	5.02 (1.74)	1.64 (1.24)	6.38 (1.02)
3TC/FTC+d4T+NVP	7	14.38 (2.73)	5.85 (1.48)	1.62 (1.50)	6.91 (0.15)
ddC+SQV+ZDV	7	14.62 (2.59)	5.32 (1.73)	2.86 (0.19)	6.43 (1.27)
3TC/FTC+d4T+SQV	6	9.25 (4.44)	2.16 (1.79)	2.96 (0.03)	4.12 (3.29)
d4T+ddI+EFV	6	12.12 (3.62)	5 (2.35)	0.66 (1.19)	6.46 (1.17)
3TC/FTC+APV/FPV+TDF	5	12.65 (6.95)	4.95 (2.88)	2.11 (1.21)	5.60 (3.13)
3TC/FTC+ATV+TDF	5	13.59 (2.58)	4.78 (1.59)	2.97 (0.04)	5.84 (1.60)
3TC/FTC+ABC+LPV	4	15.46 (2.26)	5.96 (1.59)	2.81 (0.22)	6.69 (0.47)
3TC/FTC+d4T+IDV	4	12.69 (5.34)	4.79 (2.36)	2.93 (0.09)	4.97 (3.35)
3TC/FTC+d4T+LPV	4	13.38 (6.31)	5.41 (2.98)	2.72 (0.52)	5.25 (3.50)
d4T+ddI	4	9.90 (3.58)	3.43 (2.74)	2.98 (0.02)	3.48 (4.02)
d4T+ddI+NVP	4	11.25 (2.72)	4.48 (2.42)	0.86 (1.44)	5.91 (2.16)
ddI+NVP+ZDV	4	11.60 (2.42)	2.88 (1.68)	1.79 (0.97)	6.92 (0.15)
3TC/FTC	3	12.57 (0.39)	3.23 (0.39)	2.85 (0.25)	6.49 (0.88)
3TC/FTC+NVP+TDF	3	13.95 (3.92)	5.39 (2.55)	1.96 (1.70)	6.60 (0.42)
d4T+ddI+IDV	3	14.28 (2.22)	5.59 (1.30)	1.69 (1.53)	7 (0.00)
d4T+ddI+SQV	3	10.56 (5.78)	3.72 (2.98)	2.96 (0.04)	3.88 (3.48)
ddI+EFV+TDF	3	9.15 (4.99)	4.89 (1.60)	0.00 (0.00)	4.26 (3.41)

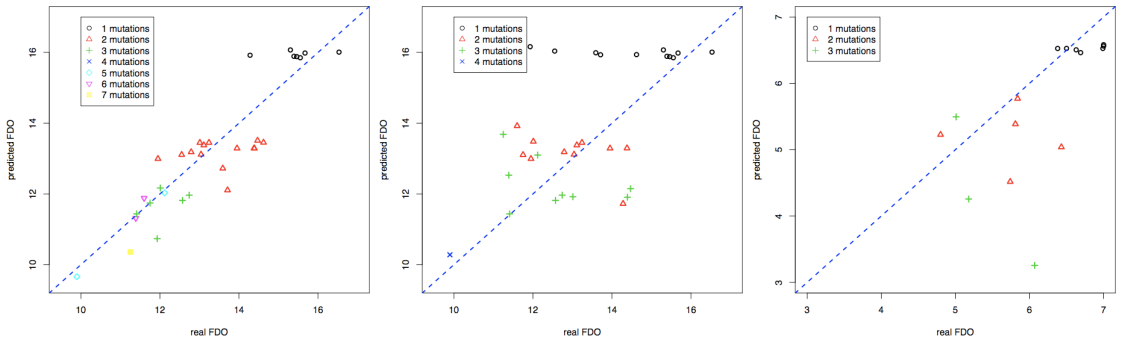
Table 6.2: First line antiretroviral treatments and observed future drug options (FDO). The column count states the sample size that served to estimate the FDO. The column all, NRTI, NNRTI, and PI state the FDOs for all drugs, NRTIs, NNRTIs, and PIs, respectively. Values are the mean among all observed cases and standard deviation is state in parenthesis.

drug combination	count	all	NRTI	NNRTI	PI
ZDV	95	14.47 (2.55)	4.98 (2.10)	2.67 (0.69)	6.82 (0.87)
ddC+ZDV	43	14.39 (2.70)	4.94 (2.24)	2.81 (0.33)	6.64 (1.31)
3TC/FTC+ZDV	40	13.01 (2.49)	3.49 (2.13)	2.64 (0.77)	6.88 (0.34)
3TC/FTC+LPV+ZDV	38	15.55 (1.97)	6.17 (1.30)	2.88 (0.24)	6.50 (0.99)
3TC/FTC+NVP+ZDV	34	13.11 (3.13)	4.60 (2.07)	1.63 (1.23)	6.88 (0.25)
3TC/FTC+NFV+ZDV	33	11.75 (3.06)	4.04 (1.41)	2.70 (0.57)	5.01 (2.17)
3TC/FTC+IDV+ZDV	30	13.71 (3.05)	5.10 (1.79)	2.87 (0.20)	5.74 (2.06)
3TC/FTC+EFV+ZDV	22	13.24 (3.26)	5.01 (1.99)	1.61 (1.38)	6.62 (0.64)
3TC/FTC+LPV+TDF	21	15.39 (2.62)	6.40 (0.97)	2.60 (0.72)	6.38 (1.54)
3TC/FTC+ABC+ZDV	20	12.01 (3.13)	2.60 (2.60)	2.68 (0.77)	6.73 (0.72)
ddI+ZDV	20	12.74 (2.59)	3.23 (2.24)	2.55 (0.88)	6.97 (0.06)
3TC/FTC+SQV+ZDV	18	12.55 (3.06)	3.83 (1.95)	2.91 (0.14)	5.81 (2.01)
3TC/FTC+d4T+NFV	12	11.95 (4.08)	4.28 (1.78)	2.87 (0.20)	4.80 (2.68)
3TC/FTC+d4T	11	11.39 (1.67)	2.02 (1.24)	2.53 (0.86)	6.84 (0.34)
3TC/FTC+EFV+TDF	10	12.79 (4.09)	4.89 (2.11)	1.83 (1.40)	6.07 (1.25)
d4T+ddI+NFV	8	11.41 (3.77)	3.34 (2.28)	2.88 (0.07)	5.18 (2.37)
3TC/FTC+d4T+EFV	7	13.04 (2.10)	5.02 (1.74)	1.64 (1.24)	6.38 (1.02)
3TC/FTC+d4T+NVP	7	14.38 (2.73)	5.85 (1.48)	1.62 (1.50)	6.91 (0.15)
ddC+SQV+ZDV	7	14.62 (2.59)	5.32 (1.73)	2.86 (0.19)	6.43 (1.27)
d4T+ddI+EFV	6	12.12 (3.62)	5 (2.35)	0.66 (1.19)	6.46 (1.17)
3TC/FTC+ATV+TDF	5	13.59 (2.58)	4.78 (1.59)	2.97 (0.04)	5.84 (1.60)
3TC/FTC+ABC+LPV	4	15.46 (2.26)	5.96 (1.59)	2.81 (0.22)	6.69 (0.47)
3TC/FTC+APV/FPV+TDF	4	15.67 (1.89)	6.18 (0.96)	2.49 (0.99)	7 (0.00)
3TC/FTC+d4T+SQV	4	11.93 (1.97)	2.90 (1.76)	2.96 (0.04)	6.07 (1.68)
d4T+ddI	4	9.90 (3.58)	3.43 (2.74)	2.98 (0.02)	3.48 (4.02)
d4T+ddI+NVP	4	11.25 (2.72)	4.48 (2.42)	0.86 (1.44)	5.91 (2.16)
ddI+NVP+ZDV	4	11.60 (2.42)	2.88 (1.68)	1.79 (0.97)	6.92 (0.15)
3TC/FTC	3	12.57 (0.39)	3.23 (0.39)	2.85 (0.25)	6.49 (0.88)
3TC/FTC+NVP+TDF	3	13.95 (3.92)	5.39 (2.55)	1.96 (1.70)	6.60 (0.42)
3TC/FTC+d4T+IDV	3	15.30 (1.38)	5.76 (1.63)	2.91 (0.10)	6.63 (0.62)
3TC/FTC+d4T+LPV	3	16.53 (0.63)	6.90 (0.05)	2.63 (0.59)	6.99 (0.01)
d4T+ddI+IDV	3	14.28 (2.22)	5.59 (1.30)	1.69 (1.53)	7 (0.00)

Table 6.3: First line antiretroviral treatments and observed future drug options (FDO). The column count states the sample size that served to estimate the FDO. The column all, NRTI, NNRTI, and PI state the FDOs for all drugs, NRTIs, NNRTIs, and PIs, respectively. Values are the mean among all observed cases and standard deviation is state in parenthesis.

top N	1	10	100	1,000	10,000
best fit	0.784	0.876	0.907	0.943	0.951
manual	0.148	0.351	0.561	0.588	0.572

Table 6.4: Pearson correlation between predicted and actual FDOs. Top N indicates how many most likely variants were used for computing the predicted drug options. The two methods for finding the number of mutations to be introduced are *best fit* and *manual*.



(a) 100 variants (best fit) (b) 100 variants (manual) (c) 100 variants protease (best fit)

Figure 6.14: Scatter plot between real FDO and predicted FDO using 100 most likely variants. FDO for all 17 drugs, and the number of mutations was set using the best fit approach (a) or the manual heuristic (b). FDO for PIs only, the number of mutations was set using the best fit approach on all 17 drugs (c).

corresponding scatter plot for only PI containing regimens.

Discussion

The best fit approach explains the future drug options very well, while requiring only a small number of additional mutations for the majority of treatments to explain the observed FDO value. Moreover, the best fit (in terms of mutations) regarding the FDO of all drugs reflects also the class-wise FDOs with correlations raging from 0.599 to 0.809, indicating that the underlying mutation model reflects the resistance development in the different drug targets *in vivo* to a satisfying extent.

Indeed, the measured FDOs exhibit a high variance themselves, most likely owing to differing treatment lengths, differing patient adherence, and also differing mutations in the virus preexisting at treatment start. Clearly, these sources of uncertainty add to the difficulty of fitting a single FDO to a single drug combination. Consideration of the other factors, however, would either lead to overfitting the model to the noisy data (e.g. introduction of a treatment length parameter) or unreliable estimations of the FDOs due to low sample numbers (e.g. requirement of baseline genotype). Moreover, for other external factors like adherence no data are available.

We made a further assumption which does certainly not hold: every patient started with a wild type consensus B virus. As can be seen from the first extraction of data from the database, a substantial number of cases showed an FDO of 5.0 or less and, therefore,

were excluded from the analysis. Moreover, also the cleaned set of instances contains unlikely post-treatment FDOs. For instance, patients treated with ddI+d4T only have only 3.5 active PIs at treatment failure, on average. A change in FDO among PIs after a PI-less combination therapy can only be explained by either preexisting drug resistance mutations or errors in the database, e.g. the treatment contained a PI or stop date of the first treatment is incorrect and the genotype was actually obtained during the following PI-based regimen.

The best fit estimation clearly serves as an upper bound of the model performance. Despite the many sources of uncertainty originating from the gold standard, i.e. the real data, the model explains the observations using only few mutations. And, most importantly, the resistance development in the different drug classes is captured well. Thus, the mutation model provides the means to simulate resistance development in antiretroviral combination treatments, an essential basis for the sequencing of treatments.

6.5 Remarks

We introduced five different mutation models together with a framework for quickly simulating resistance development in HIV against combination treatments. The mutation models demonstrated, to varying extent, that prediction of response to the one after next regimen can be achieved at moderate performance. Moreover, a second validation setting showed that observed future drug options can be explained using one of the models ($g2p_{mix}$) by the introduction of only few mutations. Here, the overall resistance development was reflected well by resistance development within drug classes.

A crucial variable contributing to success and failure of the model is the number of mutations that are introduced by a treatment. The amount of mutations depends on a range of parameters, including activity of the regimen, potency of the drugs in the regimen, duration of the regimen, and the patient's adherence to the regimen. The exact number is therefore hard to estimate. Within the first validation setting, we explored different functions to infer a suitable number of mutations based on activity and potency of the regimen. These approaches showed an improved behavior compared to the introduction of a fixed number of mutations. However, there is clearly room for further improvements, for instance by functions based on adherence of the patient. Of note, inferring the number of mutations from the duration of the regimen alone is not likely to succeed. For example, one can expect that a long treatment duration results in a higher number of mutations. But, on the other hand, a long duration indicates good compliance of the patient and, as a consequence, a sustained treatment success resulting in only few mutations. Thus, only a measure of patient adherence paired with treatment length is likely to succeed.

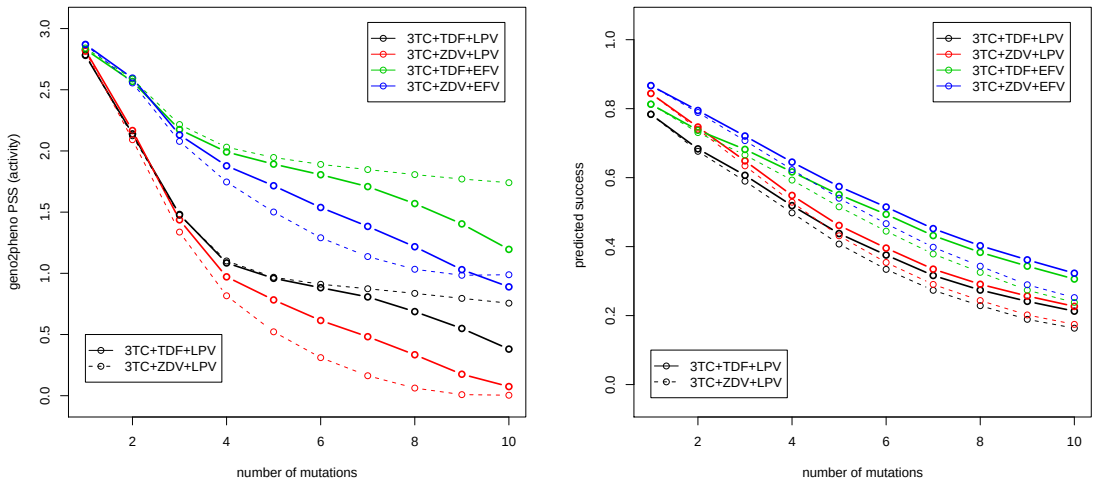
The introduced model can still provide insight on how to sequence treatments even when the number of mutations is unknown. More precisely, one can observe the development of the estimated success probability of potential nth line treatments with respect to $(n-1)$th line regimen. For instance, on a Dual-Core AMD Opteron^{TM} processor with 2.6 GHz and 32 GB memory, the simulation from 1 to 10 additional mutations including storage of the 1,000 most likely variants takes approximately 35 seconds. The interpretation of all variants using the EV EuResist engine requires (due to suboptimal implementation for this task) another 70 seconds. Furthermore, the framework for simulating the evolution is

easily scalable and one can achieve further speedup by limiting the number of variants (6 seconds with 100 variants instead of 1000), the number of different mutations (17 seconds with 5 mutations instead of 10) or a combination of both (3 seconds). Hence, we reached a computational performance that enables the construction of web services. Such a service might allow to upload the sequence of a patient isolate as well as the selection of a few potential treatments and possible follow-up regimens. As a result the service will provide the change in treatment options (using different measures) as a consequence of the selected regimen. In the following we provide a case-study of four typical first-line treatments using three different measures for future treatment options.

Figure 6.15 depicts the development of four different regimens (different colors) in response to two potential first-line regimens (different line styles) in terms of PSS (a) and predicted success by the EV EuResist engine (b). The simulation started with a consensus B wild type sequence, up to 10 mutations were introduced and the 1,000 most likely variants were analyzed. Figure 6.15 a) reveals that up to six mutations there is no difference for TDF-containing regimens (black and green) whether the preceding treatment contained TDF or ZDV. For the ZDV-containing regimens, however, the TDF-containing pre-treatments exhibit a higher score (around 0.3). Hence, first using TDF and then ZDV seems to be a better choice. Also the analysis in terms of estimated success probability (Figure 6.15 b) demonstrates that the use of 3TC+ZDV+LPV provides a slightly worse perspective for future regimens than administration of 3TC+TDF+LPV (i.e. all dashed lines are below solid lines of same color). The benefit of 3TC+TDF over 3TC+ZDV in terms of resistance at treatment failure is well recognized in the medical community, and indeed, TDF-containing regimens are suggested for first-line regimens (http://www.aidsinfo.nih.gov/Guidelines/). The advantage of TDF is that it selects for the mutation K65R, which causes hypersusceptibility of ZDV and other NRTIs (Parikh et al., 2006). Of note, also the applied mutation model assigns a high probability to the appearance of K65R during TDF treatment (see Figure 6.7). The simulation of both treatments (including interpretation) took approximately 3.5 minutes (using only the 100 most likely variants the computational time is reduced to 30 seconds, while maintaining the same qualitative result).

Instead of studying the development of the predicted success rate for a set of fixed regimens, one can also focus on resistance against individual drugs. For instance, Figure 6.16 shows the decrease in FDO as response to the same four drug combinations. By inspecting the FDO for all drugs no difference between the use of TDF or ZDV is visible. When focusing on the FDO in the class of NRTIs, however, the inferiority of ZDV becomes evident, as there is a difference of approximately half a drug with four and five mutations.

Concluding, the transducer-based framework provides the computational means for efficiently simulating resistance development during anti-HIV therapy, and thereby allows the search for optimal orderings of antiretroviral therapies. The computational performance can be even further increased by using the C++ interface instead of the command-line tool; the latter requires to store transducers on the hard drive, while the former allows to store transducers in the main memory and therefore avoid the I/O overhead. Furthermore, the mutation models, on which are simulation depends, can also be based on expert-based rating systems, as opposed to geno2pheno. Likewise, depending on the users' preference the used rating scheme for computing FDOs can be based on expert-based systems.



(a) geno2pheno PSS

(b) success probability

Figure 6.15: Development of geno2pheno PSS (a) and predicted success probability (b) of four different regimens (different colors) after treatment with two typical first-line regimens (different line styles)

6.6 Discussion

Transducers provide a solid framework for simulating development of resistance mutations in HIV. The problem of finding the most likely viral variant, which may emerge during a combination therapy, is modeled as a shortest path problem, in which edge costs are related to the probability of single mutations to emerge during treatment. Due to the application of transducers in the domain of natural language processing (e.g. spoken language recognition and text translation) a number of free and efficient implementations of algorithms that are based on transducers are available. Moreover, the transducer framework can easily be scaled to the available computational hardware. For instance, the number of mutations to be added, the number of most likely viral variants to be tracked, and the amino acid positions of the HIV sequence to be considered (positions have to be simply omitted from the linear input acceptor representing the baseline sequence) can be adapted without modifying the mutation models.

Most importantly though, transducers allow to easily implement different mutation models. In addition to the models presented here, mutation models based on Hidden Markov Models (HMMs) can be realized (see for instance Healy and Degruttola (2007)). The capability of representing HMMs follows from the fact that weighted finite state transducers comprise output labels and weights which can be used to model observation (emissions of the HMM) and probabilities for transitions and emissions, respectively.

A major drawback of using transducers for implementing mutation models is the practical requirement to model mutations independently from each other. That is, each mutation will lead to the same increase in total path cost; independent of the mutations that are already present in the genotype. In theory it is possible to construct mutation transduc-

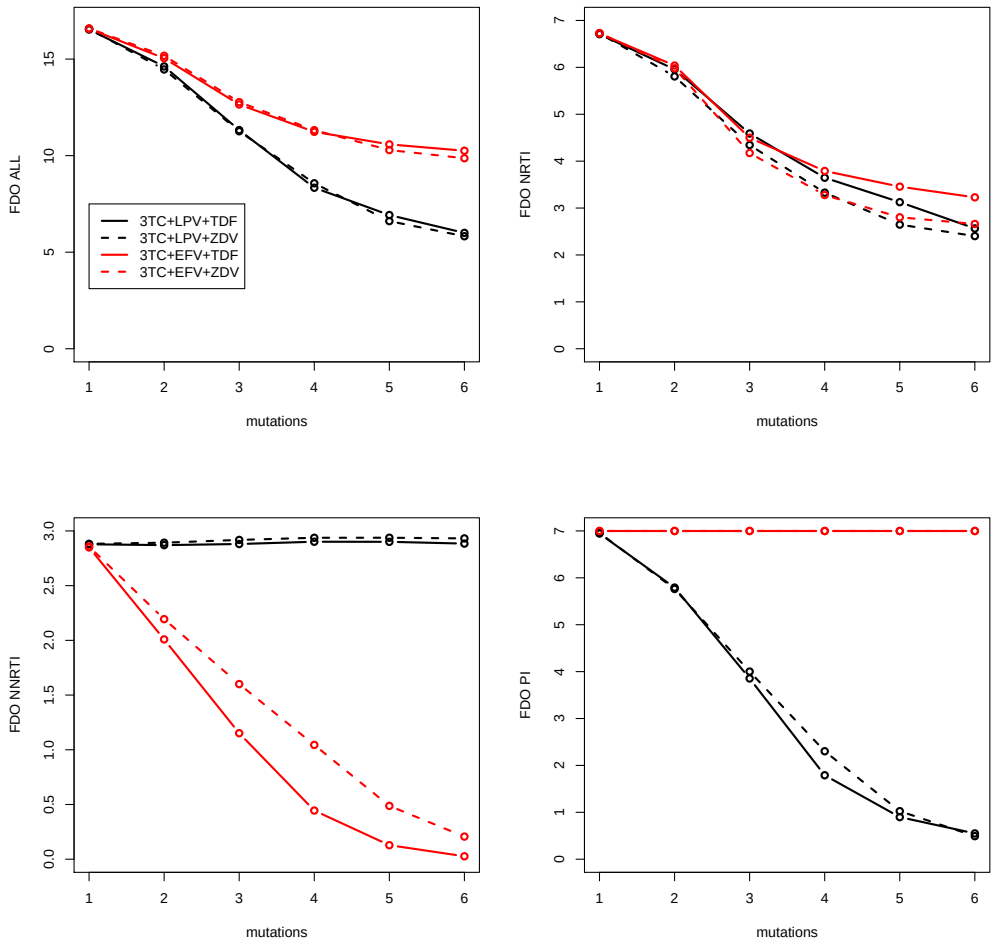


Figure 6.16: Change of future drug options (FDO) as response to four typical first-line treatments. The panels depict the FDO for all drugs (top left), NRTIs (top right), NNRTIs (bottom left), and PIs (bottom right).

ers that implement mutation scores that are dependent on present mutations; however, these transducers are complicated to construct, and the results are in general very large automata (i.e. many additional states and edges are required to represent alternative paths that go along preexisting mutations). Of course, large transducers increase the required computation time (and the I/O overhead).

Another method that takes preexisting mutations into account is the approach introduced by Deforche et al. (2008) for estimating *in vivo* viral fitness landscapes. Here, a Bayesian network model for an individual drug is trained on viral sequences that were obtained during treatment with that drug. This network is used to capture interactions between mutations. Furthermore, in a second training phase the conditional probabilities of the Bayesian network are replaced with conditional fitness contributions of each mutation. This fitness function is applied together with a simple finite ideal Wright-Fisher population model (Wright, 1931) for simulating the resistance development in the viral population during anti-HIV treatment. However, the approach has to model and simu-

late the resistance development for each viral variant in the population (Deforche et al. (2008) used a population size of 10^4), and is therefore likely not to meet the computational demands for an interactive system.

The major problem for all approaches is the estimation of mutation models for drug combinations instead of individual drugs. For instance, longitudinal *in vivo* data is usually scarce (see Figure 6.8); the available data decreases even further when one is restricted to a single drug combination. Thus, constructing models for individual drugs is likely to achieve more robust models. Estimation of such models from *in vivo* data, however, is problematic since drugs are given in combination and the causal inference between drugs and emerging mutations is hard to solve. In the end, the most stable approach is the estimation of mutation models for individual drugs from *in vitro* data followed by a combination approach to realize models for drug combinations.

Finally, another possibility for finding sequences of beneficial drug combinations is to omit the explicit resistance development in the viral population and directly learn from the sequences of combination therapies in the patients' pasts. For instance, one can focus on the success of a treatment given the $N-1$ preceding combinations. We recently investigated such an N-gram graph of therapy sequences as an potential encoding for a patient's therapy history (Müller, 2008). Unfortunately, only few treatment sequences occurred sufficiently often. However, it cannot be excluded that with the significantly larger EuResist integrated database, which is available today, and a more elaborate statistical learning method (or encoding of therapy sequences) information about beneficial orders of drug combinations can be extracted without explicitly simulating the viral population.

7 A Solution to the HIV Pandemic: A Vaccine

Current treatment options do not provide a cure to HIV infections. Hence, the ultimate solution to the HIV pandemic is to prevent further infections. One established instrument for preventing infections are vaccines, and therefore the development of a successful HIV vaccine is of major interest today. In this chapter we will present a bioinformatics approach to aide the search for immunogens that elicit broadly neutralizing antibodies against HIV. In Section 7.1 we will very briefly summarize the major elements of the immune system and the molecular basis of immunity conferred by vaccination. This is followed by a short overview on HIV vaccine research, including a focus on the viral spike (gp160), which is the target for neutralizing antibodies (Section 7.2). Section 7.3 focuses on predicting antibody-mediated neutralization from gp160 genotype based on statistical models. These models are also used to identify features of gp160 that increase susceptibility to neutralizing antibodies. This knowledge will be used to find gp160 variants that can be used as immunogens in a successful HIV vaccine. Section 7.4 concludes the chapter with an outlook.

7.1 The Immune System and Vaccines

In the following we provide a brief overview on the immune system based on DeFranco et al. (2007). The immune system of an organism is organized as a layered defense. The first layer comprises physical barriers (e.g. shells, skin) and mechanical mechanisms (e.g. coughing, sneezing). Once a pathogen manages to surmount the first line of defense, it faces mechanisms of the innate immune system and the adaptive immune system, where the latter one is a characteristic of jawed vertebrates.

Innate Immune System

The cells and molecules of the innate immune system recognize features that are conserved across broad groups of pathogens. This system reacts to pathogens in a generic (non-specific) way and it confers neither long-lasting immunity nor increased efficacy after previous exposure to a pathogen. The innate immune system includes different defense mechanisms: humoral and chemical barriers of the innate system include inflammation and the complement system. Inflammation is a process for recruiting immune cells to the sites of infection; it is produced by eicosanoids and cytokines that are released by injured or infected cells. The complement system consists of about 30 serum and membrane proteins that can trigger a variety of immune reactions by a biochemical signal cascade. Furthermore, the innate response depends on a range of leukocytes (white blood cells). A major subgroup of leukocytes comprises phagocytes, which act by engulfing particles or pathogens: the pathogen is trapped in a vesicle and subsequently "destroyed" by digestive enzymes. Dendritic cells and macrophages (a type of phagocyte) present parts of the digested pathogen on their surface. These antigens (*anti*body *gen*erators) bridge the gap

to the adaptive immune system by triggering the response of T cells and B cells, which constitute the main building block of the adaptive immune system.

Adaptive Immune System

T cells and B cells belong to the group of lymphoctytes, a special type of leukocytes. T cells are involved in a cell-mediated immune response, whereas B cells are part of the humoral response. Both T and B cells carry receptor molecules that recognize only specific targets, and each cell produces only one type of antigen-recognizing receptor. T cells are divided into two subgroups: killer or cytotoxic T cells and helper T cells. T cells only respond to antigens that are bound to a protein of the major histocompatibility complex (MHC), killer T cells ($CD8^+$ T cells) require antigens coupled to Class I MHC molecules, whereas helper T cells ($CD4^+$ T cells) rely on Class II MHC molecule-antigen complexes. As their name suggests cytotoxic T cells release substances, which are toxic for cells (cytotoxins), and thereby directly kill infected (e.g. by viruses) and damaged or dysfunctinal (e.g. cancer) cells. Helper T cells activate other cells of the immune system and thereby regulate both the innate and adaptive immune responses. For instance, helper T cells stimulate killer T cells and enhance the microbicidal function of macrophages. Moreover, helper T cells also activate B cells to produce neutralizing antibodies. Both helper and killer T cells originate from naïve T cells, and the required proliferation into effector T cells is mediated by the dendritic cells from the innate immune system. Naïve B cells carry B cell antigen receptors (BCR) on their surface that contain immunoglobulins (antibodies). Once the BCR binds to its specific antigen (e.g. a virus surface protein) the complex undergoes endocytosis. The antigen is cleaved into small peptides that are presented by Class II MHC molecules on the surface of the cell. After recognition and subsequent activation by a helper T cell the B cell becomes activated and undergoes differentiation into an effector B cell that secretes copies of exactly the antibody, which recognized the antigen. Antibodies can directly inactivate (neutralize) pathogens by binding to its surface proteins that are used to infect target cells or to toxins secreted by bacteria. In addition, bound antibodies mark their target for destruction by the innate immune system, i.e. complement activation or processing by phagocytes. In principle, only antigens that can be recognized by the preexisting pool of B cells can trigger an immune response. However, on successful activation, the antigens produced by B cells undergo a process termed affinity maturation in which somatic hypermutations lead to modified binding affinities of the antibody to the antigen, and antibodies with higher affinity are selected. In contrast to the innate immune system, the adaptive immune system maintains specific cells targeting specific pathogens, or more precisely: specific antigens. Moreover, an immunologic memory is established by offsprings of T and B cells that begin to replicate after activation. These memory cells can survive for decades and therefore the immune system is prepared for a challenge with the same pathogen or more precisely by pathogens presenting the same antigens.

The aim of a vaccine is to induce immunity against a pathogen, that is, invoking a strong immune response against the pathogen upon recognition. While the main goal of a vaccine is to achieve sterilizing immunity (complete protection from infection), it is often sufficient to substantially reduce the amount of active pathogens in the host and therefore limit disease-related symptoms. Protection against a pathogen is elicited by challenging

the adaptive immune system with adequate antigens that are able to trigger a response by the immune system and consequently establish the immunologic memory in the form of T and B cells.

An immune response can be triggered by mild versions of the pathogen that usually do not cause the disease (live attenuated vaccines), by killed pathogens (inactivated vaccines), or by parts of the pathogens (compound vaccines) (Plotkin, 2005). The general idea is to mimic an attack by the pathogen and thereby create the immunity that would be naturally established after the real challenge. In general, the more similar the immunogen comprising the vaccine is to the real pathogen, the better and longer lasting is the protection by the vaccines. Hence, live attenuated vaccines are more effective than inactivated vaccine, which, in turn, are more effective than compound vaccines (Lambert et al., 2005). For instance, multiple shots of the new hepatitis B vaccine (a compound vaccine) are required to establish immunity. Inactivated and compound vaccines, however, are safer than live attenuated vaccines, since they are incapable of causing the disease.

Most of today's successful vaccines rely on the establishment of memory B cells and therefor on the rapid generation of matching neutralizing antibodies (nAbs) upon a challenge by the pathogen. Vaccines are a major instrument for controlling diseases, and since the introduction of the first vaccine against smallpox in 1798, smallpox have been eradicated and other pathogens have almost disappeared. Today, vaccination is extended from infectious to noninfectious diseases (immunotherapy). Among the infectious diseases HIV, malaria, and tuberculosis are still of leading interest, while research in the area of noninfectious diseases currently focuses on various types of cancer, autoimmune diseases (e.g. multiple sclerosis), atherosclerosis, and Alzheimer's disease (Plotkin, 2005).

7.2 Vaccines and HIV

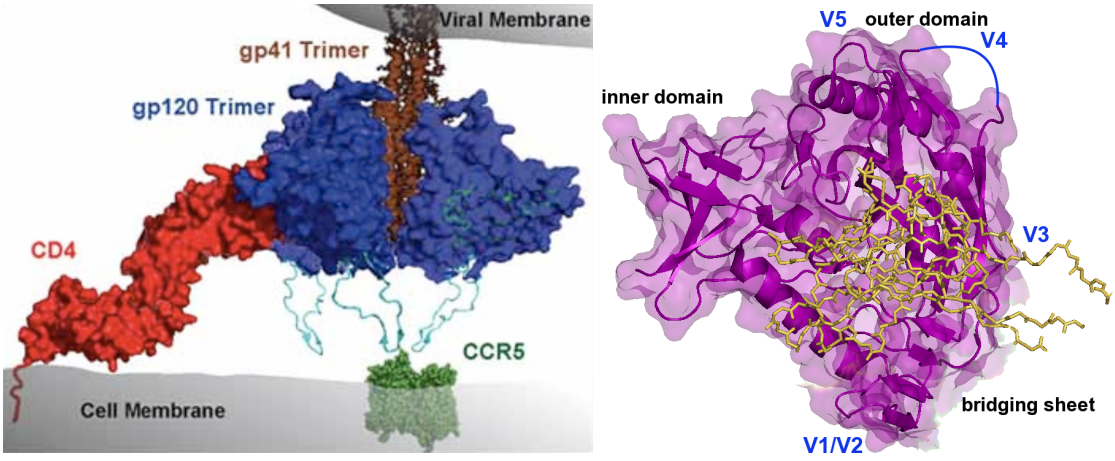
After identification of the pathogen that causes AIDS, researchers expected to have a vaccine within a few years. Indeed, on April 23 in 1994 Magaret Heckler (then U.S. Health Secretary) proclaimed publicly that a vaccine candidate will be available for testing within two years (Walker and Burton, 2008). Now, 28 years after the discovery of HIV there is still no vaccine. Even worse, there is not even a promising candidate in close sight. The challenges that researchers are facing during the development of an HIV vaccine are manifold: First, HIV is extremely diverse, which makes it hard to identify conserved epitopes, and due to the error-prone replication process HIV can quickly evade the immune response by mutating epitopes that are recognized by the adaptive immune system. Second, a characteristic of HIV is the unique choice of target cells: the virus infects cells of the immune system, in particular $CD4^+$ helper T cells that coordinate the immune response. Consequently, the virus slowly and steadily destroys is natural antagonist. Third, HIV has evolved mechanisms to counter immune responses, e.g. the accessory protein *Nef* down-regulates the synthesis of MHC molecules, which are crucial for the identification of infected cells by T cells (Yang et al., 2002), and the surface envelope proteins exhibit multiple features for successfully avoiding antibody recognition. A further complication for vaccine design is resulting from the fact that HIV integrates its genetic information into the host genome rapidly after infecting a cell. When the provirus is in its latent state, the cell shows no signs of infection and is immunologically silent and thus unrecognized

by the immune system. This way the virus can survive for decades in memory cells of the immune system, its latent viral reservoir.

All these unique features of HIV lead to following requirements for a successful vaccine: recognition of an array of diverse viruses and fast response to the challenge in order to prevent the establishment of viral reservoirs. If the vaccine fails to prevent the latter, it has at least to assist the natural immune response in preventing the destruction of $CD4^+$ cells during the acute phase (Walker and Burton, 2008). This would enable the patient to better control the infection, i.e. maintain a low viremia, and thereby substantially delay progression to AIDS and, in addition, reduce the risk of transmission to further individuals.

Because cells present fragments of proteins they produce on their surface via their MHC molecules, HIV vaccines aiming at eliciting T cell-mediated immunity can (in theory) be based on any of the viral proteins. Vaccines that rely on eliciting neutralizing antibodies, on the other hand, can only target the viral spike (gp160). As mentioned earlier (Section 2.2) the viral spike comprises two sub-units, gp41 (transmembrane) and gp120, that occur in trimeric form (Figure 7.1 a) (Liu et al., 2008). During cell entry, the glycoprotein gp120 binds to the CD4 receptor and the coreceptors, while gp41 is responsible for membrane fusion. The viral spike is heavily shielded against the host immune system by multiple mechanisms. For instance, the glycoprotein, as its name suggests, is densely covered with sugar molecules that shield conserved epitopes against the adaptive immune response. The subunit gp120 contains five variable regions (V1-V5) that are inter-spaced with conserved regions. These five variable loops exhibit a very high sequence diversity and provide further protection of the conserved parts of gp120 against immune recognition. The variable surface region of gp120 is also referred to as the "silent face" because only few antibodies are elicited against it (Karlsson Hedestam et al., 2008). Figure 7.1 b) shows the core of the gp120 protein in contact with the CD4 receptor. Also its own structure protects the viral spike against the immune response. Conserved regions in gp41, for instance, are inaccessible, simply because there is too little space between gp120 and the viral membrane to allow for binding of the B cell receptor. A more advanced protection mechanism is termed conformational masking. Here, conserved epitopes are only accessible after conformational changes in the protein. For instance, the coreceptor binding site of gp120 is only accessible after the conformational change induced by CD4 binding (Douek et al., 2006).

Upon infection, a race between HIV and the adaptive immune system commences. Unfortunately, the virus is always one step ahead. More precisely, the adaptive immune system inflicts selective pressure on the evolution of HIV, which drives HIV to evolve into variants that cannot be neutralized by the available array of neutralizing antibodies. Richman et al. (2003) studied blood sera and viruses from patients at different time points. Briefly, their analysis showed that the blood serum at time point t is very well capable of neutralizing older viral variants, i.e. obtained at time point $t-1$ and earlier. However, the sera are only insufficiently capable or incapable of neutralizing the current or future variants, respectively. Thus, antibodies typically generated during a chronic HIV infection cannot be used as basis for a vaccine. Fortunately, some chronically infected individuals elicit antibodies that are capable of neutralizing a broad range of HIV variants. These broadly neutralizing antibodies are of major interest for the design of a successful HIV vaccine because they provide the means of establishing sterilizing immunity. In addition, currently existing neutralizing antibodies demonstrate that the viral spike has a few weak



(a) The viral spike

(b) Structure of the gp120 core

Figure 7.1: A schematic representation of the HIV-1 envelope glycoprotein trimer in contact with the CD4 receptor (red) and the coreceptor (green) of the host cell (a). The trimer comprises three copies of the transmembrane glycoprotein gp41 (brown), the three copies of the surface glycoprotein gp120 (blue) are attached to gp41. Figure reprinted from Phogat and Wyatt (2007) with kind permission from Bentham Science Publishers. Structure of core gp120 (purple) in complex with the CD4 receptor (orange) (b). The figure was rendered with PyMOL on the basis of PDB entry 2nxy. Estimated position of the five loops is given in blue font.

spots that can be exploited by rational vaccine design. This knowledge can help to develop an immunogen that elicits broadly neutralizing antibodies (bnAbs) *in vivo*.

Until now, a number of broadly neutralizing antibodies have been identified in chronically infected individuals (Burton et al., 2004) and the search for new bnAbs is still going on (Walker et al., 2009). One of the first bnAbs discovered, which is called b12, targets an epitope that substentially overlaps with the conserved CD4 binding site on gp120 (Burton et al., 1994). Consequently, upon binding to gp120 the nAb prevents CD4 attachment, and thus the conformational change that is a prerequisite for successful entry of HIV into the target cell. The nAbs 2F5 and 4E10 recognize conserved linear epitopes on gp41. It is expected that these nAbs act like fusion inhibitors, i.e. they exhibit their neutralizing effect by preventing the conformational change in gp41 leading to fusion of viral and cell membranes. Figure 7.2 depicts the viral spike and epitopes recognized by some bnAbs. Recently, Trkola et al. (2008) demonstrated the efficacy of bnAbs *in vivo* with the aim to determine clinically relevant titers of the antibodies.

HIV Vaccine Candidates

So far, there have been a few vaccine candidates that were tested in clinical trials. The most prominent examples are AIDSVAX and the STEP study.

Within STEP (launched in 2004) a candidate vaccine using a replication-deficient adenovirus type 5 vector that expresses some HIV clade (subtype) B proteins (*Gag, Pol, Nef*) was investigated. As it is clear from the list of HIV proteins involved, the candidate was

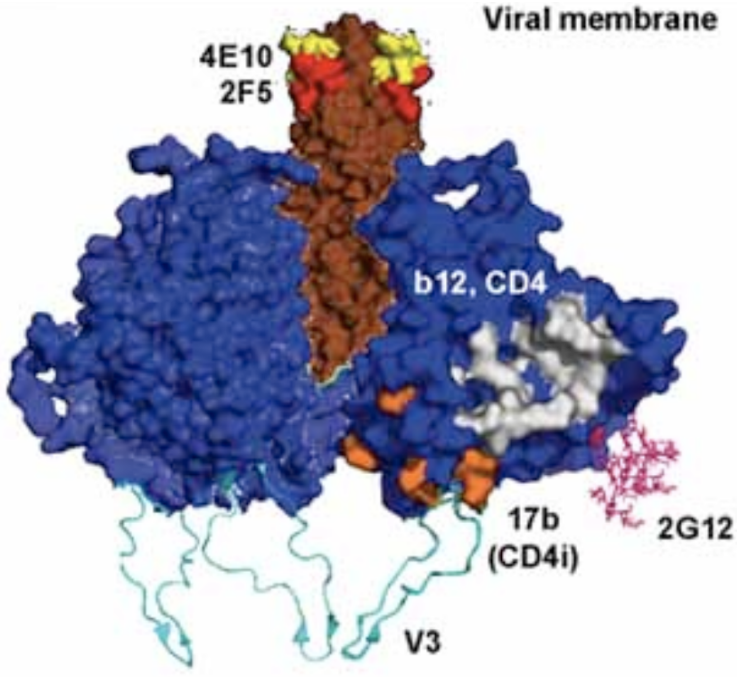


Figure 7.2: Structure of the envelope trimer with highlighted epitopes of some broadly neutralizing antibodies: 2F5 (red), 4E10 (yellow), b12 (grey), and 17b (orange). Figure reprinted from Phogat and Wyatt (2007) with kind permission from Bentham Science Publishers.

targeted to elicit T cell mediated immunity. The clinical trial, however, was stopped in 2007 by the Data and Safety Monitoring Board. The reason for the premature end of the trial was that individuals that received the vaccine exhibited an enhanced rate of HIV-1 acquisition compared to people in the control group (Barouch, 2008). This, of course, was an immense setback for the HIV vaccine research.

The AIDSVAX vaccine candidate developed by the company VaxGen uses monomeric gp120 to elicit humoral immune response, i.e. neutralizing antibodies. The candidate managed to elicit type-specific antibody response. A protective effect, however, could not be detected in two clinical trials in 2005 and 2006 (Barouch, 2008).

Recently, AIDSVAX received further media coverage when in September 2009, researchers reported a success in the search for an HIV vaccine. In a study started in October 2003 in Thailand, about 16,000 participants of HIV negative men and women received a combination of the two earlier tested vaccine candidates AIDSVAX and ALVAC-HIV. ALVAC-HIV uses a canarypox vector that, like the candidate in the STEP trial, expresses HIV (poly-)proteins (here: *Gag, Env,* PR). At the endpoint of the study, 76 participants from the control group were infected with HIV, compared to 56 of the vaccinated participants. Hence, the vaccine's efficacy was 26.4%. This result, however, did not reach statistical significance. Statistical significance and 31.2% vaccine efficacy was achieved after excluding seven individuals that were found to have HIV at baseline (Rerks-Ngarm et al., 2009). Although a glimmer of hope, the protection conferred by the vaccine is only marginal and the analysis of the trial is regarded a "data crunch" for achieving statistical significance[1].

[1] http://www.nature.com/news/2009/091021/full/news.2009.1035.html

Despite the immense promise given by bnAbs, it has not been possible yet to identify an immunogen that elicits bnAbs *in vivo*. With the work presented in the following we investigate what determines the susceptibility of an *Env* variant against three of the available bnAbs.

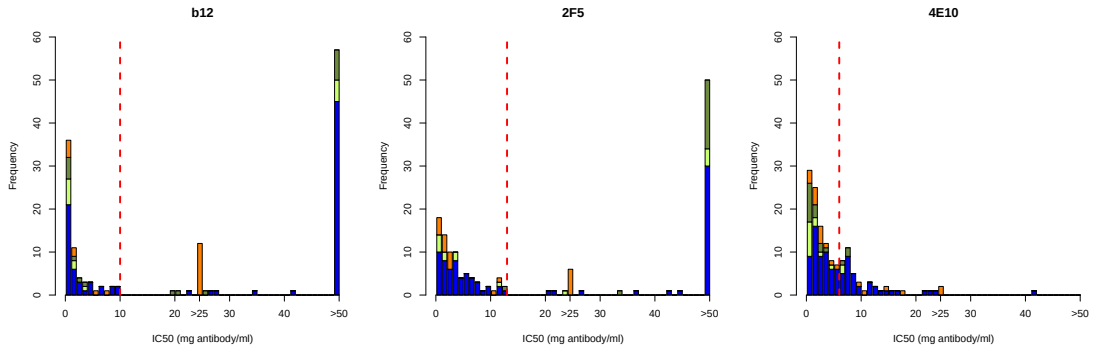
7.3 Predicting Neutralization from Genotype

This section describes a joint work with Fabian Müller, Oliver Sander, Thomas Lengauer, Klaus Überla, and Francisco Domingues. Here, using an array of statistical learning methods, we investigate the genotype-phenotype-relationship between the *Env* genotype and the neutralization phenotype of 147 viral variants with three different bnAbs. The experimental assay for determining the neutralization phenotype is very similar to the approach used for assessing *in vitro* drug resistance phenotypes (Section 2.5) and thus not described in further detail (Richman et al. (2003) for example used an adapted version of the original drug resistance assay developed by Petropoulos et al. (2000)). We selected statistical learning methods that allow for assessing feature importance and analyze statistically relevant positions in terms of susceptibility to neutralization by bnAbs. Finally, based on the knowledge derived from interpreting the statistical learning methods we propose *in vitro* experiments for experimental validation of our model. The focus of our investigation is the antibody b12, which targets an epitope that largely overlaps with the CD4 binding site. The model will be used to optimize an immunogen in an HIV vaccine development project.

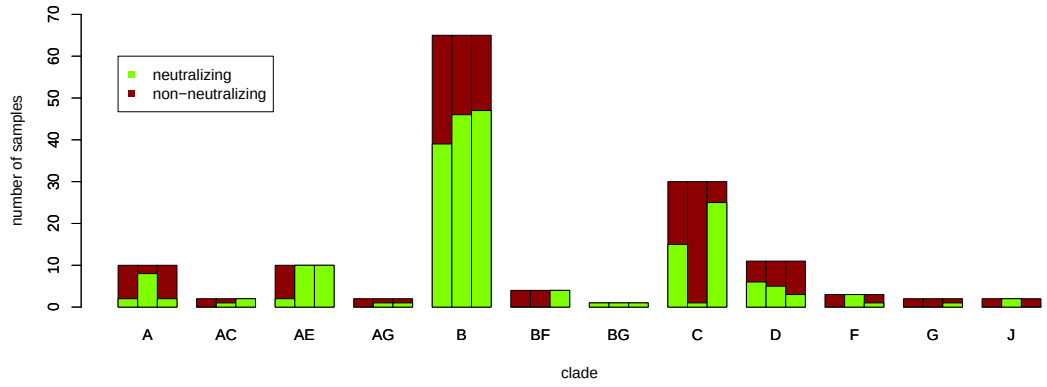
7.3.1 Material and Methods

Data

The data for this study originate from four publications. Binley et al. (2004) studied the neutralization phenotype regarding a panel of nAbs of 90 viral variants from different clades, Li et al. (2005, 2006) analyzed 19 and 18 variants from clade B and C, respectively, and Schweighardt et al. (2007) examined another 20 variants from subtype B. The datasets are named according to their origin and subtype of focus (*Binley, Li.B, Li.C, and Schweighardt.B*). Due to an overlap of viral variants studied in Binley and Li.B, five of the total 147 samples were excluded from the analysis. All publications provided IC_{50} values for the three nAbs b12, 2F5, and 4E10. The continuous neutralization phenotypes were dichotomized to "neutralizing" and "non-neutralizing" using an antibody-dependent cutoff; IC_{50} values below the cutoff indicate neutralization by the antibody. For b12 and 2F5 the cutoff selection was straightforward, since a large fraction of the samples was clearly not susceptible to the antibody (values ">50" and ">25" of mg antibody/ml, respectively), and variants with intermediate susceptibility were rare. The neutralization effect of 4E10, however, was so broad that only very few viruses were completely resistant. Thus, a cutoff between very susceptible and intermediate susceptible isolates was selected based on the distribution of IC_{50} values. Figure 7.3 a) depicts the distribution of IC_{50} values for all three bnAbs and the applied cutoffs, and Figure 7.3 b) shows the distribution of the binary neutralization phenotype within each clade in the dataset. The b12 phenotype data for clades B, C, and D is almost balanced, while other clades (represented by a few isolates only) are mainly non-neutralizing.



(a) IC_{50} distribution



(b) Clade distribution

Figure 7.3: Histograms of IC_{50} values for antibodies b12, 2F5, and 4E10 (a). Origin of the data is color-coded: Binley (blue), Li.B (light green), Li.C (dark green), Schweighardt.B (orange). Cutoffs are depicted by red dashed lines. Distribution of the binary neutralization phenotype within each clade (b). Left, middle, and right bars correspond to the b12, 2F5, and 4E10 phenotype, respectively.

Amino-acid sequence data for the 142 *Env* variants were obtained from GenBank using the accession numbers listed in the four publications. The *Env* gene comprises the signal peptide, gp120, and gp41. The sequences were aligned using MUSCLE (Edgar, 2004) according to a guide tree based on the HIV phylogenetic tree published by Leitner et al. (2005). The sequence of HXB2 was added to the multiple sequence alignment as an alignment independent reference. For details on the sequence processing please refer to the Master thesis of Fabian Müller (Müller, 2009). The final alignment comprised 975 amino acid positions.

Feature Encodings

The amino acids at each of the 975 alignment positions were encoded in two different ways. The *categorical encoding* was used for statistical methods that can operate on categorical data. Here, each alignment position was represented as a single character: one of the 20

amino acids, a gap symbol (-), symbols for uncertain amino acids (B and Z, representing presence of Asparagine or Aspartic acid and Glutamine or Glutamic acid, respectively), or a symbol representing any amino acid (X). Hence, each viral variant was represented by one feature for each alignment position. The *binary encoding* was input to support vector machines (SVMs), which are unable to operate on categorical features with more than two categories. Consequently, we used the same encoding as in GENO2PHENO (Section 2.5.2), i.e. we encoded an amino acid sequence using binary indicator variables. More precisely, one indicator variable was used for each possible symbol at each alignment position. Thus, a sequence was represented as a binary vector of length 24×975. We refer to the categorical encoding and the binary encoding as the *sequence* encoding.

A property that is only implicitly captured by the sequence encoding is the location of potential N-linked glycosylation sites. As stated before, sugar molecules attached to the surface of gp120 mask conserved epitopes from recognition by the immune system. The glycosylation processes is carried out by the host cell at Asparagines (N) that are followed by any amino acid (X) and Serine (S) or Threonine (T), i.e. the typical NX[S,T] pattern of N-linked glycosylation. There were 268 positions in the alignment showing at least one N among all sequences. Of those, 148 fulfilled the NX[S,T] criterion in at least one *Env* variant. The respective 148 binary features constitute the *glycosylation encoding*, which were added to the sequence encoding (*glyco*).

Apart from regions of high sequence variability, the multiple sequence alignment was very stable. Unfortunately, regions of high sequence variability comprise exactly the five variable loops that are believed to play an important role in determining antibody neutralization (Wyatt and Sodroski, 1998; Pantophlet and Burton, 2006). In order to further investigate the influence of variable loops on the ability to predict neutralization by antibodies, they were completely removed from the sequence encoding (*no loops*). In addition, we used sequence encodings that maintained sequence information on exactly one variable loop, for all five loops (*only Vx*). Furthermore, as a replacement for the sequence information, we encoded the five loops using alignment-independent features meant to capture physicochemical properties, which can be derived from the sequence information. More precisely, for *variable loop features* we used the molecular weight (sum of all molecular weights of residues in the loop; one continuous feature per loop), the charge (number of positively and negatively charged amino acids or the overall charge; three integer features per loop), the length (number of residues in the loop, one integer variable per loop), the hydropathy (sum of all hydropathy indices of the residues (Kyte and Doolittle, 1982); one continuous feature per loop), the number of potential hydrogen bonds (absolute number of residues potentially participating in hydrogen bonds, and sum of all donors and acceptors; two integer values per loop), the N-linked glycosylation (number of potential glycosylation sites; one integer per loop), the fold index (based on Prilusky et al. (2005); one continuous variable per loop), and simply the abundance of each amino acid in the loop (20 integer values per loop). These variable loop features were added to the *no loops* encoding (*loop feat*).

Finally, we also investigated subsets of the sequence encoding that were restricted to the three parts of *Env*: the signal peptide (*signal*), gp120, and gp41.

All applied encodings are briefly summarized in Table 7.1.

encoding	description
sequence	comprises sequence information of the *Env* gene: for the RF, amino acids sequences were encoded using categorical features; for the SVM, sequences were represented by a binary encoding
glyco	sequence encoding and additional 148 binary features indicating potential glycosylation sites
no loops	sequence information removed from all variable loops
loop feat	sequence information removed from all variable loops; instead loops are represented by physicochemical properties (see text)
only Vx	sequence information removed from all variable loops, but Vx, with $x \in \{1, \ldots, 5\}$
signal	sequence encoding restricted to signal peptide
gp120	sequence encoding restricted to gp120
gp41	sequence encoding restricted to gp41

Table 7.1: List of feature encodings. The features are briefly summarized, for a full description see the text.

Statistical Learning Methods

We employed two statistical learning methods for predicting the binary neutralization phenotype of viral *Env* variants. The Random Forests (RF) method (Breiman, 2001) trains B decision trees on B bootstrap replicates of the training data. At each node, only m randomly chosen variables of all p variables are considered for the split. The variable that reduces the impurity of the labels the most is selected for splitting the node. In RF the impurity is measured using the Gini index: $\sum_k p_k(1-p_k)$, where k and p_k are the class and the frequency of that class at the node to be split, respectively. All B trees are fully grown and left unpruned. For the two parameters of the method, the number of features checked at each split (m) and the number of trees (B) we chose standard values $\lfloor\sqrt{p}\rfloor$ and 500, respectively. For the RF classifier the categorical sequence encoding was used since RF can operate on categorical variables.

As second classifier we applied linear support vector machines (SVMs). For the parameter C, which penalizes misclassified samples in the training set, values of $2^{-7}, 2^{-6}, \ldots, 2^2$ were tested in a 10-fold cross-validation. The choice of the parameter did not influence the performance substantially, and was therefore set to 1. Since linear SVMs do not support categorical data the binary sequence encoding was applied. For achieving optimal performance, constant features were removed, and the built-in scaling option for features in the libSVM implementation (Chang and Lin, 2001) was used.

The classification performance for both methods was assessed using the area under the ROC curve (AUC) in a 10 times five-fold cross-validation setting. Both classifiers were used together with all feature encodings for predicting neutralization by all three antibodies. Statistical significance is assessed using a one-sided Wilcoxon test on the 10 mean AUCs obtained in the five-fold cross-validation runs.

Measures of Variable Importance

We used Mutual Information (MI) and p-values from Fisher's exact test (Fisher, 1922) as univariate measures of feature importance. Both measures were applied only to the categorical sequence encoding. Furthermore, we used the mean decrease in Gini index (MDG) extracted from the RF models and z-scores of features weights derived from the linear SVM (Section 3.1) as multivariate measures of feature importance. Briefly, the MDG of a feature is the decrease in Gini index observed when a node was split using that feature – averaged over all B trees in the forest. Due to the binary encoding of the sequence for the SVM, the feature importance (z-scores) was provided for each amino acid at each alignment position. In order to make the measure comparable to the measures provided by the other three methods, which provide one value for each alignment position, only the maximum z-score at each alignment position was used as a measure of feature importance.

As recently described, the MDG is biased towards features with a high number of categories (Strobl et al., 2007), e.g. an uninformative variable with 10 categories receives a higher MDG than an uninformative variable with two categories. Consequently, informative features with few categories may go unnoticed among uninformative features with many categories. This bias is of particular interest for our problem, since alignment positions in unstable regions of the alignment show a higher number of different amino acids, i.e. categories, than positions in stable regions. Thus, MDG for amino acids located in the variable loops are likely to be boosted by this artifact.

In order to correct for this bias, we developed a heuristic based on permutations. Briefly, the response vector is randomly permuted k times, and for each permutation the RF classifier is trained and MDG for each variable is computed. Mean μ and standard deviation σ of MDG for one feature are estimated from the k models. Under the assumption that the MDG of an uninformative variable is normally distributed, given μ and σ, the probability of observing the MDG value obtained in combination with the original response vector serves as a p-value. We termed this approach *permutation importance* (PImp; Altmann et al. (2010)). For correcting the MDG importance, we used 1000 permutations of the original response vector.

For MI and the combined z-scores we observed the same bias as for MDG (Müller, 2009), and consequently, we applied PImp for correcting it. For both MI and SVM-based z-scores the PImp p-value was based on 1000 iterations.

All measures of variable importance were applied to the complete set of 142 training samples, i.e. not derived from cross-validation models.

In vitro Validation

In order to experimentally validate our b12 prediction model, site-directed mutagenesis is performed to confirm important positions derived from the feature importance analysis. To this end, the b12 neutralization phenotype is measured before (original viral variant) and after the introduction of up to three mutations. In total, 18 sets of mutations were introduced into 10 different viral variants.

7.3.2 Results

Classification Performance

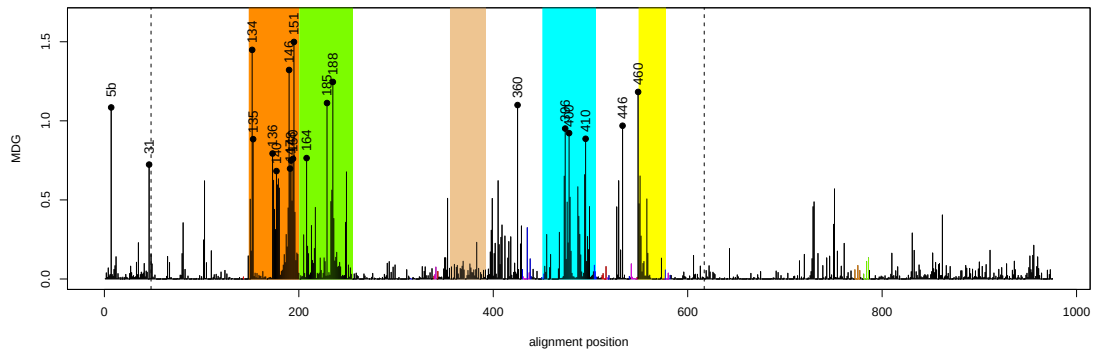
Table 7.2 lists the mean AUC and its standard deviation assessed in 10 repetitions of 5-fold cross-validation. Based on sequence information alone both classifiers achieve an AUC of around 0.8 for predicting the b12 neutralization phenotype. Adding information on putative glycosylation sites slightly decreased the performance for both methods. Removal of the sequence information from the five variable loops led to a substantial and significant decrease in performance for the RF ($p = 0.00098$) and the SVM model ($p = 0.00195$). Replacement of the sequence information of the variable loops with the variable loop features led to a further decrease in performance. Maintaining the sequence information of the V2 loop resulted in re-establishment (and even slight improvement) of the performance obtained with the complete sequence. Maintaining one of the other variable regions did not improve performance substantially over the model without any variable loop information (data not shown). The model based on the gp120 region alone achieved a performance comparable to the sequence encoding, while models based on the signal peptide or gp41 alone were clearly inferior (both $p = 0.00098$). For most encodings (except "signal" and "gp120") the SVM outperformed the RF model significantly ($p < 0.03223$).

RF performs significantly better than the linear SVM ($p < 0.00977$) for the prediction of neutralization by nAb 2F5. With AUCs above 0.9 the performance for the RF models is better than that for the b12 model. Addition of glycosylation features resulted in a slight decrease in performance, while removal of variable loop information, unlike in the case of b12, resulted in a slight increase. The best performance was achieved for both learners when the sequence information was restricted to gp41. For both methods the model restricted to gp41 sequence information model significantly better than the sequence encoding ($p < 0.004883$).

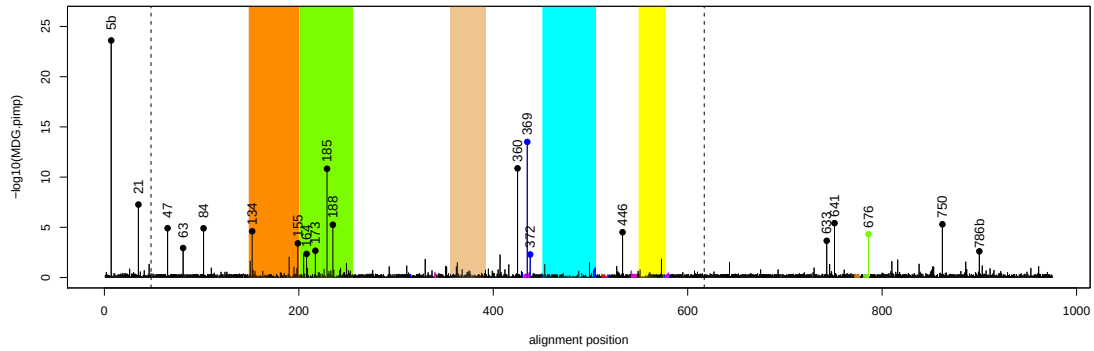
With values around 0.72 the performance of 4E10 neutralization prediction is slightly worse than the performance observed for the b12 models. Here, addition of glycosylation and removal of the variable regions slightly increased the model performance. As in the case of 2F5, the best models were obtained when the information was restricted to the target of the nAb: gp41. The improvement reached statistical significance only for the RF classifier ($p = 0.001953$).

Variable Importance

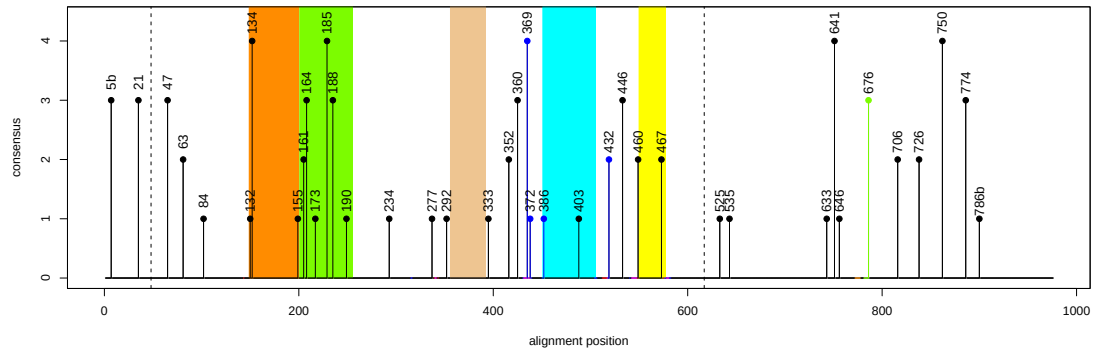
Figure 7.4 a) depicts the raw MDG importance extracted from the RF model predicting b12 neutralization. A large number of alignment positions within the variable loops (primarily V1,V2, and V4) received high MDG values. The PImp correction of the MDG (Figure 7.4 b), however, yields a different picture. Here, most of the variable loops do not contain any information. The exception is V2, which contains four of the top 20 most important positions. In general, according to this model, only few positions are linked to the b12 neutralization phenotype. Surprisingly, among the top 20, two important positions are located in the signal peptide, and another five are in gp41. One of the positions in the signal peptide (5b; an insertion compared to HXB2) is indeed the top-most important position. The second most important position is located in the b12 binding site.



(a) MDG



(b) PImp MDG



(c) consensus

Figure 7.4: Feature importance derived from models that predict b12 neutralization: RF MDG (a), PImp of MDG (b), consensus of all four methods (c). The regions of *Env* (signal peptide, gp120, and gp41) are separated by vertical dashed lines. The variable regions within gp120 are highlighted in the background orange (V1), green (V2), beige (V3), light blue (V4), and yellow (V5). Furthermore, known epitopes are highlighted with red (CD4 binding), blue (b12 binding), purple (CD4 and b12), orange (2F5 binding), and green (4E10 binding). For the top 20 important residues, the position in reference to HXB2 is provided.

	b12		2F5		4E10	
	RF	SVM	RF	SVM	RF	SVM
sequence	0.791 (0.018)	0.808 (0.020)	0.914 (0.013)	0.840 (0.012)	0.717 (0.026)	0.718 (0.018)
glyco	0.782 (0.021)	0.803 (0.020)	0.905 (0.019)	0.836 (0.014)	0.724 (0.032)	0.723 (0.020)
no loops	0.765 (0.013)	0.781 (0.019)	0.927 (0.016)	0.845 (0.015)	0.736 (0.025)	0.719 (0.022)
loop feat	0.755 (0.015)	0.773 (0.021)	0.911 (0.017)	0.849 (0.013)	0.719 (0.023)	0.710 (0.021)
only V2	0.801 (0.017)	**0.813 (0.019)**	0.924 (0.011)	0.829 (0.011)	0.727 (0.028)	0.708 (0.029)
signal	0.708 (0.024)	0.579 (0.031)	0.766 (0.016)	0.693 (0.021)	0.621 (0.038)	0.587 (0.037)
gp120	0.788 (0.015)	0.792 (0.026)	0.825 (0.028)	0.807 (0.016)	0.694 (0.023)	0.680 (0.026)
gp41	0.734 (0.019)	0.752 (0.020)	**0.956 (0.013)**	0.866 (0.016)	**0.754 (0.019)**	0.726 (0.028)

Table 7.2: Classification performance. The mean (sd) classification performance measured using AUC and assessed via 10 repetitions of 5-fold cross-validation for the two classifiers (RF and SVM) and the three different antibodies. Rows correspond to different feature sets (see Table 7.1 for a brief description). The best performance for each antibody is indicated by bold font.

Every measure for feature importance provided a slightly different picture (data not shown). Thus, for the final interpretation we will rely on a consensus of the four methods. The consensus (Figure 7.4 c) counts how often each position occurs among the top 20 most important positions (thus, only values of $0, \ldots, 4$ are possible). Figure 7.5 depicts the location of important positions (i.e. appearing in two or more top 20 lists) on the gp120 structure (PDB ID 2nxy). Of the eight positions with a consensus count of two or larger in gp120 and outside the variable loops, two cannot be mapped to the structure (HXB2 positions 47 and 63), two are located in the b12 binding site (positions 369 and 432), three are surrounding the b12 binding side (positions 352, 360, and 460), and one is located on the opposite site of the b12 binding site (position 446). Of note, position 446 is a potential glycosylation site, and also, mutations at this position might influence the glycosylation status of position 444. Due to the trimeric structure of gp120 (see for instance Liu et al. (2008)), glycans attach at this position might influence binding of b12. Of note, two additional b12 binding sites appear among the top 20 list of one method (positions 372 (RF) and 386 (SVM)).

Figure 7.6 a) shows the PImp MDG from the RF 2F5 model. One positions (665) in gp41 receives outstanding importance and is located in the 2F5 epitope (Zwick et al., 2001). Among the top 20, two additional positions (662 and 667) belonging the linear epitope are discovered. Likewise, Figure 7.6 b) depicts the PImp MDG importance for the 4E10 model. Important positions are more scattered in the complete *Env* gene, than in the case of 2F5. The most important position is located in gp41 and the second most important one between the variable loops V3 and V4. With position 674, one known 4E10 binding site position (Zwick et al., 2001) is ranked third by the corrected MDG measure.

In vitro Validation

Table 7.3 lists the sets of mutations that were selected for experimental validation. With pathways of three mutations we aim at observing a substantial change in the b12 neutralization phenotype. Mutations in the signal peptide and in gp41 are used to verify, whether these positions are statistical artifacts or have biological impact. Further mutations are used to assess the effect of adding and removing potential glycosylation sites. The remain-

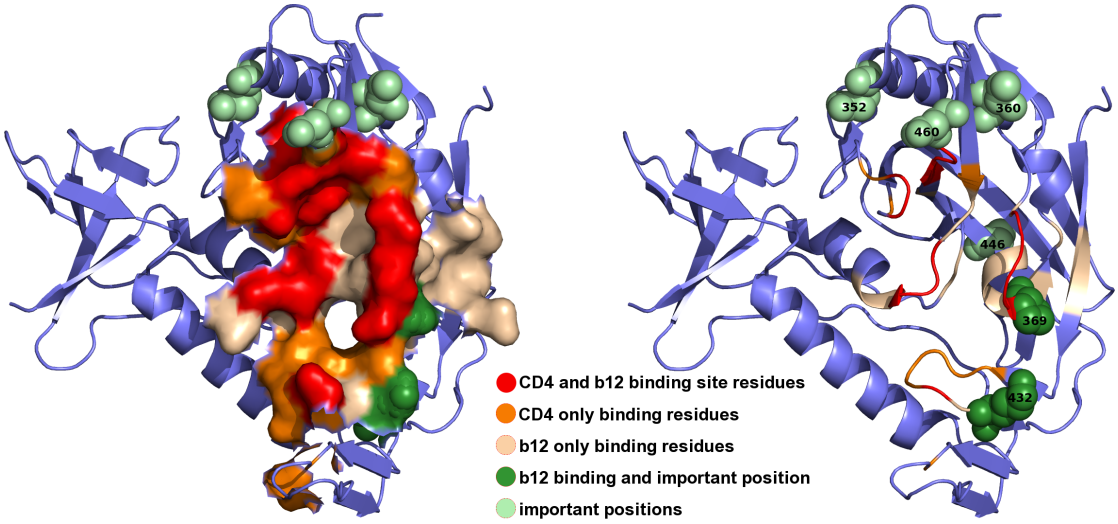
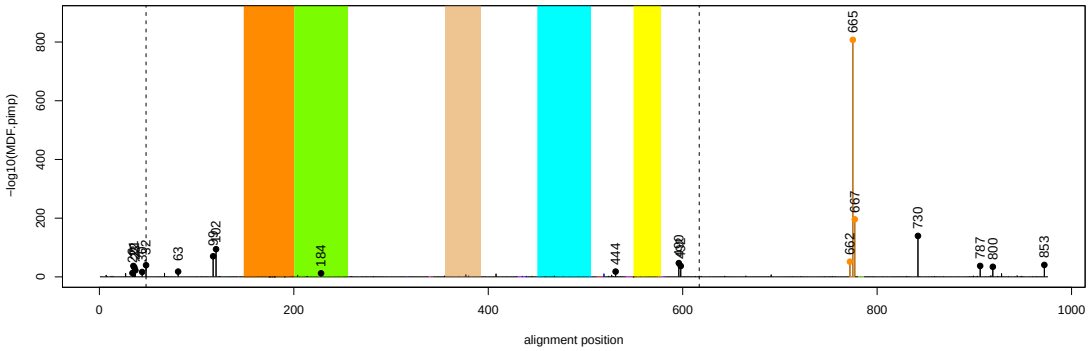
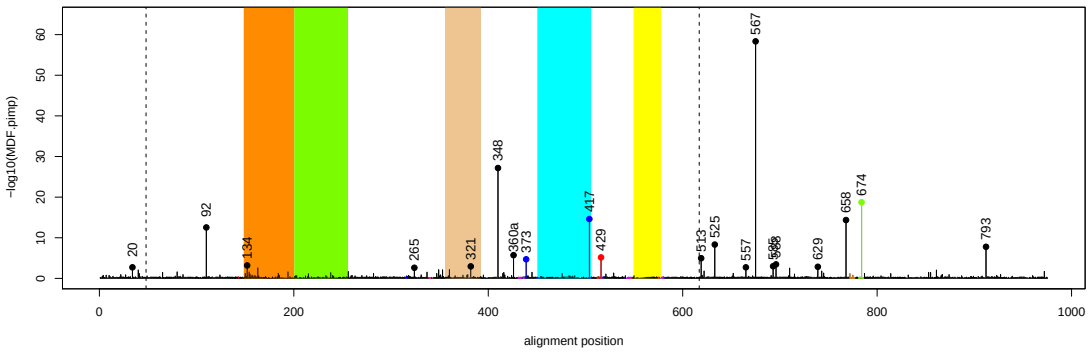


Figure 7.5: Important positions for b12 neutralization. Important positions were mapped to the structure of gp120 (PDB ID 2nxy). The right picture depicts the surface of b12 and CD4 binding sites, the left picture provides the location of the important positions on the HXB2 reference sequence.



(a) PImp MDG 2F5



(b) PImp MDG 4E10

Figure 7.6: Feature importance from models that predict 2F5 and 4E10 neutralization: PImp of RF MDG for 2F5 (a) and 4E10 (b). For further information see caption of Figure 7.4.

ing experiments shall clarify the effect of mutations close to the b12 binding site and within conserved parts of V1 and V2. For further information on the mutation selection process please refer to the Master thesis of Fabian Müller (Müller, 2009).

The mutagenesis and neutralization experiments are currently being conduced at Prof. Dr. Klaus Überla's laboratory at the Ruhr-University in Bochum, Germany.

7.3.3 Discussion

We trained statistical learning models that predict the neutralization of viral variants by three nAbs based on the *Env* amino acid sequence. With AUC values from 0.717 to 0.914 all models achieved good to very good prediction performance using the sequence encoding.

The removal of sequence information of the five variable loops resulted in a sensible loss of performance for the RF model in the prediction of neutralization by b12. The dependence on information on the variable loop V2 for predicting the b12 phenotype is confirmed by the re-established performance after providing the V2 sequence information and by the consensus variable importance (Figure 7.4). Here, several positions within V2 receive high importance values. The influence of the V2 loop on the neutralization efficacy by b12 was recognized soon after the discovery of b12 (Roben et al., 1994). Of note, the corrected MDG measure concurs better with the increase in performance due to the addition of single loops than the uncorrected MDG measure. The variable loop features derived from sequence information that aimed at capturing physicochemical properties of the loops failed to improve the model. In fact, inclusion of these features resulted in decreased performance with both statistical learning methods.

Several of the important sequence positions in the conserved region of gp120 were located either in the b12 binding site or structurally close to the b12 and CD4 binding site. The virus has to maintain the capability to bind the CD4 receptor, hence mutations in the CD4 binding site are unlikely. Consequently, substitutions near the or at the b12 binding site, but not at the CD4 binding site, are likely to affect susceptibility to nAb b12.

For other important positions that do not afford an obvious explanation, we will conduct *in vitro* neutralization experiments that will either confirm their importance or mark them as an artifact of the statistical methods.

The interpretation of the 2F5 prediction models revealed that decisions of the RF model were mainly based on a single position within the known 2F5 epitope. Also the analysis of the 4E10 RF model reveals a known position in the binding epitope among the top three important positions. Given that the cutoff selection for 4E10 was not as straightforward as for the other two antibodies, the rediscovery of a known epitope is reassuring. The agreement between the positions relevant for 2F5 and 4E10 neutralization and their known epitopes confirm the validity of the approach.

Unfortunately, the described method is susceptible to evolutionary bias. Here, evolutionary bias occurs in the form of two (or more) mutations, with one mutation having a causal link to the phenotype, and the other mutations merely being correlated to the first one due to emergence in a common ancestor or accumulation in the same lineage. Of course, the statistical learning models cannot discriminate between a causal link and a correlation to the phenotype. These correlations might explain the positions falsely assessed as being important by the method. Mutations that are located in the signal peptide or in gp41

isolate	accession	mutations	expected effect	measured effect	comment
6535	AY835438	T132S, L134A	more resistant		path of three mutations
6535	AY835438	T132S, L134A, P369L	more resistant		path of three mutations
6535	AY835438	D386N	more resistant (SVM)		add glycosylation; RF model shows no effect
ZM197M	DQ388515	K360V, L369P	more susceptible		path of three mutations
ZM197M	DQ388515	K360V, L369P, R432K	more susceptible		path of three mutations
ZM197M	DQ388515	N460I	more susceptible		remove glycosylation
CAP210.2.00.E8	DQ435683	S132T, A134L	more susceptible		path of three mutations
CAP210.2.00.E8	DQ435683	S132T, A134L, L369P	more susceptible		path of three mutations
Du172	DQ411853	V134A	more resistant		mutations in V1, V2
Du172	DQ411853	V134A, D185N	more resistant		mutations in V1, V2
Du172	DQ411853	K446T	more resistant		close to two glycosylation sites
ZM233M	DQ388517	T185D	more susceptible		mutations in V1, V2
ZM233M	DQ388517	T185D, M161T	more susceptible		mutations in V1, V2
CAAN	AY835452	N386D	more susceptible		remove glycosylation
Du156	DQ411852	H352I	more resistant		near binding site
ZM53M.PB12	AY423984	N462S	more susceptible		remove glycosylation
THRO	AY835448	K5bQ	more resistant		mutation in signal peptide
TRJO	AY835448	D750N	more susceptible (SVM)		mutation in gp41, RF model shows no effect

Table 7.3: Site-directed mutagenesis experiments. The table lists the name of the isolate, its accession number, the mutations that have to be introduced (sequence position is given in reference to HXB2; capital letters before and after the position denote wild type amino acid found in the isolate and the mutated amino acid, respectively), the expected effect, the measured effect, and the effect that is to be tested (comment).

and are supposed to be determinants of b12 binding are likely examples for this bias. For instance, position 5b in the signal peptide is highly correlated with position 352 in the proximity of the b12 binding site (Müller, 2009). Moreover, best performance was usually achieved when the sequence information was restricted to the *Env* region targeted by the nAb. Models based on the remaining regions that accommodate (potential) correlated mutations were clearly inferior. With such a restricted dataset a cautious interpretation of the results is required. Also the clade membership is a strong indicator for this evolutionary bias. In order to study a less biased dataset we repeated the RF analysis of b12 neutralization prediction with a dataset restricted to sequences from clades B, C, and D that show a balanced neutralization phenotype. The dataset comprised 106 instances, and the AUC of the sequence encoding was 0.75 ± 0.037, corresponding to a slight decrease in comparison to the full dataset. The qualitative findings, however, could be confirmed: decrease of performance without information on variable loops (0.67 ± 0.038), models restricted to signal peptide (0.578 ± 0.054) and gp41 (0.63 ± 0.043) achieved worst performance, and sequence information on the V2 loop is pivotal for prediction performance (0.764 ± 0.028).

A drawback of this study was the fact that experimental data were pooled from four different publications using different assays. Different neutralization assays may show a different phenotype in combination with some or all nAbs (Fenyö et al., 2009). However, dichotomizing the neutralization phenotype to "neutralizing" and "non-neutralizing" using a cutoff is likely to remove most of the variation originating from different assays. Moreover, we did not use the misclassification rate as a performance measure, but the area under the ROC curve. The latter measure can be interpreted as the probability that a randomly chosen positive (neutralizing) sample receives a higher prediction score than a randomly selected negative (non-neutralizing) sample. More precisely, the AUC is computed by first sorting all samples according to their predicted value. Hence, even if, due to experimental variation, a sample is mislabeled, it is less likely to substantially perturb the performance measure. Simply because samples close to the cutoff are more likely to receive prediction values in-between clear negatives and clear positives.

Despite the fact that only relatively few training data were available, we were able to build statistical models with high predictive power. Furthermore, interpretation of these models revealed sequence positions that were known to be in the binding site of the studied nAbs. Therefore, this seems to be a valid approach for identifying residues that determine antibody neutralization.

In terms of vaccine design our approach can be used of identify viral variants that are potential immunogens for eliciting a broadly neutralizing antibody. More precisely, assuming that variants that are very susceptible to a given nAb are also potential potent immunogens, which elicit the nAb *in vivo*, our *in silico* prediction method can help to screen large available sequence databases for interesting candidate immunogens.

7.4 Outlook

In the search for new broadly neutralizing antibodies, comparatively large panels of different viruses are tested against the new candidate antibody, and the three antibodies studied here are typically used as a reference, see for instance the work by Walker et al. (2009). Hence, if made public, additional data can be used to improve the prediction models intro-

duced here, or alternatively, the additional samples can be used as an independent validation set. Moreover, our approach can also be extended to study the genotype-phenotype relation of newly discovered nAbs.

An additional approach to improving models that predict neutralization by antibodies is the utilization of features derived from the protein structure instead of the sequence. Clearly, successful binding of a nAb to its epitope on the surface of the viral spike is determined by shape and chemical complementarity. The protein sequence is only an indicator for the structure of the protein *in vivo*. As a consequence of its high impact for vaccine design, the structure of the gp120 core was studied in great detail, and today there are a number of experimentally generated protein crystal structures available. Some of these structures contain even the variable loop V3 (Huang et al., 2005) that is crucial in coreceptor binding. And, precisely this structure of the V3 loop was previously used to derive structural descriptors for the prediction of coreceptor usage (Sander et al., 2007). In this previous work, a statistical learning model using structural descriptors could outperform a sequence-based approach, and a combination of both, structure and sequence, provided the best model. Likewise, the prediction of neutralization by nAbs could benefit from the application of structural descriptors. This approach is of particular interest for b12, as a gp120 structure in complex with b12 is available (Zhou et al., 2007).

The proposed approach to model nAb neutralization will be applied to assist vaccine development. The knowledge extracted from the statistical learning models and the models themselves can help to construct a large library of *Env* variants that are likely to be neutralized by the nAb. The library of *Env* variants is then used in an immunogen optimization protocol to develop an HIV vaccine. In particular, the variants in the library will be screened for their ability to evoke a strong immune response and additional statistical models will also be implemented for refining the search of candidate immunogens.

The approach for finding immunogens that elicit a desired type of antibodies can be extended to find nAbs for other pathogens or even for specific types of cancer.

8 Conclusion and Outlook

We have presented methods for improving HIV patient care. The developed models use techniques from statistical learning and focus on the prediction of response to combination treatment as opposed to resistance against single drugs.

In particular, we investigated the benefit of features encoding the viral evolution during therapy. The most predictive feature was the genetic barrier to drug resistance, which quantifies the likelihood that the virus will escape from the drug by developing further resistance mutations. A statistical model was trained on clinical data from the United States of America comprising viral genotype, treatment, and outcome of the regimen. The use of the genetic barrier together with indicators for mutations in protease and reverse transcriptase as well as indicators for the usage of antiretroviral drugs in the regimen outperforms commonly used rules-based systems in finding successful treatments. Precisely, a retrospective analysis on data from Europe demonstrated that our tool geno2pheno-THEO reaches a true positive rate (TPR) of 64% at a false positive rate (FPR) of 20%. By contrast, the rules-based approaches yield only 44% to 48% TPR at the same FPR, i.e. our method identifies up to 20% more successful combination treatments than standard approaches.

Within the EuResist project, which integrated multiple HIV resistance databases storing treatment related information, we further developed geno2pheno-THEO. In particular, we improved the genetic barrier by relying on a large number of predicted resistance phenotypes rather than a small number of measured phenotypes for defining mutation patterns that correspond to complete resistance. And, more importantly, we made use of the patient's treatment history (i.e. drugs the patient has been exposed to) and baseline viral load to improve prediction to combination therapy. Apart from geno2pheno-THEO, which is (due to the employed features based on viral evolution) also referred to as evolutionary engine, two other prediction models were developed within the project. In order to provide a single prediction for one request, we explored ways for optimally combining the output of all three engines. Here, it turned out that simply using the mean of all predictions is an efficient and robust way. On a small set of 25 treatment change episode the predictions of the combined prediction engine were compared to the predictions of ten international human HIV treatment experts. Our system performed as well as the best human expert and also as well as the consensus of all human experts.

The availability of three prediction engines trained on the same training data unveiled a serious limitation of the current definition to of treatment response. The standard datum definition focuses on short-term response measured at eight weeks (4-12) of therapy. The cutoff for success was a viral load below 500 cp/ml at the time-point closest to eight weeks. A substantial number of treatments, however, which were labeled as failure, reached a VL below the threshold at a later time during the treatment. Interestingly, all three engines captured that trend frequently, and disagreed with the label of the treatment.

As a consequence, we studied modified treatment response definitions: sustained response at 24 weeks (16-32) of treatment, the area under the viral load curve during one year of therapy, and a labeling that combines short-term and sustained response. The latter definition rejects treatments that have a discordant response at eight and 24 weeks of therapy, and therefore filters instances with potential adherence problems. In this setting our methodology achieves an AUC of 0.85 compared to an AUC of 0.77 on the original response definition (using the richest model).

Furthermore, we addressed the update-problem of data-driven decision support systems. Briefly, the update-problem originates from the fact that data-collection efforts providing the training data for those systems lag behind the introduction of novel drugs. More precisely, novel drugs tend to be prescribed to patients in an advanced state of the disease. Hence, it takes time to collect a sufficient amount of training data for the data-driven systems. In order to circumvent that problem, we introduced a new covariate, which represents the activity of novel drugs in the regimen. Here, activity of novel drugs is assessed by a rules-based system and when applied to treatments comprising newly licensed drugs the modified system outperforms the purely rules-based approach – manifested in an increase of the TPR by approximately 25% at a FPR of 20%.

We also investigated treatment history as a potential replacement for the viral sequence. The rationale behind the approach is that the prescribed drugs shape the genetic makeup of the viral population. Consequently, the treatment history, i.e. list of drugs the patient was exposed to, is a strong predictor of treatment response. Here, we found that prediction models based on history information alone are slightly inferior (AUC of 0.75) to genotype-based predictions (AUC of 0.77). History and genotype information seem to be partially complementary, since combining genotype-based and history-based predictions resulted in a further increase in performance (AUC of 0.79), regardless of the mode of combination.

We made first progress towards the sequencing of anti-HIV therapies. To this end, we used methods from large vocabulary language processing to develop a framework that allows rapid search for likely viral variants arising at treatment failure. We developed five mutation models on the basis of *in vitro* phenotypic resistance data and emergence of resistance mutations observed *in vivo*. The framework was challenged to separate successful from failing treatments on the basis of mutations caused by the immediately preceding regimen. The performance was moderate (AUC of 0.63) but constitutes a substantial improvement over the baseline (AUC of 0.55). Additionally, we could demonstrate that resistance development is correctly captured within drug classes. Simulating resistance development during antiretroviral treatment with a maximum of five mutations takes in the new framework only 3 seconds while keeping track of the 100 most likely viral variants. Hence, the framework affords web services with acceptable response time.

Finally, we developed statistical models that predict neutralization of HIV variants by three broadly neutralizing antibodies based on the *Env* genotype. The predictive performance of the models using sequence information of complete *Env* ranged from 0.72 AUC to 0.91 AUC, depending on the antibody. Interpretation of the statistical learning methods for all three antibodies recovered at least one known position located in the antibody-specific epitope. For the antibody b12, sequence information on the variable loop V2 is pivotal to maintain model performance. Furthermore, selected mutations from the b12 model are currently tested *in vitro* to further validate the approach. Knowledge derived from the

statistical models and the models themselves can be used to build focused libraries that screen *in vitro* for potent immunogens. Our methodology can, in theory, be applied to novel HIV targeting neutralizing antibodies or even be extended to other pathogens and cancer.

The work described in this thesis resulted in a number of freely available web services. Geno2pheno-THEO is available as part of the GENO2PHENO web suite (http://www.geno2pheno.org). Part of this web suite is also a prediction system for resistance against integrase inhibitors and can be directly accessed at http://integrase.geno2pheno.org. Finally, our contribution to the EURESIST prediction engine can be queried as part of the complete system using the website http://engine.euresist.org.

Outlook

A most straightforward direction is the extension of the models developed in this thesis to the newly available drugs from different classes. In particular, integrase inhibitors and entry inhibitors. A further challenge is the development of prediction tools for treatment naïve patients. This group faces the unique problem of transmitted drug resistance mutations. Transmitted mutations may, due to replicative disadvantage, vanish from the majority of the viral population, and consequently, go unnoticed using standard interpretation tools. However, traces of those mutations in viral genome are likely to be present at nucleotide or in rare cases even at amino acid level. In addition, it is of particular interest to infer drug adherence problems from stored treatment data (primarily viral load trajectories). Such information will not only enhance response prediction, but, more importantly, will also provide the means for treating clinicians to identify patients having trouble in correctly taking their medication. Here, one can provide a tool for detering development of resistance mutations.

Personalized treatment in the HIV field has, so far, only focused on the interaction between the drug and virus. With our approaches for predicting response to antiretroviral drugs *in vivo*, a third player enters the interaction network: the patient. Until now, the patient was almost neglected in the prediction of treatment response. However, in order to make further advancements in personalized anti-HIV therapy the interplay between virus and host as well as between drugs and host has to be considered. Indeed, initial studies focusing on the relationship between human genomics and the control of HIV infection have already been conduced (Telenti and Goldstein, 2006). The next steps in this direction should involve the systematic exploitation of available resistance databases that offer a wealth of treatment data. Hence, it is comparatively uncomplicated to updated these databases gradually with the missing human genetics data. This will offer a wealth of possibilities, including pharmocogenetic studies uncovering the interaction between drugs and host-related differences in the involved metabolic pathways (Telenti and Zanger, 2008)[1]. The resulting knowledge will help to provide the right dosage of drugs for the patient in addition to the right drug attacking the virus. The correct dosing is of particular interest, since drug-related side-effects are a major obstacle in the life-long HIV treatment. Also the use of next generation sequencing techniques will allow to study the evolution of the

[1] http://www.hiv-pharmacogenomics.org/

viral population during treatment and result in more advanced decision support tools.

The evaluation of clinical decision support systems used for assessing drug resistance should be carried out on standardized publicly available datasets. To date, validation of such tools is based on retrospective analysis of clinical data – with different sets of data used for different tools. Consequently, the performance of the various number of tools is hard to compare. Especially since the datasets are prone to local differences in drug prescriptions or widely differing extent of pre-treatment of involved patients. Hence, a publicly available benchmark dataset with fair distributions of drug usage and certain level of pre-treatment (i.e. not to many "simple cases") is needed. Moreover, publicly available datasets attract a variety of scientists applying available methods or developing new tailored methods to the problem. For instance, the freely available (standardized) genotype-phenotype dataset (Rhee et al., 2006) caused a number of follow-up publications (Saigo et al., 2007; Kjaer et al., 2008; Kierczak et al., 2009).

The methods developed in this theses can also be applied to optimization of treatment for other viral diseases such as Hepatitis B and Hepatitis C. While data-collection efforts and statistics will remain the same, the extension to these viral diseases is not straightforward, as clinically relevant definitions of response to treatment have to be identified and agreed upon. Furthermore, using matched genotype-phenotype data it is comparably simple to identify resistance mutations. Understanding their mechanism of resistance, however, still remains a challenge.

Bibliography

(2006). Denying science. *Nature Medicine*, 12(4):369.

(2008). The cost of silence? *Nature*, 456(7222):545.

Abbadessa, G., Accolla, R., Aiuti, F., Albini, A., Aldovini, A., Alfano, M., Antonelli, G., Bartholomew, C., Bentwich, Z., Bertazzoni, U., Berzofsky, J. A., Biberfeld, P., Boeri, E., Buonaguro, L., Buonaguro, F. M., Bukrinsky, M., Burny, A., Caruso, A., Cassol, S., Chandra, P., Ceccherini-Nelli, L., Chieco-Bianchi, L., Clerici, M., Colombini-Hatch, S., de Giuli Morghen, C., de Maria, A., de Rossi, A., Dierich, M., Della-Favera, R., Dolei, A., Douek, D., Erfle, V., Felber, B., Fiorentini, S., Franchini, G., Gershoni, J. M., Gotch, F., Green, P., Greene, W. C., Hall, W., Haseltine, W., Jacobson, S., Kallings, L. O., Kalyanaraman, V. S., Katinger, H., Khalili, K., Klein, G., Klein, E., Klotman, M., Klotman, P., Kotler, M., Kurth, R., Lafeuillade, A., Placa, M. L., Lewis, J., Lillo, F., Lisziewicz, J., Lomonico, A., Lopalco, L., Lori, F., Lusso, P., Macchi, B., Malim, M., Margolis, L., Markham, P. D., McClure, M., Miller, N., Mingari, M. C., Moretta, L., Noonan, D., O'Brien, S., Okamoto, T., Pal, R., Palese, P., Panet, A., Pantaleo, G., Pavlakis, G., Pistello, M., Plotkin, S., Poli, G., Pomerantz, R., Radaelli, A., Robertguroff, M., Roederer, M., Sarngadharan, M. G., Schols, D., Secchiero, P., Shearer, G., Siccardi, A., Stevenson, M., Svoboda, J., Tartaglia, J., Torelli, G., Tornesello, M. L., Tschachler, E., Vaccarezza, M., Vallbracht, A., van Lunzen, J., Varnier, O., Vicenzi, E., von Melchner, H., Witz, I., Zagury, D., Zagury, J.-F., Zauli, G., and Zipeto, D. (2009). Unsung hero Robert C. Gallo. *Science*, 323(5911):206–7.

Agopian, A., Gros, E., Aldrian-Herrada, G., Bosquet, N., Clayette, P., and Divita, G. (2009). A new generation of peptide-based inhibitors targeting HIV-1 reverse transcriptase conformational flexibility. *J Biol Chem*, 284(1):254–64.

Agrawal, R., Imielinski, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. *In Proceedings of ACM SIGMOD Conference on Management of Data*, pages 207–216.

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules. *Proc. 20th Int. Conf. Very Large Data Bases*, pages 487–499.

Aharoni, E., Altmann, A., b Borgulya, D'utilia, R., Incardona, F., Kaiser, R., Kent, C., Lengauer, T., Neuvirth, H., Peres, Y., Petroczi, A., Prosperi, M., Rosen-Zvi, M., Schulter, E., Sing, T., Sonnerborg, A., Thompson, R., and Zazzi, M. (2007). Integration of viral genomics with clinical data to predict response to anti-HIV treatment. *IST-Africa 2007 Conference Proceedings, Paul Cunningham and Miriam Cunningham (Eds)*.

Aho, A., Sethi, R., and Ullman, J. (1986). Compilers: Principles, techniques, and tools. *Addison-Wesley*.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, AC-19:716–723.

Allauzen, C. and Mohri, M. (2002). On the determinizability of weighted automata and transducers. *In Proceedings of the workshop Weighted Automata: Theory and Application (WATA)*.

Allauzen, C., Riley, M., Schalkwyk, J., Skut, W., and Mohri, M. (2007). OpenFst: A general and efficient weighted finite-state transducer library. *CIAA 2007, Lecture Notes in Computer Science*, 4783:11–23.

Altmann, A., Beerenwinkel, N., Sing, T., Savenkov, I., Däumer, M., Kaiser, R., Rhee, S.-Y., Fessel, W. J., Shafer, R. W., and Lengauer, T. (2007a). Improved prediction of response to antiretroviral combination therapy using the genetic barrier to drug resistance. *Antivir Ther (Lond)*, 12(2):169–78.

Altmann, A., Däumer, M., Beerenwinkel, N., Peres, Y., Schülter, E., Büch, J., Rhee, S.-Y., Sönnerborg, A., Fessel, W. J., Shafer, R. W., Zazzi, M., Kaiser, R., and Lengauer, T. (2009a). Predicting the response to combination antiretroviral therapy: retrospective validation of geno2pheno-THEO on a large clinical database. *J Infect Dis*, 199(7):999–1006.

Altmann, A., Sing, T., Vermeiren, H., Winters, B., Craenenbroeck, E. V., der Borght, K. V., Rhee, S. Y., Shafer, R. W., Schuelter, E., Kaiser, R., Peres, Y., Sonnerborg, A., Fessel, W. J., Incardona, F., Zazzi, M., Bacheler, L., Vlijmen, H. V., and Lengauer, T. (2007b). Inferring virological response from genotype: with or without predicted phenotypes? *Antivir Ther (Lond)*, 12(5):S169–S169.

Altmann, A., Sing, T., Vermeiren, H., Winters, B., Craenenbroeck, E. V., der Borght, K. V., Rhee, S.-Y., Shafer, R. W., Schülter, E., Kaiser, R., Peres, Y., Sönnerborg, A., Fessel, W. J., Incardona, F., Zazzi, M., Bacheler, L., Vlijmen, H. V., and Lengauer, T. (2009b). Advantages of predicted phenotypes and statistical learning models in inferring virological response to antiretroviral therapy from HIV genotype. *Antivir Ther (Lond)*, 14(2):273–83.

Altmann, A., Thielen, A., Frentz, D., Laethem, K. V., Bracciale, L., Incardorna, F., Sönnerborg, A., Zazzi, M., Kaiser, R., and Lengauer, T. (2009c). Keeping models that predict response to antiretroviral therapy up-to-date: fusion of pure data-driven approaches with rules-based methods. *Reviews in Antiviral Therapy*, page A92.

Altmann, A., Toloşi, L., Sander, O., and Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–7.

Andreola, M. L., Pileur, F., Calmels, C., Ventura, M., Tarrago-Litvak, L., Toulmé, J. J., and Litvak, S. (2001). DNA aptamers selected against the HIV-1 RNase H display in vitro antiviral activity. *Biochemistry*, 40(34):10087–94.

Arthos, J., Cicala, C., Martinelli, E., Macleod, K., Ryk, D. V., Wei, D., Xiao, Z., Veenstra, T. D., Conrad, T. P., Lempicki, R. A., McLaughlin, S., Pascuccio, M., Gopaul, R.,

McNally, J., Cruz, C. C., Censoplano, N., Chung, E., Reitano, K. N., Kottilil, S., Goode, D. J., and Fauci, A. S. (2008). HIV-1 envelope protein binds to and signals through integrin alpha4 beta7, the gut mucosal homing receptor for peripheral T cells. *Nat Immunol*, 9(3):301–9.

Baelen, K. V., Salzwedel, K., Rondelez, E., Eygen, V. V., Vos, S. D., Verheyen, A., Steegen, K., Verlinden, Y., Allaway, G. P., and Stuyver, L. J. (2009). Susceptibility of human immunodeficiency virus type 1 to the maturation inhibitor bevirimat is modulated by baseline polymorphisms in gag spacer peptide 1. *Antimicrob Agents Chemother*, 53(5):2185–8.

Barouch, D. H. (2008). Challenges in the development of an HIV-1 vaccine. *Nature*, 455(7213):613–9.

Barré-Sinoussi, F., Chermann, J. C., Rey, F., Nugeyre, M. T., Chamaret, S., Gruest, J., Dauguet, C., Axler-Blin, C., Vézinet-Brun, F., Rouzioux, C., Rozenbaum, W., and Montagnier, L. (1983). Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science*, 220(4599):868–71.

Beerenwinkel, N., Däumer, M., Oette, M., Korn, K., Hoffmann, D., Kaiser, R., Lengauer, T., Selbig, J., and Walter, H. (2003a). Geno2pheno: Estimating phenotypic drug resistance from HIV-1 genotypes. *Nucleic Acids Res*, 31(13):3850–5.

Beerenwinkel, N., Däumer, M., Sing, T., Rahnenfuhrer, J., Lengauer, T., Selbig, J., Hoffmann, D., and Kaiser, R. (2005a). Estimating HIV evolutionary pathways and the genetic barrier to drug resistance. *J Infect Dis*, 191(11):1953–60.

Beerenwinkel, N., Lengauer, T., Däumer, M., Kaiser, R., Walter, H., Korn, K., Hoffmann, D., and Selbig, J. (2003b). Methods for optimizing antiviral combination therapies. *Bioinformatics*, 19 Suppl 1:i16–25.

Beerenwinkel, N., Rahnenführer, J., Däumer, M., Hoffmann, D., Kaiser, R., Selbig, J., and Lengauer, T. (2005b). Learning multiple evolutionary pathways from cross-sectional data. *J Comput Biol*, 12(6):584–98.

Beerenwinkel, N., Rahnenführer, J., Kaiser, R., Hoffmann, D., Selbig, J., and Lengauer, T. (2005c). Mtreemix: a software package for learning and using mixture models of mutagenetic trees. *Bioinformatics*, 21(9):2106–7.

Beerenwinkel, N., Schmidt, B., Walter, H., Kaiser, R., Lengauer, T., Hoffmann, D., Korn, K., and Selbig, J. (2002). Diversity and complexity of HIV-1 drug resistance: a bioinformatics approach to predicting phenotype from genotype. *Proc Natl Acad Sci USA*, 99(12):8271–6.

Bellman, R. (1958). On a routing problem. *Quaterly of Applied Mathematics*, 16(1):87–90.

Berger, E. A., Murphy, P. M., and Farber, J. M. (1999). Chemokine receptors as HIV-1 coreceptors: roles in viral entry, tropism, and disease. *Annu Rev Immunol*, 17:657–700.

Bickel, S., Bogojeska, J., Lengauer, T., and Scheffer, T. (2008). Multi-task learning for HIV therapy screening. *Proceedings of the 25th International Conference on Machine Learning*, pages 56–63.

Binley, J. M., Wrin, T., Korber, B., Zwick, M. B., Wang, M., Chappey, C., Stiegler, G., Kunert, R., Zolla-Pazner, S., Katinger, H., Petropoulos, C. J., and Burton, D. R. (2004). Comprehensive cross-clade neutralization analysis of a panel of anti-human immunodeficiency virus type 1 monoclonal antibodies. *J Virol*, 78(23):13232–52.

Boffito, M., Acosta, E., Burger, D., Fletcher, C. V., Flexner, C., Garaffo, R., Gatti, G., Kurowski, M., Perno, C. F., Peytavin, G., Regazzi, M., and Back, D. (2005). Therapeutic drug monitoring and drug-drug interactions involving antiretroviral drugs. *Antivir Ther (Lond)*, 10(4):469–77.

Boser, B., Guyon, I., and Vapnik, V. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on Computational Learning Theory*, pages 144–152.

Bratt, G., Karlsson, A., Leandersson, A. C., Albert, J., Wahren, B., and Sandström, E. (1998). Treatment history and baseline viral load, but not viral tropism or CCR-5 genotype, influence prolonged antiviral efficacy of highly active antiretroviral treatment. *AIDS*, 12(16):2193–202.

Breiman, L. (2001). Random forests. *Machine learning*.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). Classification and regression trees. *Wadsworth, Belmont*.

Bromley, S. K., Burack, W. R., Johnson, K. G., Somersalo, K., Sims, T. N., Sumen, C., Davis, M. M., Shaw, A. S., Allen, P. M., and Dustin, M. L. (2001). The immunological synapse. *Annu Rev Immunol*, 19:375–96.

Brun-Vézinet, F., Costagliola, D., Khaled, M. A., Calvez, V., Clavel, F., Clotet, B., Haubrich, R., Kempf, D., King, M., Kuritzkes, D., Lanier, R., Miller, M., Miller, V., Phillips, A., Pillay, D., Schapiro, J., Scott, J., Shafer, R., Zazzi, M., Zolopa, A., and DeGruttola, V. (2004). Clinically validated genotype analysis: guiding principles and statistical concerns. *Antivir Ther (Lond)*, 9(4):465–78.

Burton, D. R., Desrosiers, R. C., Doms, R. W., Koff, W. C., Kwong, P. D., Moore, J. P., Nabel, G. J., Sodroski, J., Wilson, I. A., and Wyatt, R. T. (2004). HIV vaccine design and the neutralizing antibody problem. *Nat Immunol*, 5(3):233–6.

Burton, D. R., Pyati, J., Koduri, R., Sharp, S. J., Thornton, G. B., Parren, P. W., Sawyer, L. S., Hendry, R. M., Dunlop, N., and Nara, P. L. (1994). Efficient neutralization of primary isolates of HIV-1 by a recombinant human monoclonal antibody. *Science*, 266(5187):1024–7.

Carries, S., Muller, F., Muller, F. J., Morroni, C., and Wilson, D. (2007). Characteristics, treatment, and antiretroviral prophylaxis adherence of south african rape survivors. *J Acquir Immune Defic Syndr*, 46(1):68–71.

Chang, C. and Lin, C. (2001). LIBSVM: a library for support vector machines. *Software avaliable at http://www.csie.ntu.edu.tw/~cjlin/libsvm*.

Chapelle, O. and Zien, A. (2005). Semi-supervised classification by low density separation. *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics*, pages 57–64.

Chen, J. C., Krucinski, J., Miercke, L. J., Finer-Moore, J. S., Tang, A. H., Leavitt, A. D., and Stroud, R. M. (2000). Crystal structure of the HIV-1 integrase catalytic core and C-terminal domains: a model for viral DNA binding. *Proc Natl Acad Sci USA*, 97(15):8233–8.

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

Cozzi-Lepri, A., Phillips, A. N., Ruiz, L., Clotet, B., Loveday, C., Kjaer, J., Mens, H., Clumeck, N., Viksna, L., Antunes, F., Machala, L., Lundgren, J. D., and Group, E. S. (2007). Evolution of drug resistance in HIV-infected patients remaining on a virologically failing combination antiretroviral therapy regimen. *AIDS*, 21(6):721–32.

Crum, N. F., Riffenburgh, R. H., Wegner, S., Agan, B. K., Tasker, S. A., Spooner, K. M., Armstrong, A. W., Fraser, S., Wallace, M. R., and Consortium, T. A. C. (2006). Comparisons of causes of death and mortality rates among HIV-infected persons: analysis of the pre-, early, and late HAART (highly active antiretroviral therapy) eras. *J Acquir Immune Defic Syndr*, 41(2):194–200.

Dalgleish, A. G., Beverley, P. C., Clapham, P. R., Crawford, D. H., Greaves, M. F., and Weiss, R. A. (1984). The CD4 (T4) antigen is an essential component of the receptor for the AIDS retrovirus. *Nature*, 312(5996):763–7.

Däumer, M., Beerenwinkel, N., Hoffmann, D., Selbig, J., Lengauer, T., Oette, M., Fätkenheuer, G., Rockstroh, J. K., Pfister, H. J., and Kaiser, R. (2007). Geno2pheno: Determination of clinically relevant cut-offs. *European Journal of Medical Research*, 12(Supplement III):1–2.

Däumer, M. P., Kaiser, R., Klein, R., Lengauer, T., Thiele, B., and Thielen, A. (2008). Inferring viral tropism from genotype with massively parallel sequencing: qualitative and quantitative analysis. *Antivir Ther (Lond)*, 13(4):A101–A101.

De Luca, A., Cingolani, A., Giambenedetto, S. D., Trotta, M. P., Baldini, F., Rizzo, M. G., Bertoli, A., Liuzzi, G., Narciso, P., Murri, R., Ammassari, A., Perno, C. F., and Antinori, A. (2003). Variable prediction of antiretroviral treatment outcome by different systems for interpreting genotypic human immunodeficiency virus type 1 drug resistance. *J Infect Dis*, 187(12):1934–43.

De Luca, A., Giambenedetto, S. D., Romano, L., Gonnelli, A., Corsi, P., Baldari, M., Pietro, M. D., Menzo, S., Francisci, D., Almi, P., Zazzi, M., and Group, A. R. C. A. S. (2006). Frequency and treatment-related predictors of thymidine-analogue mutation patterns in HIV-1 isolates after unsuccessful antiretroviral therapy. *J Infect Dis*, 193(9):1219–22.

De Luca, A., Vendittelli, M., Baldini, F., Giambenedetto, S. D., Trotta, M. P., Cingolani, A., Bacarelli, A., Gori, C., Perno, C. F., Antinori, A., and Ulivi, G. (2004). Construction, training and clinical validation of an interpretation system for genotypic HIV-1 drug resistance based on fuzzy rules revised by virological outcomes. *Antivir Ther (Lond)*, 9(4):583–93.

Deforche, K., Camacho, R., Laethem, K. V., Lemey, P., Rambaut, A., Moreau, Y., and Vandamme, A.-M. (2008). Estimation of an in vivo fitness landscape experienced by HIV-1 under drug selective pressure useful for prediction of drug resistance evolution during treatment. *Bioinformatics*, 24(1):34–41.

DeFranco, A. L., Locksley, R. M., and Robertson, M. (2007). Immunity: the immune response in infectious and inflammatory disease. *New Science Press*.

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3):837–45.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39:1–38.

Descamps, D., Apetrei, C., Collin, G., Damond, F., Simon, F., and Brun-Vézinet, F. (1998). Naturally occurring decreased susceptibility of HIV-1 subtype G to protease inhibitors. *AIDS*, 12(9):1109–11.

Desper, R., Jiang, F., Kallioniemi, O. P., Moch, H., Papadimitriou, C. H., and Schäffer, A. A. (1999). Inferring tree models for oncogenesis from comparative genome hybridization data. *J Comput Biol*, 6(1):37–51.

Dijkstra, E. (1959). A note on two problems in connexion with graphs. *Numerische mathematik*, 1:269–271.

DiRienzo, A. G., DeGruttola, V., Larder, B., and Hertogs, K. (2003). Non-parametric methods to predict HIV drug susceptibility phenotype from genotype. *Statistics in medicine*, 22(17):2785–98.

Dorigo, M. and Gambardella, L. (1997). Ant colony system: A cooperative learning approach to the traveling salesman problem. *IEEE Transactions on evolutionary computation*.

Douek, D. C., Kwong, P. D., and Nabel, G. J. (2006). The rational design of an AIDS vaccine. *Cell*, 124(4):677–81.

Drăghici, S. and Potter, R. B. (2003). Predicting HIV drug resistance with neural networks. *Bioinformatics*, 19(1):98–107.

Durant, J., Clevenbergh, P., Halfon, P., Delgiudice, P., Porsin, S., Simonet, P., Montagne, N., Boucher, C. A., Schapiro, J. M., and Dellamonica, P. (1999). Drug-resistance genotyping in HIV-1 therapy: the VIRADAPT randomised controlled trial. *Lancet*, 353(9171):2195–9.

Dybul, M., Fauci, A., Bartlett, J., Kaplan, J., and Pau, A. (2002). Guidelines for using antiretroviral agents among HIV-infected adults and adolescents - the panel on clinical practices for treatment of HIV. *Ann Intern Med*, 137(5):381–433.

Edgar, R. C. (2004). Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, 32(5):1792–7.

Edmonds, J. (1967). Optimum branchings. *J Res Nbs B Math Sci*, B 71(4):233–&.

Esté, J. A. and Telenti, A. (2007). HIV entry inhibitors. *Lancet*, 370(9581):81–8.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, 27:861–874.

Fellay, J., Shianna, K. V., Ge, D., Colombo, S., Ledergerber, B., Weale, M., Zhang, K., Gumbs, C., Castagna, A., Cossarizza, A., Cozzi-Lepri, A., Luca, A. D., Easterbrook, P., Francioli, P., Mallal, S., Martinez-Picado, J., Miro, J. M., Obel, N., Smith, J. P., Wyniger, J., Descombes, P., Antonarakis, S. E., Letvin, N. L., McMichael, A. J., Haynes, B. F., Telenti, A., and Goldstein, D. B. (2007). A whole-genome association study of major determinants for host control of HIV-1. *Science*, 317(5840):944–7.

Fenyö, E. M., Heath, A., Dispinseri, S., Holmes, H., Lusso, P., Zolla-Pazner, S., Donners, H., Heyndrickx, L., Alcami, J., Bongertz, V., Jassoy, C., Malnati, M., Montefiori, D., Moog, C., Morris, L., Osmanov, S., Polonis, V., Sattentau, Q., Schuitemaker, H., Sutthent, R., Wrin, T., and Scarlatti, G. (2009). International network for comparison of HIV neutralization assays: the NeutNet report. *PLoS ONE*, 4(2):e4505.

Fields, B. N., Knipe, D. M., and Howley, P. M. (2007). Fields virology. *Lippincott Williams & Wilkins*.

Fiscus, S. A., Cheng, B., Crowe, S. M., Demeter, L., Jennings, C., Miller, V., Respess, R., Stevens, W., and for Collaborative HIV Research Alternative Viral Load Assay Working Group, F. (2006). HIV-1 viral load assays for resource-limited settings. *PLoS Med*, 3(10):e417.

Fisher, R. (1922). On the interpretation of χ^2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society*.

Freed, E. O. (2004). HIV-1 and the host cell: an intimate association. *Trends Microbiol*, 12(4):170–7.

Gallo, R. C. (2002). Historical essay. the early years of HIV/AIDS. *Science*, 298(5599):1728–30.

Gallo, R. C., Sarin, P. S., Gelmann, E. P., Robert-Guroff, M., Richardson, E., Kalyanaraman, V. S., Mann, D., Sidhu, G. D., Stahl, R. E., Zolla-Pazner, S., Leibowitch, J., and Popovic, M. (1983). Isolation of human T-cell leukemia virus in acquired immune deficiency syndrome (AIDS). *Science*, 220(4599):865–7.

Gao, F., Chen, Y., Levy, D. N., Conway, J. A., Kepler, T. B., and Hui, H. (2004). Unselected mutations in the human immunodeficiency virus type 1 genome are mostly nonsynonymous and often deleterious. *J Virol*, 78(5):2426–33.

Gareiss, P. C. and Miller, B. L. (2009). Ribosomal frameshifting: an emerging drug target for HIV. *Current opinion in investigational drugs (London, England : 2000)*, 10(2):121–8.

Gilks, C. F., Crowley, S., Ekpini, R., Gove, S., Perriens, J., Souteyrand, Y., Sutherland, D., Vitoria, M., Guerma, T., and Cock, K. D. (2006). The WHO public-health approach to antiretroviral treatment against HIV in resource-limited settings. *Lancet*, 368(9534):505–10.

Glass, T. R., Geest, S. D., Hirschel, B., Battegay, M., Furrer, H., Covassini, M., Vernazza, P. L., Bernasconi, E., Rickenboch, M., Weber, R., Bucher, H. C., and Study, S. H. C. (2008). Self-reported non-adherence to antiretroviral therapy repeatedly assessed by two questions predicts treatment failure in virologically suppressed patients. *Antivir Ther (Lond)*, 13(1):77–85.

Green, D. R., Droin, N., and Pinkoski, M. (2003). Activation-induced cell death in T cells. *Immunol Rev*, 193:70–81.

Grobler, J. A., Mckenna, P. M., Ly, S., Stillmock, K. A., Bahnck, C. M., Danovich, R. M., Dornadula, G., Hazuda, D. J., and Miller, M. D. (2009). Hiv integrase inhibitor dissociation rates correlate with efficacy in vitro. *Antivir Ther (Lond)*, 14(4):A27–A27.

Grossman, Z., Meier-Schellersheim, M., Paul, W. E., and Picker, L. J. (2006). Pathogenesis of HIV infection: what the virus spares is as important as what it destroys. *Nature Medicine*, 12(3):289–95.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, 46:389–422.

Hahn, B. H., Shaw, G. M., Taylor, M. E., Redfield, R. R., Markham, P. D., Salahuddin, S. Z., Wong-Staal, F., Gallo, R. C., Parks, E. S., and Parks, W. P. (1986). Genetic variation in HTLV-III/LAV over time in patients with AIDS or at risk for AIDS. *Science*, 232(4757):1548–53.

Hall, M. (1999). Correlation-based feature selection for machine learning. *Ph.D. Thesiss, Department of Computer Science, University of Waikato, Hamilton, New Zealand.*

Hammer, S. M., Eron, J. J., Reiss, P., Schooley, R. T., Thompson, M. A., Walmsley, S., Cahn, P., Fischl, M. A., Gatell, J. M., Hirsch, M. S., Jacobsen, D. M., Montaner, J. S. G., Richman, D. D., Yeni, P. G., Volberding, P. A., and Society-USA, I. A. (2008). Antiretroviral treatment of adult HIV infection: 2008 recommendations of the international AIDS society-USA panel. *JAMA*, 300(5):555–70.

Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundacker, H., Schooley, R. T., Haubrich, R. H., Henry, W. K., Lederman, M. M., Phair, J. P., Niu, M., Hirsch, M. S., and Merigan, T. C. (1996). A trial comparing nucleoside monotherapy with combination

therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *N Engl J Med*, 335(15):1081–90.

Han, J., Pei, J., Yin, Y., and Mao, R. (2004). Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data Mining and Knowledge Discovery*, 8:53–87.

Harrigan, P. R., Hogg, R. S., Dong, W. W. Y., Yip, B., Wynhoven, B., Woodward, J., Brumme, C. J., Brumme, Z. L., Mo, T., Alexander, C. S., and Montaner, J. S. G. (2005). Predictors of HIV drug-resistance mutations in a large antiretroviral-naive cohort initiating triple antiretroviral therapy. *J Infect Dis*, 191(3):339–47.

Hart, P., Nilsson, N., and Raphael, B. (1968). A formal basis for the heuristic determination of minimum cost paths. *IEEE transactions on Systems Science and Cybernetics*, SSC4(2):100–107.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2001). The elements of statistical learning: data mining, inference, and prediction. *Springer Series in Statistics*.

Hazenberg, M. D., Otto, S. A., van Benthem, B. H. B., Roos, M. T. L., Coutinho, R. A., Lange, J. M. A., Hamann, D., Prins, M., and Miedema, F. (2003). Persistent immune activation in HIV-1 infection is associated with progression to AIDS. *AIDS*, 17(13):1881–8.

Hazuda, D. J., Felock, P., Witmer, M., Wolfe, A., Stillmock, K., Grobler, J. A., Espeseth, A., Gabryelski, L., Schleif, W., Blau, C., and Miller, M. D. (2000). Inhibitors of strand transfer that prevent integration and inhibit HIV-1 replication in cells. *Science*, 287(5453):646–50.

Healy, B. and Degruttola, V. (2007). Hidden markov models for settings with interval-censored transition times and uncertain time origin: application to HIV genetic analysis. *Biostatistics*, 8(2):438–452.

Heeney, J. L., Dalgleish, A. G., and Weiss, R. A. (2006). Origins of HIV and the evolution of resistance to AIDS. *Science*, 313(5786):462–6.

Hetherington, L. (2004). The MIT finite-state transducer toolkit for speech and language processing. *Proceeding of the 8th International Conference on Spoken Language Processing*.

Huang, C., Tang, M., Zhang, M.-Y., Majeed, S., Montabana, E., Stanfield, R. L., Dimitrov, D. S., Korber, B., Sodroski, J., Wilson, I. A., Wyatt, R., and Kwong, P. D. (2005). Structure of a V3-containing HIV-1 gp120 core. *Science*, 310(5750):1025–8.

Huang, Y. and Suen, C. (1995). A method of combining multiple experts for the recognition of unconstrained handwritten numberals. *Ieee T Pattern Anal*, 17:90–94.

Jacks, T., Power, M. D., Masiarz, F. R., Luciw, P. A., Barr, P. J., and Varmus, H. E. (1988). Characterization of ribosomal frameshifting in HIV-1 gag-pol expression. *Nature*, 331(6153):280–3.

Jenwitheesuk, E. and Samudrala, R. (2005). Prediction of HIV-1 protease inhibitor resistance using a protein-inhibitor flexible docking approach. *Antivir Ther (Lond)*, 10(1):157–66.

Jiang, H., Deeks, S. G., Kuritzkes, D. R., Lallemant, M., Katzenstein, D., Albrecht, M., and DeGruttola, V. (2003). Assessing resistance costs of antiretroviral therapies via measures of future drug options. *J Infect Dis*, 188(7):1001–8.

Joachims, T. (1999). Transductive inference for text classifcation using support vector machines. *International Conference on Machine Learning (ICML)*.

Johnson, V. A., Brun-Vezinet, F., Clotet, B., Conway, B., Kuritzkes, D. R., Pillay, D., Schapiro, J. M., Telenti, A., and Richman, D. D. (2005). Update of the drug resistance mutations in HIV-1: Fall 2005. *Topics in HIV medicine : a publication of the International AIDS Society, USA*, 13(4):125–31.

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

Jordan, R., Gold, L., Cummins, C., and Hyde, C. (2002). Systematic review and meta-analysis of evidence for increasing numbers of drugs in antiretroviral combination therapy. *BMJ*, 324(7340):757.

Kamber, M. and Han, J. (2001). Data mining: Concepts and techniques. *Morgan Kaufmann*.

Kanki, P. J., Hamel, D. J., Sankalé, J. L., c Hsieh, C., Thior, I., Barin, F., Woodcock, S. A., Guèye-Ndiaye, A., Zhang, E., Montano, M., Siby, T., Marlink, R., NDoye, I., Essex, M. E., and MBoup, S. (1999). Human immunodeficiency virus type 1 subtypes differ in disease progression. *J Infect Dis*, 179(1):68–73.

Kanthak, S. and Ney, H. (2004). FSA: an efficient and flexible C++ toolkit for finite state automata using on-demand computation. *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pages 510–517.

Kantor, R., Katzenstein, D. A., Efron, B., Carvalho, A. P., Wynhoven, B., Cane, P., Clarke, J., Sirivichayakul, S., Soares, M. A., Snoeck, J., Pillay, C., Rudich, H., Rodrigues, R., Holguin, A., Ariyoshi, K., Bouzas, M. B., Cahn, P., Sugiura, W., Soriano, V., Brigido, L. F., Grossman, Z., Morris, L., Vandamme, A.-M., Tanuri, A., Phanuphak, P., Weber, J. N., Pillay, D., Harrigan, P. R., Camacho, R., Schapiro, J. M., and Shafer, R. W. (2005). Impact of HIV-1 subtype and antiretroviral therapy on protease and reverse transcriptase genotype: results of a global collaboration. *PLoS Med*, 2(4):e112.

Karlsson Hedestam, G. B., Fouchier, R. A. M., Phogat, S., Burton, D. R., Sodroski, J., and Wyatt, R. T. (2008). The challenges of eliciting neutralizing antibodies to HIV-1 and to influenza virus. *Nat Rev Microbiol*, 6(2):143–55.

Katzenstein, D. A., Bosch, R. J., Hellmann, N., Wang, N., Bacheler, L., Albrecht, M. A., and Team, A. . S. (2003). Phenotypic susceptibility and virological outcome in nucleoside-experienced patients receiving three or four antiretroviral drugs. *AIDS*, 17(6):821–30.

Kauder, S. E., Bosque, A., Lindqvist, A., Planelles, V., and Verdin, E. (2009). Epigenetic regulation of HIV-1 latency by cytosine methylation. *PLoS Pathog*, 5(6):e1000495.

Kierczak, M., Ginalski, K., Draminski, M., Koronacki, J., Rudnicki, W., and Komorowski, J. (2009). A rough set-based model of HIV-1 reverse transcriptase resistome. *Bioinformatics and Biology Insights*.

Kirkpatrick, S., Gelatt, C., and Vecchi, M. (1983). Optimization by simulated annealing. *Science*, 220(4598):671–680.

Kittler, J., Hatef, M., Duin, R., and Matas, J. (1998). On combining classifiers. *Ieee T Pattern Anal*, 20(3):226–239.

Kjaer, J., Høj, L., Fox, Z., and Lundgren, J. D. (2008). Prediction of phenotypic susceptibility to antiretroviral drugs using physiochemical properties of the primary enzymatic structure combined with artificial neural networks. *HIV Med*, 9(8):642–52.

Korber, B., Muldoon, M., Theiler, J., Gao, F., Gupta, R., Lapedes, A., Hahn, B. H., Wolinsky, S., and Bhattacharya, T. (2000). Timing the ancestor of the HIV-1 pandemic strains. *Science*, 288(5472):1789–96.

Kumar, G. N., Rodrigues, A. D., Buko, A. M., and Denissen, J. F. (1996). Cytochrome p450-mediated metabolism of the HIV-1 protease inhibitor ritonavir (ABT-538) in human liver microsomes. *J Pharmacol Exp Ther*, 277(1):423–31.

Kuncheva, L. (2002). Switching between selection and fusion in combining classifiers: anexperiment. *IEEE Transactions on Systems Man And Cybernetics, Part B-cybernetics*, 32(2):146–156.

Kuncheva, L., Bezdek, J., and Duin, R. (2001). Decision templates for multiple classifier fusion: an experimental comparison. *Pattern Recognition*, 34(2):299–314.

Kyte, J. and Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J Mol Biol*, 157(1):105–32.

Lambert, P.-H., Liu, M., and Siegrist, C.-A. (2005). Can successful vaccines teach us how to induce efficient protective immune responses? *Nature Medicine*, 11(4 Suppl):S54–62.

Landwehr, N., Hall, M., and Frank, E. (2005). Logistic model trees. *Machine learning*, 59:161–205.

Lapins, M., Eklund, M., Spjuth, O., Prusis, P., and Wikberg, J. E. S. (2008). Proteochemometric modeling of HIV protease susceptibility. *BMC Bioinformatics*, 9:181.

Larder, B., Degruttola, V., Hammer, S., Harrigan, R., Wegner, S., Winslow, D., and Zazzi, M. (2002). The international HIV resistance response database initiative: a new global collaborative approach to relating viral genotype and treatment to clinical outcome. *Antivir Ther (Lond)*, 7:S111–S111.

Larder, B., Wang, D., Revell, A., and Lane, C. (2003). Neural network model identified potentially effective drug combinations for patients failing salvage therapy. *2nd IAS Conference on HIV Pathogenesis and Treatment 13–17 July2003, Paris, France. Poster LB39.*

Larder, B., Wang, D., Revell, A., Montaner, J., Harrigan, R., Wolf, F. D., Lange, J., Wegner, S., Ruiz, L., Pérez-Elías, M. J., Emery, S., Gatell, J., Monforte, A. D., Torti, C., Zazzi, M., and Lane, C. (2007). The development of artificial neural networks to predict virological response to combination HIV therapy. *Antivir Ther (Lond)*, 12(1):15–24.

Larder, B. A., Darby, G., and Richman, D. D. (1989). HIV with reduced sensitivity to zidovudine (AZT) isolated during prolonged therapy. *Science*, 243(4899):1731–4.

Larder, B. A. and Kemp, S. D. (1989). Multiple mutations in HIV-1 reverse transcriptase confer high-level resistance to zidovudine (AZT). *Science*, 246(4934):1155–8.

Lathrop, R. and Pazzani, M. (1999). Combinatorial optimization in rapidly mutating drug-resistant viruses. *Journal of Combinatorial Optimization*, 3:301–320.

Le, T., Chiarella, J., Simen, B. B., Hanczaruk, B., Egholm, M., Landry, M. L., Dieckhaus, K., Rosen, M. I., and Kozal, M. J. (2009). Low-abundance HIV drug-resistant viral variants in treatment-experienced persons correlate with historical antiretroviral use. *PLoS ONE*, 4(6):e6079.

Le Cessie, S. and Van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Applied Statistics*, 41:191–201.

Leitner, T., Korber, B., Daniels, M., Calef, C., and Foley, B. (2005). HIV-1 subtype and circulating recombinant form (CRF) reference sequences, 2005. *HIV sequence compendium.*

Lengauer, T., Sander, O., Sierra, S., Thielen, A., and Kaiser, R. (2007). Bioinformatics prediction of HIV coreceptor usage. *Nat Biotechnol*, 25(12):1407–10.

Lengauer, T. and Sing, T. (2006). Bioinformatics-assisted anti-HIV therapy. *Nat Rev Microbiol*, 4(10):790–7.

Li, M., Gao, F., Mascola, J. R., Stamatatos, L., Polonis, V. R., Koutsoukos, M., Voss, G., Goepfert, P., Gilbert, P., Greene, K. M., Bilska, M., Kothe, D. L., Salazar-Gonzalez, J. F., Wei, X., Decker, J. M., Hahn, B. H., and Montefiori, D. C. (2005). Human immunodeficiency virus type 1 env clones from acute and early subtype B infections for standardized assessments of vaccine-elicited neutralizing antibodies. *J Virol*, 79(16):10108–25.

Li, M., Salazar-Gonzalez, J. F., Derdeyn, C. A., Morris, L., Williamson, C., Robinson, J. E., Decker, J. M., Li, Y., Salazar, M. G., Polonis, V. R., Mlisana, K., Karim, S. A., Hong, K., Greene, K. M., Bilska, M., Zhou, J., Allen, S., Chomba, E., Mulenga, J., Vwalika, C., Gao, F., Zhang, M., Korber, B. T. M., Hunter, E., Hahn, B. H., and Montefiori, D. C. (2006). Genetic and neutralization properties of subtype C human immunodeficiency virus type 1 molecular env clones from acute and early heterosexually acquired infections in Southern Africa. *J Virol*, 80(23):11776–90.

Liu, J., Bartesaghi, A., Borgnia, M. J., Sapiro, G., and Subramaniam, S. (2008). Molecular architecture of native HIV-1 gp120 trimers. *Nature*, 455(7209):109–13.

Liu, R. and Yuan, B. (2001). Multiple classifiers combination by clustering and selection. *Information Fusion*, 2:163–168.

Loizidou, E. Z., Kousiappa, I., Zeinalipour-Yazdi, C. D., Van de Vijver, D. A. M. C., and Kostrikis, L. G. (2009). Implications of HIV-1 M group polymorphisms on integrase inhibitor efficacy and resistance: genetic and structural in silico analyses. *Biochemistry*, 48(1):4–6.

Lombardy, S., Regis-Gianas, Y., and Sakarovitch, J. (2004). Introducing Vaucanson. *Theoretical Computer Science*, 328:77–96.

Lyles, R. H., Muñoz, A., Yamashita, T. E., Bazmi, H., Detels, R., Rinaldo, C. R., Margolick, J. B., Phair, J. P., and Mellors, J. W. (2000). Natural history of human immunodeficiency virus type 1 viremia after seroconversion and proximal to AIDS in a large cohort of homosexual men. *J Infect Dis*, 181(3):872–80.

Maggiolo, F., Airoldi, M., Callegaro, A., Ripamonti, D., Gregis, G., Quinzan, G., Bombana, E., Ravasio, V., and Suter, F. (2007). Prediction of virologic outcome of salvage antiretroviral treatment by different systems for interpreting genotypic HIV drug resistance. *Journal of the International Association of Physicians in AIDS Care (JIAPAC)*, 6:87–93.

Mahungu, T., Smith, C., Turner, F., Egan, D., Youle, M., Johnson, M., Khoo, S., Back, D., and Owen, A. (2009). Cytochrome p450 2b6 516g–>t is associated with plasma concentrations of nevirapine at both 200 mg twice daily and 400 mg once daily in an ethnically diverse population. *HIV Med*, 10(5):310–7.

Mallal, S., Phillips, E., Carosi, G., Molina, J.-M., Workman, C., Tomazic, J., Jägel-Guedes, E., Rugina, S., Kozyrev, O., Cid, J. F., Hay, P., Nolan, D., Hughes, S., Hughes, A., Ryan, S., Fitch, N., Thorborn, D., and Benbow, A. (2008). HLA-B*5701 screening for hypersensitivity to abacavir. *N Engl J Med*, 358(6):568–79.

Mansky, L. M. and Temin, H. M. (1995). 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):5087–94.

Markowitz, M., Mohri, H., Mehandru, S., Shet, A., Berry, L., Kalyanaraman, R., Kim, A., Chung, C., Jean-Pierre, P., Horowitz, A., Mar, M. L., Wrin, T., Parkin, N., Poles,

M., Petropoulos, C., Mullen, M., Boden, D., and Ho, D. D. (2005). Infection with multidrug resistant, dual-tropic HIV-1 and rapid progression to AIDS: a case report. *Lancet*, 365(9464):1031–8.

Mattapallil, J. J., Douek, D. C., Hill, B., Nishimura, Y., Martin, M., and Roederer, M. (2005). Massive infection and loss of memory CD4+ T cells in multiple tissues during acute SIV infection. *Nature*, 434(7037):1093–7.

McCallister, S., Lalezari, J., Richmond, G., Thompson, M., Harrigan, R., Martin, D., Salzwedel, K., and Allaway, G. (2008). HIV-1 Gag polymorphisms determine treatment response to bevirimat (PA-457). *Antivir Ther (Lond)*, 13(4):A10–A10.

McColl, D. J., Fransen, S., Gupta, S., Parkin, N., Margot, N., Chuck, S., Cheng, A. K., and Miller, M. D. (2007). Resistance and cross-resistance to first generation integrase inhibitors: insights from a Phase II study of elvitegravir (GS-9137). *Antivir Ther (Lond)*, 12(5):S11–S11.

Meynard, J.-L., Vray, M., Morand-Joubert, L., Race, E., Descamps, D., Peytavin, G., Matheron, S., Lamotte, C., Guiramand, S., Costagliola, D., Brun-Vézinet, F., Clavel, F., Girard, P.-M., and Group, N. T. (2002). Phenotypic or genotypic resistance testing for choosing antiretroviral therapy after treatment failure: a randomized trial. *AIDS*, 16(5):727–36.

Miller, M. D. and Hazuda, D. J. (2004). HIV resistance to the fusion inhibitor enfuvirtide: mechanisms and clinical implications. *Drug Resist Updat*, 7(2):89–95.

Mocroft, A., Vella, S., Benfield, T. L., Chiesi, A., Miller, V., Gargalianos, P., d'Arminio Monforte, A., Yust, I., Bruun, J. N., Phillips, A. N., and Lundgren, J. D. (1998). Changing patterns of mortality across europe in patients infected with hiv-1. eurosida study group. *Lancet*, 352(9142):1725–30.

Mohri, M. (2002). Semiring frameworks and algorithms for shortest-distance problems. *Journal of Automata, Languages and Combinatorics*, 7(3):321–350.

Mohri, M. (2005). Statistical natural language processing. *M. Lothaire, editor, Applied Combinatorics on Words. Cambridge University Press.*, pages 199–226.

Mohri, M., Pereira, F., and Riley, M. (2000). The design principles of a weighted finite-state transducer library. *Theoretical Computer Science*, 231:17–32.

Mohri, M. and Riley, M. (2001). A weight pushing algorithm for large vocabulary speech recognition. *Proceedings of the Seventh European Conference on Speech Communication and Technology*, pages 1303–1606.

Mohri, M. and Riley, M. (2002). An efficient algorithm for the N-best-strings problem. *Proceedings of the Seventh International Conference on Spoken Language Processing*, pages 1313–1316.

Montagnier, L. (2002). Historical essay. a history of HIV discovery. *Science*, 298(5599):1727–8.

Müller, F. (2008). Inferring virological response to antiretroviral combination therapy based on past treatment lines. *Bachelor Thesis, Saarland University, Saarbrücken, Germany.*

Müller, F. (2009). Assessing antibody neutralization oh HIV-1 as an initial step in the search of gp160-based immunogens. *Master Thesis, Saarland University, Saarbrücken, Germany.*

Ney, H., Essen, U., and Kneser, R. (1994). On structuring probabilistic dependences in stochastic language modelling. *Computer Speech and Language*, 8:1–38.

Nijhuis, M., van Maarseveen, N. M., Lastere, S., Schipper, P., Coakley, E., Glass, B., Rovenska, M., de Jong, D., Chappey, C., Goedegebuure, I. W., Heilek-Snyder, G., Dulude, D., Cammack, N., Brakier-Gingras, L., Konvalinka, J., Parkin, N., Kräusslich, H.-G., Brun-Vezinet, F., and Boucher, C. A. B. (2007). A novel substrate-based HIV-1 protease inhibitor drug resistance mechanism. *PLoS Med*, 4(1):e36.

Novembre, J., Galvani, A. P., and Slatkin, M. (2005). The geographic spread of the CCR5 δ32 HIV-resistance allele. *PLoS Biol*, 3(11):e339.

Nowozin, S., BakIr, G., and Tsuda, K. (2007). Discriminative subsequence mining for action classification. *11th IEEE International Conference on Computer Vision*, pages 1919–1923.

Oette, M., Kaiser, R., Däumer, M., Petch, R., Fätkenheuer, G., Carls, H., Rockstroh, J. K., Schmalöer, D., Stechel, J., Feldt, T., Pfister, H., and Häussinger, D. (2006). Primary HIV drug resistance and efficacy of first-line antiretroviral therapy guided by resistance testing. *J Acquir Immune Defic Syndr*, 41(5):573–81.

Pantaleo, G., Graziosi, C., and Fauci, A. S. (1993). New concepts in the immunopathogenesis of human immunodeficiency virus infection. *N Engl J Med*, 328(5):327–35.

Pantophlet, R. and Burton, D. R. (2006). GP120: target for neutralizing HIV-1 antibodies. *Annu Rev Immunol*, 24:739–69.

Parikh, U. M., Bacheler, L., Koontz, D., and Mellors, J. W. (2006). The K65R mutation in human immunodeficiency virus type 1 reverse transcriptase exhibits bidirectional phenotypic antagonism with thymidine analog mutations. *J Virol*, 80(10):4971–7.

Pereira, F. and Riley, M. (1997). Speech recognition by composition of weighted finite automata. *Roche, E. and Schabes, Y. (Eds.) Finite-State language processing. MIT Press*, pages 431–453.

Perelson, A. S. (2002). Modelling viral and immune system dynamics. *Nat Rev Immunol*, 2(1):28–36.

Perner, J., Altmann, A., and Lengauer, T. (2009). Semi-supervised learning for improving prediction of HIV drug resistance. *Lecture Notes in Informatics, Proceedings of the German Conference on Bioinformatics 2009*, P-157:55–65.

Petropoulos, C. J., Parkin, N. T., Limoli, K. L., Lie, Y. S., Wrin, T., Huang, W., Tian, H., Smith, D., Winslow, G. A., Capon, D. J., and Whitcomb, J. M. (2000). A novel phenotypic drug susceptibility assay for human immunodeficiency virus type 1. *Antimicrob Agents Chemother*, 44(4):920–8.

Pettit, S. C., Everitt, L. E., Choudhury, S., Dunn, B. M., and Kaplan, A. H. (2004). Initial cleavage of the human immunodeficiency virus type 1 gagpol precursor by its activated protease occurs by an intramolecular mechanism. *J Virol*, 78(16):8477–85.

Phogat, S. and Wyatt, R. (2007). Rational modifications of HIV-1 envelope glycoproteins for immunogen design. *Curr Pharm Des*, 13(2):213–27.

Picker, L. J. (2006). Immunopathogenesis of acute AIDS virus infection. *Curr Opin Immunol*, 18(4):399–405.

Plantier, J.-C., Leoz, M., Dickerson, J. E., Oliveira, F. D., Cordonnier, F., Lemée, V., Damond, F., and Simon, D. L. R. F. (2009). A new human immunodeficiency virus derived from gorillas. *Nature Medicine*.

Plotkin, S. A. (2005). Vaccines: past, present and future. *Nature Medicine*, 11(4 Suppl):S5–11.

Popovic, M., Sarngadharan, M. G., Read, E., and Gallo, R. C. (1984). Detection, isolation, and continuous production of cytopathic retroviruses (HTLV-III) from patients with AIDS and pre-AIDS. *Science*, 224(4648):497–500.

Prilusky, J., Felder, C. E., Zeev-Ben-Mordehai, T., Rydberg, E. H., Man, O., Beckmann, J. S., Silman, I., and Sussman, J. L. (2005). Foldindex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics*, 21(16):3435–8.

Prosperi, M., Giambenedetto, S. D., Trotta, M., Cingolani, A., Ruiz, L., Baxter, J., Clevenbergh, P., Perno, C., Cauda, R., Ulivi, G., Antinori, A., and De Luca, A. (2004). A fuzzy relational system trained by genetic algorithms and HIV-1 resistance genotypes/virological response data from prospective studies usefully predicts treatment outcomes. *Antivir Ther (Lond)*, 9(4):U89–U89.

Prosperi, M., Rosen-Zvi, M., Altmann, A., Aharoni, E., Peres, Y., Sönnerbord, A., Incardona, F., Struck, D., Kaiser, R., and Zazzi, M. (2009a). Antiretroviral therapy optimisation without genotype resistance testing: a perspective on treatment history based models. *Reviews in Antiviral Therapy*, 1:28–29.

Prosperi, M., Zazzi, M., Perno, C., Giambenedetto, S. D., Baxter, J., Ruiz, L., Clevenbergh, P., Ulivi, G., Antinori, A., and De Luca, A. (2005). 'Common law' applied to treatment decisions for drug resistant HIV. *Antivir Ther (Lond)*, 10(4):S62–S62.

Prosperi, M. C. F., D'Autilia, R., Incardona, F., De Luca, A., Zazzi, M., and Ulivi, G. (2009b). Stochastic modelling of genotypic drug-resistance for human immunodeficiency virus towards long-term combination therapy optimization. *Bioinformatics*, 25(8):1040–7.

Rahnenführer, J., Beerenwinkel, N., Schulz, W. A., Hartmann, C., von Deimling, A., Wullich, B., and Lengauer, T. (2005). Estimating cancer survival and clinical outcome based on genetic tumor progression scores. *Bioinformatics*, 21(10):2438–46.

Rambaut, A., Posada, D., Crandall, K. A., and Holmes, E. C. (2004). The causes and consequences of HIV evolution. *Nat Rev Genet*, 5(1):52–61.

Ratner, L., Haseltine, W., Patarca, R., Livak, K. J., Starcich, B., Josephs, S. F., Doran, E. R., Rafalski, J. A., Whitehorn, E. A., Baumeister, K., Ivanoff, L., Petteway, S. R., Pearson, M. L., Lautenberger, J. A., Papas, T. S., Ghrayeb, J., Chang, N. T., Gallo, R. C., and Wong-Staal, F. (1985). Complete nucleotide sequence of the AIDS virus, HTLV-III. *Nature*, 313(6000):277–84.

Rerks-Ngarm, S., Pitisuttithum, P., Nitayaphan, S., Kaewkungwal, J., Chiu, J., Paris, R., Premsri, N., Namwat, C., de Souza, M., Adams, E., Benenson, M., Gurunathan, S., Tartaglia, J., McNeil, J., Francis, D., Stablein, D., Birx, D., Chunsuttiwat, S., Khamboonruang, C., Thongcharoen, P., Robb, M., Michael, N., Kunasol, P., Kim, J., and the MOPH-TAVEG Investigators (2009). Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in thailand. *N Engl J Med*.

Revell, A. D., Wang, D., Harrigan, R., Gatell, J., Ruiz, L., Emery, S., Perez-Elias, M. J., Torti, C., Baxter, J., DeWolf, F., Gazzard, B., Geretti, A. M., Staszewski, S., Hamers, R., Wensing, A. M. J., Lange, J., Montaner, J. M., and Larder, B. A. (2009). Computational models developed without a genotype for resource-poor countries predict response to HIV treatment with 82% accuracy. *Antivir Ther (Lond)*, 14(4):A38–A38.

Rhee, S.-Y., Gonzales, M. J., Kantor, R., Betts, B. J., Ravela, J., and Shafer, R. W. (2003). Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res*, 31(1):298–303.

Rhee, S.-Y., Taylor, J., Wadhera, G., Ben-Hur, A., Brutlag, D. L., and Shafer, R. W. (2006). Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proc Natl Acad Sci USA*, 103(46):17355–60.

Ribeiro, R. M. and Bonhoeffer, S. (2000). Production of resistant HIV mutants during antiretroviral therapy. *Proc Natl Acad Sci USA*, 97(14):7681–6.

Richman, D. D., Wrin, T., Little, S. J., and Petropoulos, C. J. (2003). Rapid evolution of the neutralizing antibody response to HIV type 1 infection. *Proc Natl Acad Sci USA*, 100(7):4144–9.

Roben, P., Moore, J. P., Thali, M., Sodroski, J., Barbas, C. F., and Burton, D. R. (1994). Recognition properties of a panel of human recombinant Fab fragments to the CD4 binding site of gp120 that show differing abilities to neutralize human immunodeficiency virus type 1. *J Virol*, 68(8):4821–8.

Robertson, D. L., Anderson, J. P., Bradac, J. A., Carr, J. K., Foley, B., Funkhouser, R. K., Gao, F., Hahn, B. H., Kalish, M. L., Kuiken, C., Learn, G. H., Leitner, T., McCutchan, F., Osmanov, S., Peeters, M., Pieniazek, D., Salminen, M., Sharp, P. M., Wolinsky, S., and Korber, B. (2000). HIV-1 nomenclature proposal. *Science*, 288(5463):55–6.

Rogova, G. (1994). Combining the results of several neural network classifiers. *Neural networks*, 7(5):777–781.

Roomp, K., Beerenwinkel, N., Sing, T., Schuelter, E., Buech, J., Sierra-Aragon, S., Daeumer, M., Hoffmann, D., Kaiser, R., Lengauer, T., and Selbig, J. (2006). Arevir: A secure platform for designing personalized antiretroviral therapies against HIV. *Lect Notes Comput Sc*, 4075:185–194.

Rosen-Zvi, M., Altmann, A., Prosperi, M., Aharoni, E., Neuvirth, H., Sönnerborg, A., Schülter, E., Struck, D., Peres, Y., Incardona, F., Kaiser, R., Zazzi, M., and Lengauer, T. (2008). Selecting anti-HIV therapies based on a variety of genomic and clinical factors. *Bioinformatics*, 24(13):i399–406.

Rousseeuw, P. and Kaufman, L. (1990). Finding groups in data: an introduction to cluster analysis. *Wiley*.

Sacktor, N., Nakasujja, N., Skolasky, R. L., Rezapour, M., Robertson, K., Musisi, S., Katabira, E., Ronald, A., Clifford, D. B., Laeyendecker, O., and Quinn, T. C. (2009). HIV subtype D is associated with dementia, compared with subtype A, in immunosuppressed individuals at risk of cognitive impairment in kampala, uganda. *Clin Infect Dis*, 49(5):780–6.

Saigo, H., Uno, T., and Tsuda, K. (2007). Mining complex genotypic features for predicting HIV-1 drug resistance. *Bioinformatics*, 23(18):2455–62.

Salzwedel, K., Martin, D. E., and Sakalian, M. (2007). Maturation inhibitors: a new therapeutic class targets the virus structure. *AIDS reviews*, 9(3):162–72.

Sander, O., Sing, T., Sommer, I., Low, A. J., Cheung, P. K., Harrigan, P. R., Lengauer, T., and Domingues, F. S. (2007). Structural descriptors of gp120 V3 loop for the prediction of HIV-1 coreceptor usage. *PLoS Comput Biol*, 3(3):e58.

Sarkar, I., Hauber, I., Hauber, J., and Buchholz, F. (2007). HIV-1 proviral DNA excision using an evolved recombinase. *Science*, 316(5833):1912–5.

Schweighardt, B., Liu, Y., Huang, W., Chappey, C., Lie, Y. S., Petropoulos, C. J., and Wrin, T. (2007). Development of an HIV-1 reference panel of subtype B envelope clones isolated from the plasma of recently infected individuals. *J Acquir Immune Defic Syndr*, 46(1):1–11.

Shaw, G. M., Hahn, B. H., Arya, S. K., Groopman, J. E., Gallo, R. C., and Wong-Staal, F. (1984). Molecular characterization of human T-cell leukemia (lymphotropic) virus type III in the acquired immune deficiency syndrome. *Science*, 226(4679):1165–71.

Shaw, G. M., Harper, M. E., Hahn, B. H., Epstein, L. G., Gajdusek, D. C., Price, R. W., Navia, B. A., Petito, C. K., O'Hara, C. J., and Groopman, J. E. (1985). HTLV-III infection in brains of children and adults with AIDS encephalopathy. *Science*, 227(4683):177–82.

Sheehy, A. M., Gaddis, N. C., Choi, J. D., and Malim, M. H. (2002). Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. *Nature*, 418(6898):646–50.

Shehu-Xhilaga, M., Crowe, S. M., and Mak, J. (2001). Maintenance of the Gag/Gag-Pol ratio is important for human immunodeficiency virus type 1 RNA dimerization and viral infectivity. *J Virol*, 75(4):1834–41.

Shen, L., Peterson, S., Sedaghat, A. R., McMahon, M. A., Callender, M., Zhang, H., Zhou, Y., Pitt, E., Anderson, K. S., Acosta, E. P., and Siliciano, R. F. (2008). Dose-response curve slope sets class-specific limits on inhibitory potential of anti-HIV drugs. *Nature Medicine*, 14(7):762–6.

Simon, V., Ho, D. D., and Karim, Q. A. (2006). HIV/AIDS epidemiology, pathogenesis, prevention, and treatment. *Lancet*, 368(9534):489–504.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005a). ROCR: visualizing classifier performance in R. *Bioinformatics*, 21(20):3940–1.

Sing, T., Svicher, V., Beerenwinkel, N., Ceccherinig-Silberstein, F., Däumer, M., Kaiser, R., Walter, H., Korn, K., Hoffmann, D., Oette, M., Rockstroh, J., Fatkenheuer, G., Perno, C., and Lengauer, T. (2005b). Characterization of novel HIV drug resistance mutations using clustering, multidimensional scaling and SVM-based feature ranking. *Lect Notes Artif Int*, 3721:285–296.

Sluis-Cremer, N., Arion, D., and Parniak, M. A. (2000). Molecular mechanisms of HIV-1 resistance to nucleoside reverse transcriptase inhibitors (NRTIs). *Cell Mol Life Sci*, 57(10):1408–22.

Snoeck, J., Kantor, R., Shafer, R. W., Laethem, K. V., Deforche, K., Carvalho, A. P., Wynhoven, B., Soares, M. A., Cane, P., Clarke, J., Pillay, C., Sirivichayakul, S., Ariyoshi, K., Holguin, A., Rudich, H., Rodrigues, R., Bouzas, M. B., Brun-Vézinet, F., Reid, C., Cahn, P., Brigido, L. F., Grossman, Z., Soriano, V., Sugiura, W., Phanuphak, P., Morris, L., Weber, J., Pillay, D., Tanuri, A., Harrigan, R. P., Camacho, R., Schapiro, J. M., Katzenstein, D., and Vandamme, A.-M. (2006). Discordances between interpretation algorithms for genotypic resistance to protease and reverse transcriptase inhibitors of human immunodeficiency virus are subtype dependent. *Antimicrob Agents Chemother*, 50(2):694–701.

Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics*, 8:25.

Swanstrom, R., Bosch, R. J., Katzenstein, D., Cheng, H., Jiang, H., Hellmann, N., Haubrich, R., Fiscus, S. A., Fletcher, C. V., Acosta, E. P., and Gulick, R. M. (2004). Weighted phenotypic susceptibility scores are predictive of the HIV-1 RNA response in protease inhibitor-experienced HIV-1-infected subjects. *J Infect Dis*, 190(5):886–93.

Telenti, A. and Goldstein, D. B. (2006). Genomics meets HIV-1. *Nat Rev Microbiol*, 4(11):865–73.

Telenti, A. and Zanger, U. M. (2008). Pharmacogenetics of anti-HIV drugs. *Annu Rev Pharmacol Toxicol*, 48:227–56.

Thibaut, L., Rochas, S., Dourlat, J., Monneret, C., Carayon, K., Deprez, E., Mouscadet, J. F., Soma, E., and Lebel-Binay, S. (2009). New integrase binding inhibitors acting in synergy with raltegravir. *Antivir Ther (Lond)*, 14(4):A29–A29.

Trkola, A., Kuster, H., Rusert, P., von Wyl, V., Leemann, C., Weber, R., Stiegler, G., Katinger, H., Joos, B., and Günthard, H. F. (2008). In vivo efficacy of human immunodeficiency virus neutralizing antibodies: estimates for protective titers. *J Virol*, 82(3):1591–9.

Uchil, P. D. and Mothes, W. (2009). HIV entry revisited. *Cell*, 137(3):402–4.

Van Laethem, K., De Luca, A., Antinori, A., Cingolani, A., Perno, C. F., and Vandamme, A.-M. (2002). A genotypic drug resistance interpretation algorithm that significantly predicts therapy response in HIV-1-infected patients. *Antivir Ther (Lond)*, 7(2):123–9.

Vercauteren, J. and Vandamme, A.-M. (2006). Algorithms for the interpretation of HIV-1 genotypic drug resistance information. *Antiviral Res*, 71(2-3):335–42.

Verheyen, J., Verhofstede, C., Knops, E., Vandekerckhove, L., Fun, A., Brunen, D., Dauwe, K., Wensing, A., Pfister, H., Kaiser, R., and Nijhuis, M. (2009). High prevalence of bevirimat resistance mutations in protease inhibitor-resistant hiv isolates. *AIDS*.

Vermeiren, H., Van Craenenbroeck, E., Alen, P., Bacheler, L., Picchio, G., Lecocq, P., and Team, V. C. R. C. (2007). Prediction of HIV-1 drug susceptibility phenotype from the viral genotype using linear regression modeling. *J Virol Methods*, 145(1):47–55.

von Andrian, U. H. and Mackay, C. R. (2000). T-cell function and migration. two sides of the same coin. *N Engl J Med*, 343(14):1020–34.

Walker, B. D. and Burton, D. R. (2008). Toward an AIDS vaccine. *Science*, 320(5877):760–4.

Walker, L. M., Phogat, S. K., Chan-Hui, P.-Y., Wagner, D., Phung, P., Goss, J. L., Wrin, T., Simek, M. D., Fling, S., Mitcham, J. L., Lehrman, J. K., Priddy, F. H., Olsen, O. A., Frey, S. M., Hammond, P. W., Investigators, P. G. P., Kaminsky, S., Zamb, T., Moyle, M., Koff, W. C., Poignard, P., and Burton, D. R. (2009). Broad and potent neutralizing antibodies from an african donor reveal a new HIV-1 vaccine target. *Science*, 326(5950):285–9.

Walter, H., Schmidt, B., Korn, K., Vandamme, A. M., Harrer, T., and Überla, K. (1999). Rapid, phenotypic HIV-1 drug sensitivity assay for protease and reverse transcriptase inhibitors. *J Clin Virol*, 13(1-2):71–80.

Wang, D., Larder, B., Revell, A., Harrigan, R., and Montaner, J. (2003). A neural network model using clinical cohort data accurately predicts virological response and identifies regimens with increased probabilisticbility of success in treatment failures. *Antivir Ther (Lond)*, 8(3):U99–U99.

Wang, J. Y., Ling, H., Yang, W., and Craigie, R. (2001). Structure of a two-domain fragment of HIV-1 integrase: implications for domain organization in the intact protein. *EMBO J*, 20(24):7333–43.

Wang, K., Jenwitheesuk, E., Samudrala, R., and Mittler, J. E. (2004). Simple linear model provides highly accurate genotypic predictions of HIV-1 drug resistance. *Antivir Ther (Lond)*, 9(3):343–52.

Winters, B., Montaner, J., Harrigan, P. R., Gazzard, B., Pozniak, A., Miller, M. D., Emery, S., van Leth, F., Robinson, P., Baxter, J. D., Perez-Elias, M., Castor, D., Hammer, S., Rinehart, A., Vermeiren, H., Craenenbroeck, E. V., and Bacheler, L. (2008). Determination of clinically relevant cutoffs for HIV-1 phenotypic resistance estimates through a combined analysis of clinical trial and cohort data. *J Acquir Immune Defic Syndr*, 48(1):26–34.

Wong-Staal, F., Shaw, G. M., Hahn, B. H., Salahuddin, S. Z., Popovic, M., Markham, P., Redfield, R., and Gallo, R. C. (1985). Genomic diversity of human T-lymphotropic virus type III (HTLV-III). *Science*, 229(4715):759–62.

Woods, K., Kegelmeyer Jr., W. P., and Bowyer, K. (1997). Combination of multiple classifiers using local accuracy estimates. *Ieee T Pattern Anal*, 19:405–410.

Wright, S. (1931). Evolution in mendelian populations. *Genetics*, 16:97–159.

Wyatt, R. and Sodroski, J. (1998). The HIV-1 envelope glycoproteins: fusogens, antigens, and immunogens. *Science*, 280(5371):1884–8.

Yang, O. O., Nguyen, P. T., Kalams, S. A., Dorfman, T., Göttlinger, H. G., Stewart, S., Chen, I. S. Y., Threlkeld, S., and Walker, B. D. (2002). Nef-mediated resistance of human immunodeficiency virus type 1 to antiviral cytotoxic T lymphocytes. *J Virol*, 76(4):1626–31.

Yin, J., Beerenwinkel, N., Rahnenführer, J., and Lengauer, T. (2006). Model selection for mixtures of mutagenetic trees. *Statistical applications in genetics and molecular biology*, 5:Article17.

Yoder, K. E. and Bushman, F. D. (2000). Repair of gaps in retroviral DNA integration intermediates. *J Virol*, 74(23):11191–200.

Zaccarelli, M., Lorenzini, P., Ceccherini-Silberstein, F., Tozzi, V., Forbici, F., Gori, C., Trotta, M. P., Boumis, E., Narciso, P., Perno, C. F., and Antinori, A. (2009). Historical resistance profile helps to predict salvage failure. *Antivir Ther (Lond)*, 14(2):285–91.

Zazzi, M., Prosperi, M., Vicenti, I., Giambenedetto, S. D., Callegaro, A., Bruzzone, B., Baldanti, F., Gonnelli, A., Boeri, E., Paolini, E., Rusconi, S., Giacometti, A., Maggiolo, F., Menzo, S., De Luca, A., and Group, A. C. (2009). Rules-based HIV-1 genotypic resistance interpretation systems predict 8 week and 24 week virological antiretroviral treatment outcome and benefit from drug potency weighting. *J Antimicrob Chemother*, 64(3):616–24.

Zhou, T., Xu, L., Dey, B., Hessell, A. J., Ryk, D. V., Xiang, S.-H., Yang, X., Zhang, M.-Y., Zwick, M. B., Arthos, J., Burton, D. R., Dimitrov, D. S., Sodroski, J., Wyatt, R., Nabel, G. J., and Kwong, P. D. (2007). Structural definition of a conserved neutralization epitope on HIV-1 gp120. *Nature*, 445(7129):732–7.

Zhu, T., Korber, B. T., Nahmias, A. J., Hooper, E., Sharp, P. M., and Ho, D. D. (1998). An african HIV-1 sequence from 1959 and implications for the origin of the epidemic. *Nature*, 391(6667):594–7.

Zhu, X. (2007). Semi-supervised learning literature survey. *Computer Science*.

Zwick, M. B., Labrijn, A. F., Wang, M., Spenlehauer, C., Saphire, E. O., Binley, J. M., Moore, J. P., Stiegler, G., Katinger, H., Burton, D. R., and Parren, P. W. (2001). Broadly neutralizing antibodies targeted to the membrane-proximal external region of human immunodeficiency virus type 1 glycoprotein gp41. *J Virol*, 75(22):10892–905.

Appendix: List of Publications

Jasmina Bogojeska, Steffen Bickel, **André Altmann**, Thomas Lengauer (submitted). Dealing with sparse data in predicting outcomes of HIV combination therapies. *Bioinformatics (submitted).*

André Altmann*, Laura Tolosi*, Oliver Sander, Thomas Lengauer (2010). Permuation importance: a corrected feature importance measure. *Bioinformatics, 2010, 26(10): 1340-1347.*

Hendrik Weisser, **André Altmann**, Saleta Sierra, Francesca Incardona, Daniel Struck, Anders Sönnerborg, Rolf Kaiser, Maurizio Zazzi, Monika Tschochner, Hauke Walter, and Thomas Lengauer (2010). Only slight impact of predicted replicative capacity for therapy response prediction. *PLoS ONE, 2010, 5 (2): e9044.*

Thomas Lengauer, **André Altmann**, Alexander Thielen, Rolf Kaiser (2010). Chasing the AIDS virus. *Communications of the ACM, 2010, 53(3): 66-74.*

Thomas Lengauer, **André Altmann**, Alexander Thielen (2009). Bioinformatische Unterstützung der Auswahl von HIV-Therapien. *Informatik-Spektrum, 2009, 32(4): 320-331.*

Juliane Perner , **André Altmann**, Thomas Lengauer (2009). Semi-Supervised Learning for Improving Prediction of HIV Drug Resistance. *Lecture Notes in Informatics (Proceedings of GCB 2009), P-157: 55-65.*

Nadine Sichtig, Saleta Sierra, Rolf Kaiser, Martin Däumer, Stefan Reuter, Eugen Schülter, **André Altmann**, Gerd Fätkenheuer, Ulf Dittmer, Herbert Pfister, Stefan Esser (2009). Evolution of Raltegravir Resistance During Therapy. *Journal of Antimicrobial Chemotherapy, 64: 25-32.*

Mattia CF Prosperi, **André Altmann**, Michal Rosen-Zvi, Ehud Aharoni, Gabor Borgulya, Fulop Bazso, Anders Sönnerborg, Yarenda Peres, Eugen Schülter, Daniel Struck, Giovanni Ulivi, Francesca Incardona, Anne-Mieke Vandamme, Jürgen Vercauteren and Maurizio Zazzi for the EuResist and Virolab study groups (2009). Investigation of Expert Rule Bases, Logistic Regression, and Non-Linear Machine Learning Techniques for Predicting Response to Antiretroviral Treatment. *Antiviral Thearapy, 14: 433-442.*

André Altmann*, Tobias Sing*, Hans Vermeiren, Bart Winters, Elke Van Craenenbroeck, Koen Van der Borght, Soo-Yon Rhee, Robert W Shafer, Eugen Schülter, Rolf Kaiser, Yardena Peres, Anders Sönnerborg, W Jeffrey Fessel, Francesca Incardona, Maurizio Zazzi, Lee Bacheler, Herman Van Vlijmen, Thomas Lengauer (2009). Advantages of Predicted Phenotypes and Statistical Learning Models in Inferring Virological Response to Antiretroviral Therapy from HIV Genotype. *Antiviral Thearapy, 14: 273-283.*

André Altmann, Martin Däumer, Niko Beerenwinkel, Yardena Peres, Eugen Schülter, Joachim Büch, Soo-Yon Rhee, Anders Sönnerborg, Jeffrey Fessel, Robert W Shafer, Maurizio Zazzi, Rolf Kaiser and Thomas Lengauer (2009). Predicting response to combination antiretroviral therapy: retrospective validation of geno2pheno-THEO on a large clinical database. *Journal of Infectious Diseases, 199(7), 999-1006.*

André Altmann, Michal Rosen-Zvi , Mattia Prosperi, Ehud Aharoni, Hani Neuvirth, Anders Sönnerborg, Eugen Schülter, Joachim Büch, Daniel Struck, Yardena Peres, Francesca Incardona, Rolf Kaiser, Maurizio Zazzi and Thomas Lengauer (2008). Comparison of Classifier Fusion Methods for Predicting Response to Anti HIV-1 Therapy. *PLoS ONE, 3(10): e3470.*

Jasmina Bogojeska, Adrian Alexa, **André Altmann**, Thomas Lengauer and Jörg Rahnenführer (2008). Rtreemix: an R package for estimating evolutionary pathways and genetic progression scores. *Bioinformatics, 24 (20): 2391.*

Michal Rosen-Zvi, **André Altmann**, Mattia Prosperi, Ehud Aharoni, Hani Neuvirth, Anders Sönnerborg, Eugen Schülter, Daniel Struck, Yardena Peres, Francesca Incardona, Rolf Kaiser, Maurizio Zazzi and Thomas Lengauer (2008). Selecting anti-HIV therapies based on a variety of genomic and clinical factors. *Bioinformatics, 24 (13):i399-406.*

André Altmann, Niko Beerenwinkel, Tobias Sing, Igor Savenkov, Martin Däumer, Rolf Kaiser, Soo-Yon Rhee, W Jeffrey Fessel, Robert W Shafer and Thomas Lengauer (2007). Improved prediction of response to antiretroviral combination therapy using the genetic barrier to drug resistance. *Antiviral Therapy, 12 (2):169-178.*