Predicting HIV-1 Co-receptor Usage of the Viral Quasispecies Using Classifier Ensembles

Inaugural-Dissertation
zur
Erlangung des Doktorgrades
Dr. rer. nat.

der Fakultät für
Biologie
an der

Universität Duisburg-Essen

vorgelegt von
Jan Nikolaj Dybowski

aus Starnberg

März 2011

Die der vorliegenden Arbeit zugrunde liegenden Experimente wurden am Zentrum für Medizinische Biotechnologie (ZMB) in der Abteilung für Bioinformatik der Universität Duisburg-Essen durchgeführt.

1. Gutachter: Prof. Dr. Daniel Hoffmann
2. Gutachter: Prof. Dr. Jörg Timm

Vorsitzender des Prüfungsausschusses: Prof. Dr. Markus Kaiser

Tag der mündlichen Prüfung: 16. Juni 2011

«Life moves pretty fast. If you don't stop and look around once in a while, you could miss it.»

– Ferris Bueller, Ferris Bueller's Day Off

Having looked around thoroughly, I found my love eight years ago. Now there remains little to be missed, but so much more to be done together. This is for you.

– JND

Contents

List of Abbreviations 6

List of Figures 9

List of Tables 11

1 Introduction to HIV-1 and Co-receptor Tropism 12
1.1 HIV and AIDS 12
1.1.1 History of HIV and AIDS 12
1.1.2 The Global AIDS Epidemic 13
1.2 HIV Structure and Lifecycle 15
1.3 Antiretroviral Therapy 17
1.3.1 Antiretroviral Drugs 19
1.3.2 Resistances and Testing 21
1.4 HIV-1 Cell Entry and Co-receptor Tropism 21
1.4.1 Entry mechanism 21
1.4.2 Structural Data 25
1.4.3 Small-Molecule CCR5 Antagonists 28
1.4.4 Phenotypic Testing of Co-receptor Tropism 32
1.4.5 Genotypic Testing of Co-receptor Tropism 33
1.5 Research Motivation 34

2 Prediction of HIV-1 Co-receptor Tropism from Genotype 35
2.1 Introduction 35
2.1.1 Machine Learning in Biology 36
2.1.2 Random Forests 36
2.1.3 Structural Descriptor: The Electrostatic Potential 41
2.1.4 Sequence Descriptor: The Hydropathy Scale 46
2.2 Materials and Methods 47

2.3 Results 50

2.3.1 Prediction of Tropism Based on the Electrostatic Potential 50

2.3.2 Prediction of Tropism Based on the Hydropathy Scale 60

2.3.3 Fusion of Electrostatics and Hydrophobicity Models 62

2.4 Discussion 70

3 Prediction of HIV-1 Co-receptor Tropism from Quasispecies 77

3.1 Introduction 77

3.1.1 The Quasispecies 77

3.1.2 Next-Generation Sequencing 78

3.1.3 Tropism Prediction and the Viral Quasispecies 81

3.2 Materials and Methods 84

3.3 Results 87

3.3.1 Prediction of Tropism Based on the Viral Quasispecies 87

3.3.2 Structure of Viral Quasispecies as Indicator for Tropism Switches 91

3.4 Discussion 95

4 Summary and Outlook 99

4.1 Summary 99

4.2 Outlook 100

Bibliography 102

List of Publications 121

Acknowledgements 122

Curriculum Vitæ 123

Declarations 124

List of Abbreviations

ϕ	Electrostatic Potential
7TM	Transmembrane 7-Helix Bundle
AIDS	Acquired Immunodeficiency Syndrome
ART	Antiretroviral Therapy
AUC	Area Under the Curve
CCR5	Chemokine Co-Receptor 5
CD4	Cluster of Differentiation 4 Receptor
cDNA	Complementary-Desoxyribonucleic Acid
CXCR4	Chemokine Co-Receptor 4
ddNTPs	Dideoxynucleotidetriphosphates
DM	Dual-tropic/Mixed, Trofile result detecting X4 viruses
DNA	Desoxyribonucleic Acid
dNTPs	Deoxynucleotidetriphosphates
ECL	Extracellular Loop
emPCR	Polymerase Chain Reaction
ESP	Electrostic Potential
ESTA	Enhanced Sensitivity Trofile Assay
FDA	Food and Drug Administration
FPR	False Positive Rate

gp41, gp120	Glycoproteins 41 and 120
GPCR	G-Protein Coupled Receptor
HAART	Highly Active Antiretroviral Therapy
HIV	Human Immunodeficiency Virus
HTLV	T-lymphotropic virus
ICL	Intracellular Loop
II	Integrase Inhibitor
MIP1-α/β	Macrophage Inflammatory Protein-1α/β
mRNA	Messenger-Ribonucleic Acid
MVC	Maraviroc
NGS	Next-Generation Sequencing
NMR	Nuclear Magnetic Resonance
NNRTI	Non-Nucleosidic Reverse Transcriptase Inhibitor
NRTI	Nucleosidic Reverse Transcriptase Inhibitor
OBT	Optimized Background Therapy
OOB	Out-of-Bag
PBE	Poisson-Boltzmann Equation
PCR	Polymerase Chain Reaction
PI	Protease Inhibitor
PSSM	Position-specific Scoring Matrix
R5	CCR5-using HI virus
R5X4	Dual-tropic HI virus
RANTES	Regulated on Activation Normal T Cell Expressed and Secreted

RF	Random Forest
RNA	Ribonucleic Acid
ROC	Receiver Operatting Characteristic
rsCD4	Recombinant Soluble CD4
RT	Reverse Transcriptase
SAS	Solvent-Accessible Surface
SDF-1	Stromal Cell Derived Factor-1
SVM	Support-Vector Machines
TPR	True Positive Rate
UDS	Ultra-Deep Sequencing
V3	Variable Loop 3
VL	Viral Load
VVC	Vicriviroc
X4	CXCR4-using HI virus

List of Figures

1.1 AIDS epidemic progression . . . 13

1.2 AIDS epidemic prevalence . . . 14

1.3 HIV morphology . . . 16

1.4 HIV lifecycle . . . 17

1.5 Envelope spike . . . 22

1.6 GPCR schematic . . . 23

1.7 GPCR homology model in membrane . . . 24

1.8 V3 loop length distribution . . . 25

1.9 V3 sequence logos . . . 26

1.10 V3 loop structure . . . 27

1.11 Maraviroc . . . 29

1.12 Vicriviroc . . . 31

2.1 Binary decision tree . . . 38

2.2 Drug resistance decision tree . . . 39

2.3 Electrostatic potential of CCR5 and CXCR4 . . . 42

2.4 Electrostatic hull concept . . . 45

2.5 Origin and subtype of training sequences . . . 51

2.6 Electrostatic hull, volume variation . . . 52

2.7 Cutoff-dependent electrostatics-model accuracy . . . 54

2.8 ROC curve of electrostatics model . . . 55

2.9 Sequence similarities within data sets . . . 57

2.10 ROC curves comparing crossvalidation schemes . . . 58

2.11 Cutoff-dependent electrostatics-model accuracy, crossvalidation scheme comparison . . . 59

2.12 Cutoff-dependent hydrophobicity model accuracy . . . 61

2.13 ROC curve of hydrophobicity model . . . 62

2.14 Scatterplot of electrostatic vs. hydrophobicity votes . . . 64

2.15 ROC curve of fused random forest model 66
2.16 Prediction landscapes of second-level classifiers 69
2.17 Genotypic prediction method performance comparison 71
2.18 Sequence distance distribution of blind data from training data 73
2.19 Hydrophobicity model variable importance 74
2.20 Electrostatics model variable importance 76

3.1 Standard Sanger sequencing 79
3.2 454 next generation sequencing 82
3.3 Expected variant occurrence by error 86
3.4 Quasispecies co-receptor tropism progression 89
3.5 Correlation of prominent minority variants vs. population 91
3.6 Quasispecies co-receptor tropism evolution 93
3.7 Scatterplot of prediction distance vs. alignment distance 96
3.8 Performance with varying read coverage 98

List of Tables

1.1 HIV drug classes 20
1.2 Tropism prediction webservers 34

2.1 Electrostatic training data 53
2.2 Hydrophobicity training data 60
2.3 Combined training data 65

3.1 Variants analyzed for quasispecies data 88
3.2 Most prominent X4 minority sequences 90

1 Introduction to HIV-1 and Co-receptor Tropism

«Auch die Pause gehört zum Rhythmus.»

– Stefan Zweig

1.1 HIV and AIDS

1.1.1 History of HIV and AIDS

The Acquired Immunodeficiency Syndrome (AIDS) was first described in 1981, with reports of an accumulation of several rare conditions, unusual for the predominately young group of patients [45, 92, 135, 40]. Similar cases were reported over the next months, showing a distinct similarity, the drastic reduction of CD4-positive T-lymphocytes, an important part of the immune system. It became clear, that the underlying cause for the witnessed diseases was a deficiency of the immune system, making patients vulnerable to opportunistic infections. Reports of hemophiliac patients with severe immune deficiencies, prompted discussions about transmission through blood products [54]. The facts of declining numbers of CD4-positive cells and the suspicion of transmissions through blood put the focus on the human T-lymphotropic virus family (HTLV) [41], the only virus known at that time to be able to infect CD4-lymphocytes [112]. The virus responsible for the development of AIDS was successfully isolated by the group of Luc Montagnier in 1983 [6], termed lymphadenopathy-associated virus (LAV) and again in 1984 by the group of Robert Gallo [42], under the name of HTLV-III. The pathogen of the family of lentiviruses, was later termed human immunodeficiency virus (HIV). The HI virus exists in two variants, HIV-1 and HIV-2, both of which have probably emerged from the closely related simian immunodeficiency virus (SIV) [132] found in monkeys. Through phylogenetic analysis the origin was dated to the late 19th or early 20th century in west-central

Africa [166]. Since then AIDS has become a global epidemic, affecting every country of the world [54].

1.1.2 The Global AIDS Epidemic

Since the first reported case of AIDS, the number of people living with HIV has steadily increased (see Figure 1.1). In 1990 around 8 million total infections have been reported. Twenty years later, this number has increased to over 33 million cases reported in 2008 [153]. From these, 22.5 million people are living in Sub-Saharan Africa, where the combined overall prevalence, the percentage of people living with HIV, in adults (15-49 years) is 5% (see Figure 1.2). In some regions in Southern Africa the prevalence is as high as 25% [154]. In 2008 72% of all new HIV infections occurred in Sub-Saharan Africa [103]. On the continent, HIV is the most common cause of death, and in some regions has lead to a reduction of the average life-expectancy by almost 20 years [54].

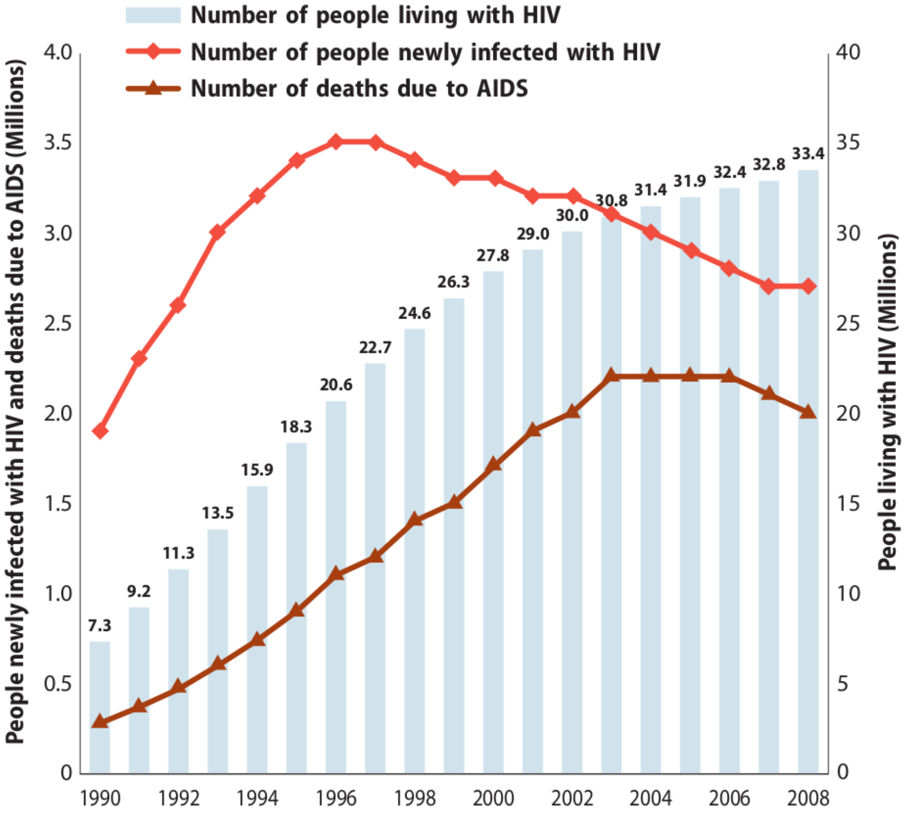

Figure 1.1: **Progression of the global epidemic.** *Source: United Nations Millennium Development Goals Report 2010 [103]*

Despite these devastating facts, the newly published UNAIDS Report on the Global AIDS Epidemic in 2010 [154] draws a more optimistic picture. The report highlights an overall deceleration of the epidemic growth. Since the peak of epidemic spreading in 1997, when 3.2 million new infections were reported, this number has steadily decreased to around 2.6 million in 2009, a reduction by 21% (see Figure 1.1). This trend is mostly driven by the positive development in Sub-Saharan Africa, where the number of new infections per year has dropped by almost 28% since 1997. The authors attribute the successful reduction to HIV prevention efforts, like safe-sex and prevention of mother-to-child-transmission. A similar development is reported for the number of AIDS-related deaths, shown in Figure 1.1. From the peak in 2004, when 2.1 million deaths were reported, the number decreased steadily to an estimated 1.8 million in 2009. Again, the trend is lead by a decrease in Sub-Saharan Africa which saw a reduction of 1.6 million AIDS-related deaths in 2004 to around 1.3 million in 2009. This development is coupled to the steadily decreasing number of new infections. It is also influenced by the improved availability of antiretroviral therapy in low- and middle-income countries, where 42% of people infected received therapy in 2008, compared to only 33% in 2007 [103]. With the exception of Eastern Europe and Central Asia the epidemic appears to have stabilized.

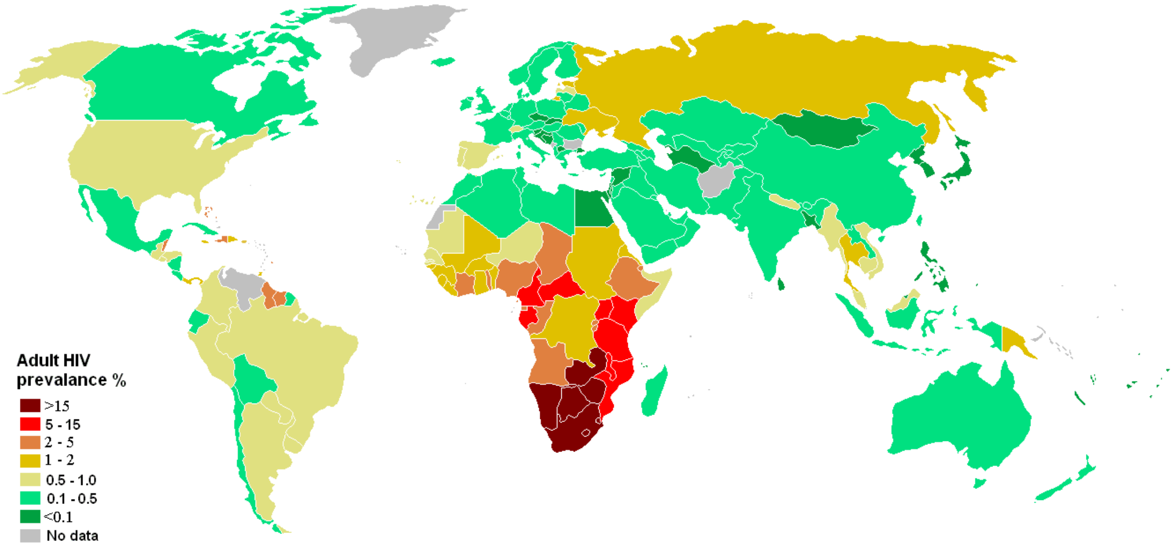

Figure 1.2: **Number of people newly infected with HIV.** Estimated HIV/AIDS prevalence among young adults (15-49) by country as of 2008 [161]. Based on the UNAIDS Global Report 2008. *Source: Wikipedia*

1.2 HIV Structure and Lifecycle

The human immunodeficiency virus consists of an envelope containing the viral core which in turn contains several viral proteins and the viral genome [11]. A schematic representation is given in Figure 1.3. The envelope, also known as coat, is of spherical shape and about 120 nm in diameter. The coat is formed by a double layer of lipids, extracted from the host cell membrane during the budding process by which new viruses leave the cell. Inserted into the lipid membrane are a varying number of envelope (Env) proteins required for virus cell entry [150]. The Env proteins are trimers of glycoproteins 120 (gp120) anchored to the lipid membrane by a trimer-bundle of alpha-helical gp41. Each Env protein is about 10 nm in diameter [54]. At the interior side of the virus envelope, matrix-proteins p17 are attached. The viral core structure, also referred to as capsid, is contained in the envelope and has cone-like shape. Its outer structure is formed by about 2000 copies of the matrix-protein p24. The capsid contains two complete copies of the virus genome as RNA bound to the nucleocapsid protein p7, as well as the viral enzymes reverse transcriptase, integrase and protease, required for the replication process.

The primary target of HIV are T-cells, but other cells like macrophages or monocytes can also be used as hosts for viral replication [16]. The life cycle of HIV goes from cell-entry and uncoating, over reverse transcription and integration of the viral genome into the host cell genome, to the assembly and maturation of new HIV virions. The complete cycle is shown schematically in Figure 1.4.

The entry of HIV into host cells is mediated by interaction of viral coat-protein gp120 and cell the T-cell surface receptor CD4. In addition to CD4, host-cell chemokine receptors are required to establish the fusion of HIV with the host cell membrane and complete cell entry. The entry process is central to this work and will be described in detail in Section 1.4. After completing membrane fusion the viral capsid is released into the cytoplasm, a process termed “uncoating“. The reverse transcriptase then detaches the viral RNA from the p7 nucleoprotein and starts with the transcription into viral DNA. This is done by synthesizing the cDNA strand complementary to the single stranded viral RNA, which is, in parallel, being degraded by the reverse transcriptase. Subsequently, the complementary *sense* strand to the *antisense* cDNA strand is created. The complete viral DNA molecules are then transported to the host cell nucleus where they are integrated into the host genome by the HIV integrase. The transcription of the integrated proviral DNA to mRNA is the next step towards the development of new virions. The mRNA is cleaved into fragments coding for small regulatory proteins Tat and Rev.

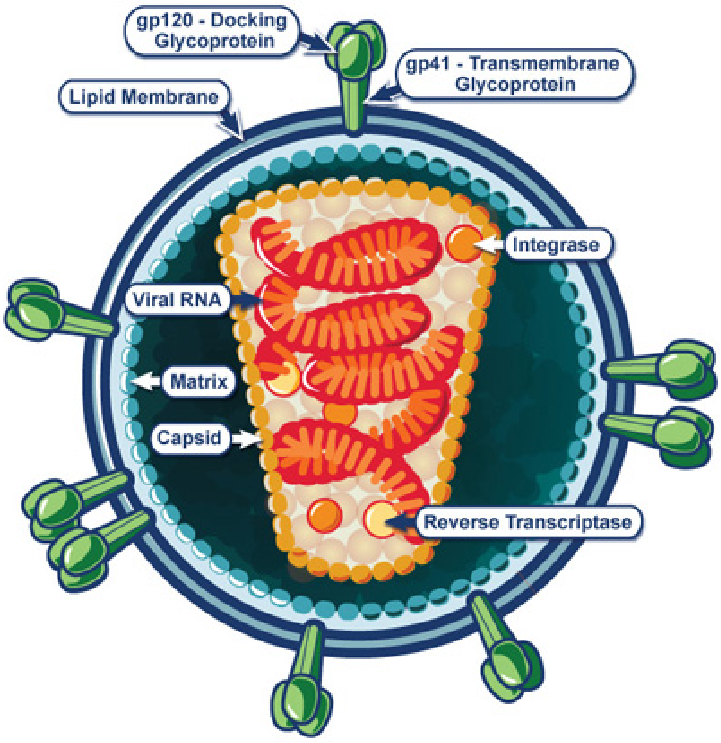

Figure 1.3: **Schematic morphology of the HIV.** Source: National Institute of Allergy and Infectious Diseases (NIAID) [102]

Tat leads to an increased expression of viral mRNA, while Rev regulates the expression of virus Gag-precursor, Gag-pol-precursor, and Env-precursor polyproteins. The Env-polyprotein is cleaved by the HIV protease into gp41 and gp120 and inserted into the host cell membrane. The Gag-, Gag-pol-precursors proteins, viral RNA, and several cellular proteins are transported to the cell membrane, associating with the inner cell membrane, where they are enclosed in the newly forming virion, starting the budding process. After budding, virions are still immature. During the following maturation process, the HIV protease cleaves the virus-membrane associated Gag-, and Gag-pol precursor-proteins into their functional structural units, which then begin to form the virus capsid. By the same protease, Gag-pol polyproteins are cleaved into functional virus enzymes. The mature virus is now infectious, closing the replication cycle.

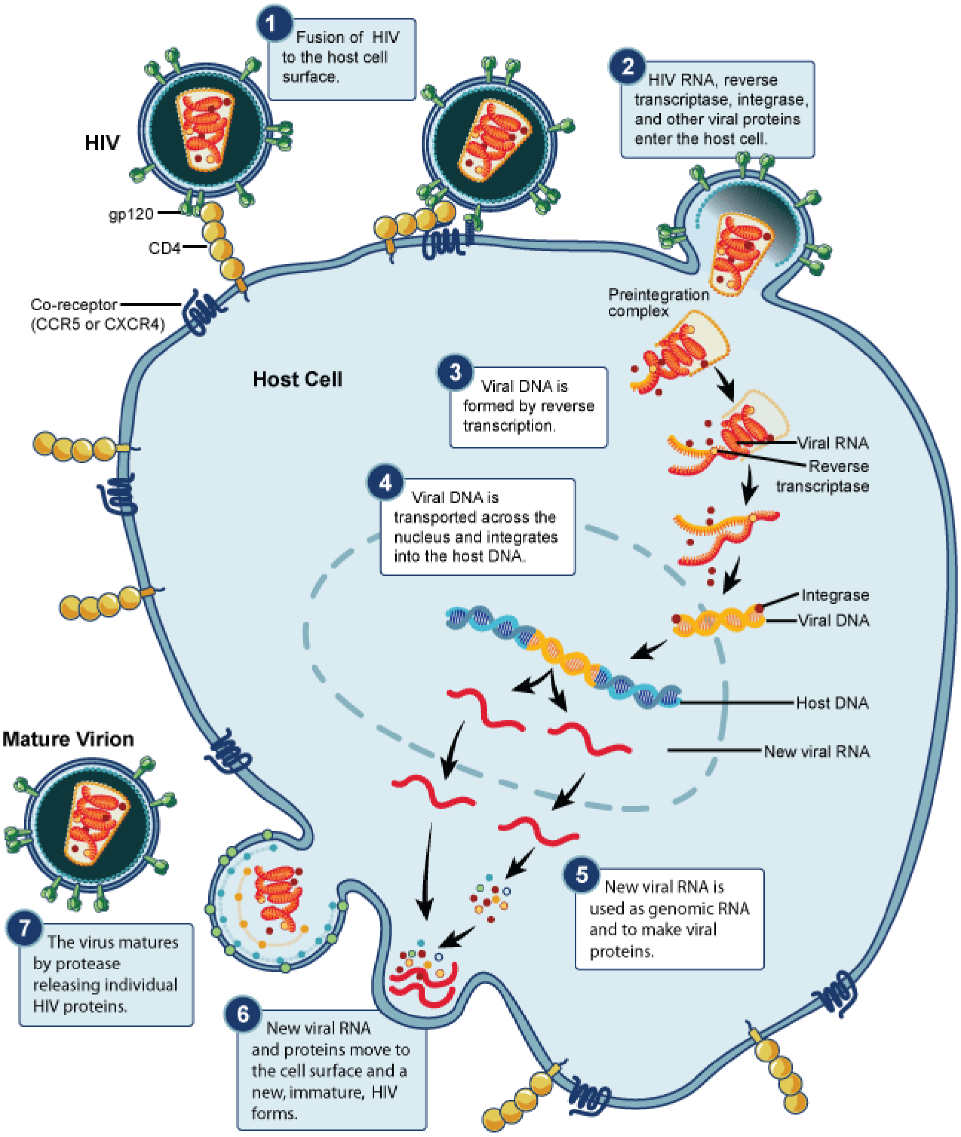
1 Fusion of HIV to the host cell surface.
2 HIV RNA, reverse transcriptase, integrase, and other viral proteins enter the host cell.
HIV
gp120
CD4
Co-receptor (CCR5 or CXCR4)
Host Cell
Preintegration complex
3 Viral DNA is formed by reverse transcription.
Viral RNA
Reverse transcriptase
4 Viral DNA is transported across the nucleus and integrates into the host DNA.
Integrase
Viral DNA
Host DNA
New viral RNA
Mature Virion
5 New viral RNA is used as genomic RNA and to make viral proteins.
7 The virus matures by protease releasing individual HIV proteins.
6 New viral RNA and proteins move to the cell surface and a new, immature, HIV forms.

Figure 1.4: **Schematic representation of the HIV lifecycle.** Source: National Institute of Allergy and Infectious Diseases (NIAID) [102]

1.3 Antiretroviral Therapy

Many of the different stages of the HIV replication are targets for antiretroviral drugs, aiming at breaking the replication cycle, and thus, enabling an effective antiretroviral therapy (ART). According to the German-Austrian Guidelines for ART in HIV-1 Infections, the goals of antiretroviral therapy today, is to suppress infection-based symptoms, slow disease progression, reestablish cellular immunity and reduce the chronic immune-activation by inhibiting the HIV-replication [27]. To ensure this, the effectiveness of the drug regimen, the combination of antiretroviral drugs, has to be sustained as long as possible. Resistance mutations of the virus have to be avoided. This can be achieved by maintaining a steady drug intake. A good patient compliance, however, is more likely when side-effects, a common problem of HIV therapy are limited to a minimum. Effective substances, causing severe side-effects are not suited for therapy, as they will probably lead to poor patient compliance. The central parameters, governing the success of antiretroviral therapy are the efficacy of drugs, the absence of resistances, and good patient compliance [130], ensuring a consequent drug intake. The first HIV therapies were mono-therapies, consisting of single drugs administered to patients. Their effectiveness were, however, often compromised by rapid development of drug resistances. Studies performed in the mid-1990s showed improved effectiveness and reduced emergence of drugs resistance using combination-therapies consisting of more than one drug [18, 48]. Today, this therapy, known as highly active antiretroviral therapy (HAART) is the accepted standard, with drug regimen consisting on average of three different antiretroviral drugs of different classes. Indicators for the initiation of HAART taken from the German-Austrian Guidelines (2010) [27] are different for patient with and without HIV-associated symptoms. In the latter case, initiation of HAART is generally recommended. In the case of asymptomatic patients, the CD4 cell count, defined as the number of CD4+ T-Lymphocyte cells per volume (CD4/μl), is used as decisive measure. A low CD4 count is a severe immunological condition, thus it is recommended that HAART should be started with CD4 counts of less than 350/μl. The immunological success of HAART is often defined as achieving an increase in CD4 cell count. Unfortunately, the immunological response is hard to influence, varies between patients, and is a weak indicator for therapy outcome. Thus, for monitoring the success of HAART, CD4 cells counts are not very well suited, because as a central part of the immune system, many factors, besides HIV, will affect the number of CD4 cells. The viral load (VL), a virological parameter defined as the number of copies of viral RNA per volume (copies/mL), is easier to measure and serves as a direct indicator for the response to therapy [54]. To circumvent a probable decrease in CD4 cell counts, it is recommended to initiate HAART at viral loads of more than 100.000 copies/mL. Virological success is achieved by pushing the VL under detection limits of 50 copies/mL, whereas virologic failure is regarded as failing to reach VL of <50 copies/mL six months after initiation of HAART, or a VL of >5000 copies/mL. Although VL and CD4 cell counts are correlated, virologic failure does not always imply immunological failure [95, 24]. In fact, it seems more important to achieve a reduction of viral load compared to the baseline (VL at treatment start), than to achieve the absolute VL of <50 copies/mL [78, 25]. Several studies have shown benefits of earlier treatment starts, such as the reduction of clinical complications and side-effects or improved immune reactions. In addition, low levels of viremia have been shown to significantly reduce the infectiousness and rate of sexual viral transmission [4]. The rate of progression towards AIDS was described as being higher for later treatment starts, in several studies [110, 144]. However, earlier treatment increases the risk for long-term toxicity and for developing drug resistances. Based on current knowledge, HAART should be regarded as a life-long, continuous therapy.

1.3.1 Antiretroviral Drugs

In clinical HIV treatment four classes of drugs are currently administered. Entry inhibitors prohibit the virus from entering the host cell or integrating itself into the host DNA where it would be replicated. Cell fusion can be inhibited by targeting virus envelope proteins gp120 or gp41, as well as host co-receptor CCR5. Shortly after viral entry, reverse transcriptase inhibitors prohibit the HIV reverse transcriptase from translating the single-stranded HIV RNA into double-stranded DNA. Reverse transcriptase inhibitors are divided into nucleosidic- and non-nucleosidic reverse transcriptase inhibitors, so-called NRTIs and NNRTIs. While NNRTIs are simply small-molecule antagonists inhibiting a reverse transcriptase active site, NRTIs are nucleoside-analogs that, when integrated, terminate chain elongation during the reverse transcription of viral RNA to DNA. Integration of viral RNA into host DNA can be prevented by a relatively new drug class blocking the HIV integrase. Another class, the protease inhibitors (PI), are acting at a later stage of the virus life cycle. The HIV protease is blocked in a way that the newly synthesized HIV precursor-polyproteins Gag, Gag-pol, or Env cannot be cleaved correctly, resulting in defective proteins. Table 1.1 gives an overview over drugs currently administered in clinical practice.

During the last years there has been a substantial focus on the development of a new drug class: entry inhibitors. When interfering with the cell entry process, one can target virus proteins, like other drug classes, or develop drugs that bind to host-cell proteins. In the latter case, the virus cannot impair the drug binding affinity, making it harder to develop resistance mutations. Targeting the entry of HIV into the host cell is not an entirely new field. In the early 1990s so called attachment inhibitors attempted to block HIV-1 gp120 by recombinant soluble CD4 (rsCD4) molecules. These were, however, discontinued due to low efficacy in phase I clinical trials [23]. The first entry inhibitor to be licensed was enfuvirtide in 2003, targeting virus envelope protein gp41. A so called

Table 1.1: **List of antiretroviral drug classes used in ART.** Adapted form the HIV-Buch 2010 [54]

Trade Name	Abbr.	Substance	Company	Class
Emtriva	FTC	Emtricitabine	Gilead	NRTI
Epivir	3TC	Lamivudine	GSK	NRTI
Retrovir	AZT	Zidovudine	GSK	NRTI
Videx	DDI	Didanosine	BMS	NRTI
Viread	TDF	Tenofovir	Gilead	NRTI
Zerit	D4T	Stavudine	BMS	NRTI
Ziagen	ABC	Abacavir	GSK	NRTI
Sustiva (Stocring)	EFV	Efavirenz	BMS/MSD	NNRTI
Viramune	NVP	Nevirapine	Boehringer	NNRTI
Intelence	ETV	Etravirine	Tibotec	NNRTI
Rescriptor	DLV	Delavirdine	Pfizer	NNRTI
Aptivus	TPV	Tipranavir*	Boehringer	PI
Crixivan	IDV	Indinavir*	MSD	PI
Invirase	SQV	Saquinavir*	Roche	PI
Kaletra	LPV	Lopinavir/Ritonavir	Abbott	PI
Norvir (as Booster)*	RTV	Ritonavir	Abbott	PI
Prezista	DRV	Darunavir*	Tibotec	PI
Reyataz	ATV	Atazanavir*	BMS	PI
Telzir (Lexivav)	FPV	Fosamprenavir*	GSK	PI
Viracept	NFV	Nelfinavir	Roche/Pfizer	PI
Celsentri (Selzentry)	MVC	Maraviroc	Pfizer	EI
Fuzeon	T-20	Enfuvirtide	Roche	EI
Isentress	RAL	Raltegravir	MSD	II
Atripla	ATP	TDF+FTC+EFV	Gilead+BMS+MSD	Comb.
Combivir	CBV	AZT+3TC	GSK	Comb.
Kivexa (Epzicom)	KVX	3TC+ABC	GSK	Comb.
Trizivir	TZV	AZT+3TC+ABC	GSK	Comb.
Truvada	TVD	TDF+FTC	Gilead	Comb.

post-attachment inhibitor, the monoclonal antibody ibalizumab binds to the host receptor CD4, limiting the molecules flexibility and hindering gp120 binding [13]. First results have shown promising results, yielding a reduction in viral load of 1.5 log_{10} copies/mL [74]. The discovery that a homozygous non-lethal deletion in the host cell chemokine co-receptor CCR5 (CCR5Δ32), results in immunity against HIV infections [59] while the heterozygous deletion delays disease progression [79], spurred the development of co-receptor antagonists, targeting CCR5. The HIV-1 cell entry mechanism, as well as targeting of host-cell co-receptors is central to the understanding of this work and will be discussed in detail in Section 1.4.

1.3.2 Resistances and Testing

In untreated patients around 10^9 new virus particles are generated every day [109]. This very high turnover rate of the HI virus, in combination with an error-prone replication process – the reverse transcriptase generates around one mutation in 10^4 to 10^5 bases per cycle – leads to a fast development of drug resistant viruses. This process can best be suppressed by maintaining a constantly low level of viremia, minimizing the amount of newly generated virus variants, and by administering at least two different drug classes, targeting different steps of the virus replication cycle. Both aspects are covered by current HAART, and patients today are less likely to develop virologic failure due to resistance mutations than in the pre-HAART days [85, 77]. Still, regular tests for drug resistance-related mutations are vital for ensuring effective treatment and to adapt the drug regimen accordingly. The German-Austrian therapy guidelines recommend resistance testing prior to treatment start in therapy-naïve patients, and generally after therapy failure in all patients [27], in order to determine the effectiveness of future treatments.

1.4 HIV-1 Cell Entry and Co-receptor Tropism

1.4.1 Entry mechanism

The entry of new host cells by the HIV virus is a multi-stage process, involving the attachment, co-receptor binding and membrane fusion. The major viral component of these processes is the viral spike, a trimeric complex of HIV envelope glycoproteins (gp) gp41 and gp120. During HIV replication the glycoprotein gp160 is expressed as polyprotein and subsequently cleaved into gp41 and gp120 by the HIV protease. The viral spike consists of a multi-helix bundle structure of gp41 which is non-covalently linked to the trimer of gp120.

The entry process begins by binding of the gp120 bridging sheet to the host-cell CD4-receptor. This event triggers the detachment of the variable loop 3 (V3) of gp120 which is now free to associate with a nearby host cell chemokine co-receptor, further stabilizing the attachment of the virus to the host cell (see Figure 1.5). Next, the previously inaccessible gp41 is inserted into the host cell membrane and mediates fusion of the

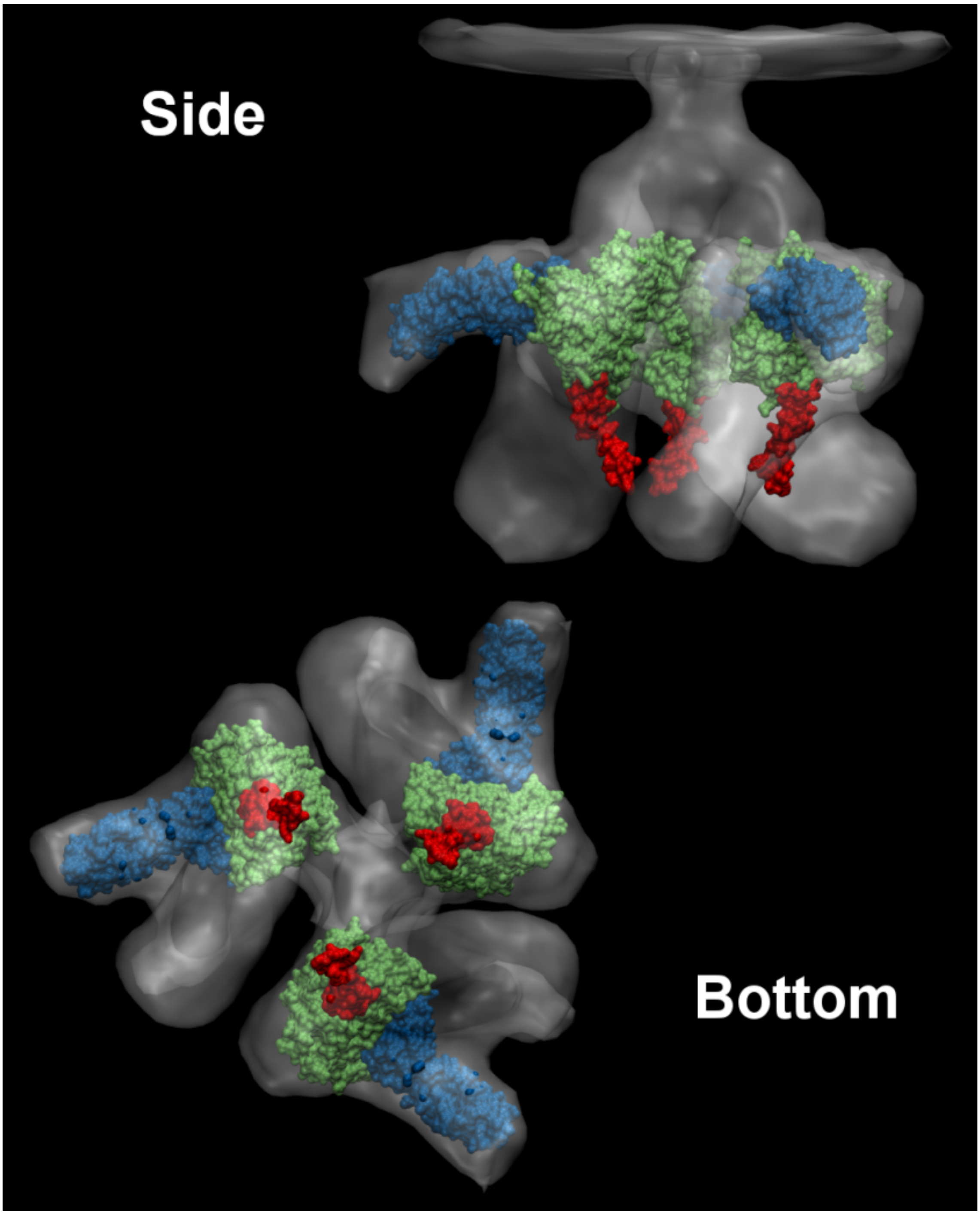

Figure 1.5: **HIV envelope spike structure.** Structure of the trimeric envelope spike. Trimeric gp120 (green) is bound to cell receptors CD4 (blue). The V3 loop (red) is detached and extends into the solvent after CD4 binding. Based on PDB structures 2QAD [56], electron microscopy data (white shade) from Liu *et al.* [83]

lipid membrane of the virus and host cell [60]. Glycoprotein gp120 can interact with different classes of chemokine co-receptors, although most commonly either Chemokine Co-receptors CCR5 or CXCR4 are used. Both belong to the family of G-Protein Coupled Receptors (GPCR), the largest and most diverse group of receptors involved in signaling in eukaryotes [89]. The human genome holds over 1000 GPCRs, including receptors for odors, hormones, nucleotides, pheromones and many more. GPCRs are involved in many biological processes, making them major targets for medical drugs. Over 26% of drugs approved by the American Food and Drug Association (FDA) are targeting Rhodopsin-like GPCRs [107].

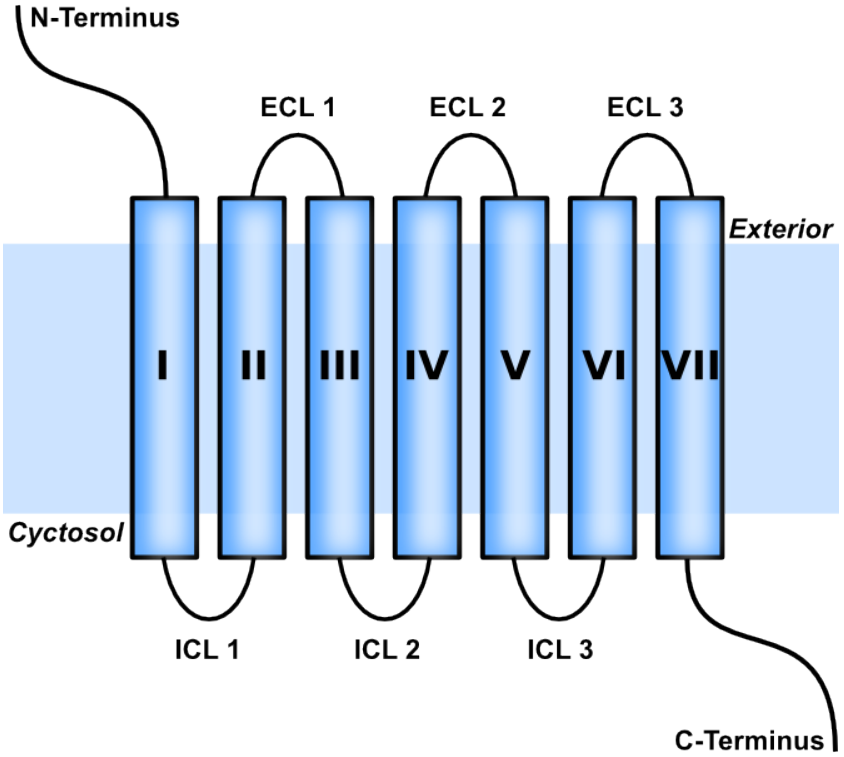

Figure 1.6: **Schematic representation of a G-Protein Coupled Receptor (GPCR).** The seven membrane-spanning helices (I-VII) are connected by three extracellular loops (ECL 1-3), three intracellular loops (ICL 1-3) while the first (I) and last (VII) transmembrane helix have an N-terminal and C-terminal domain, respectively.

The structure of these integral membrane proteins consists of seven membrane-spanning, or transmembrane (7TM) α-helices linked by loops of varying length and structure (see Figure 1.6). Accoding to their location, these loops are referred to as extracellular or intracellular loops (three ECLs and three ICLs, respectively). In chemokine co-receptors, cysteines forming stabilizing disulfide bridges are common features of the extracellular loops. The N-terminal tail of both CCR5 and CXCR4 extend into the cell exterior, while the C-terminus reaches into the cytosol. Certain residues on both termini are often subject to post-translational modifications like serine and threonine to phosphorylation or tyrosine to sulfation. A homology model of the CCR5 co-receptor embedded in a lipid membrane is shown seen in Figure 1.7. CCR5 is most commonly expressed on T-Cells, macrophages, dendritic cells and microglia, the primary ligands are pro-inflammatory cytokines MIP-1α, MIP-1β, and RANTES, all of which have been found to act as competitive inhibitors against HIV infection [86].

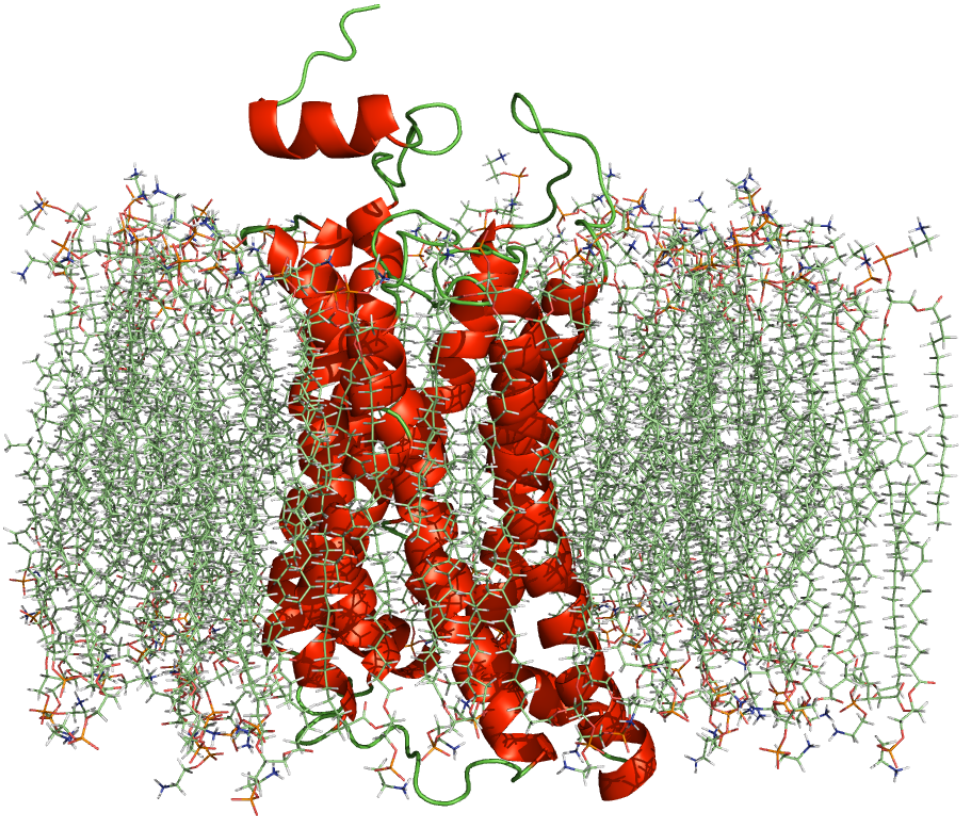

Figure 1.7: **Model of CCR5 in lipid bilayer.** The picture shows a homology model of the human CCR5 co-receptor (red transmembrane helices and green loops) embedded into a lipid bilayer slab (mint green).

The type of co-receptor used by the virus to induce cell fusion is strongly dependent of the V3 loop of gp120. This loop of, on average, 35 aminoacids is very variable in primary sequence, mutations including substitutions, deletions, and insertions (see Figures

1.8 and 1.9). The Los Alamos HIV Sequence Database (http://www.hiv.lanl.gov/) currently contains around 36.000 unique V3 sequences. Next generation sequencing experiments of HIV positive patients have revealed quasispecies diversity of hundreds of unique sequences in some patients [151, 121]. The extent, however, to which minor variants are caused by sequencing errors is still unclear.

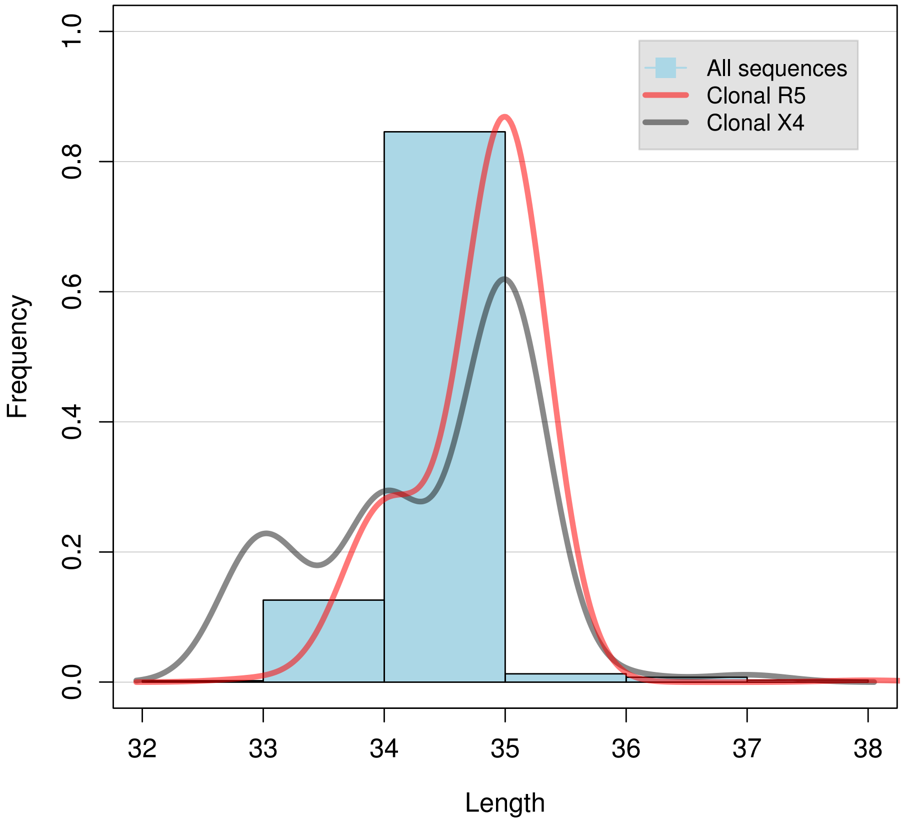

Figure 1.8: **Length distribution of the variable loop 3 (V3) of HIV gp120.** Sequence lengths of the two classes (R5 and X4) show a clear difference, with the X4 sequences having a tendency towards shorter (33) loop sizes. Sequences with known co-receptor tropism were taken from the Los Alamos HIV Sequence Database (http://www.hiv.lanl.gov/).

1.4.2 Structural Data

In recent years, various groups have successfully determined the structure of proteins involved in the HIV cell entry mechanism. In 1996 Wu *et al.* [168] solved the structure of T-Cell CD4, including various mutants unable to bind gp120. Two years later,

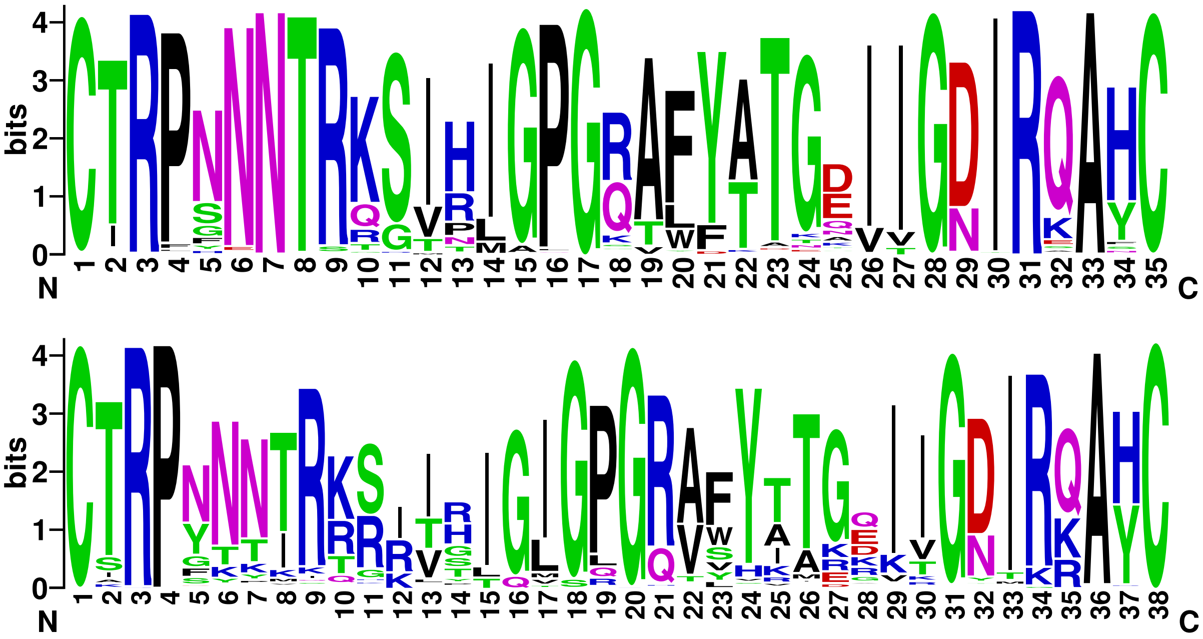

Figure 1.9: **Weblogos of clonal V3 sequences.** Weblogos [21] showing the frequency of aminoacids for the different sequence positions of the V3 loop. Weblogos were generated for both the R5 (top) and X4 (bottom) tropic sequences. Both logos show the strong heterogeneity of the V3 loop. Sequences with known tropism were acquired from the Los Alamos HIV Sequence Database.

CD4 was crystallized by Kwong *et al.* [75], bound to gp120 and a neutralizing antibody. This study shed light on the interaction of gp120 and CD4, however the flexible V3 region was not captured and structural features remained unknown until 2005, when the group of Huang [57] successfully solved the structure of gp120 in complex with CD4 and an antibody. Earlier, several groups had already experimented with V3-like peptides [159, 152, 131, 119]. In 2008, gp120 was already well known and studied as a monomer, when Liu *et al.* [83] solved the biological trimeric gp120 of the viral spike bound to CD4, by means of cryo-electron tomography. The published electron density map covers the complete HIV viral spike, including gp41. Just recently, Liu *et al.* [84] have solved the monomeric form of the C-terminal ectodomain of gp41, responsible for interaction with gp120. These advances are very promising and suggest the complete structural determination of the viral spike within the next few years. Unfortunately, the remaining proteins involved in the fusion process, host chemokine co-receptors CCR5 and CXCR4, are not well described structurally, as crystallization of integral membrane proteins like GPCRs is challenging [170]. A major challenge is finding a suitable type of lipid to form a host micelle for the protein. Then, proteins have to be purified without losing lipid

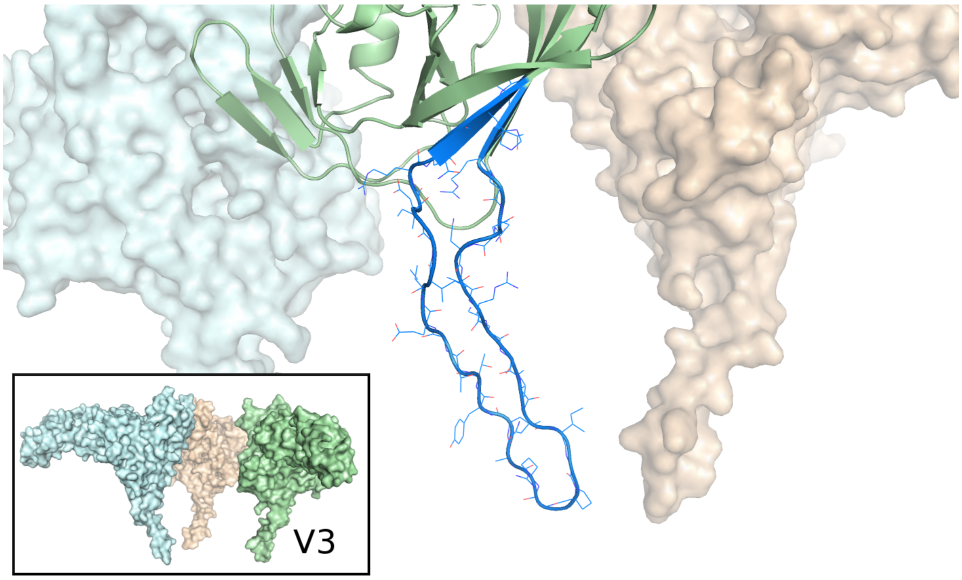

Figure 1.10: **Structure of the V3 loop.** The V3 loop is shown as blue ribbon and extends from one of the gp120 envelope proteins of the gp120 trimeric structure which is shown in the box in the lower left corner. V3 loop structure taken from PDB structure 2QAD [56].

association. Additionally, exposed and flexible loops, often found in GPCRs, are subject to protease degradation. Until recently, the only structural entry of human chemokine co-receptors CCR5 and CXCR4 were short fragments of the N-terminal extracellular domain. In 2007 Huang *et al.* [56] presented the crystal structure of gp120 bound to CD4 and a tyrosine-sulfated antibody targeting the V3 loop. They found that one of the sulfated tyrosines seems to mimic the binding of the N-terminal co-receptor domain. To test their hypothesis, a short N-terminal peptide of CCR5 including residues 5-12 was solved by Nuclear Magnetic Resonance (NMR). Docking of the peptide to the crystallized V3 loop yielded the first experimentally determined structural model for the interaction of gp120 with a co-receptor. The structure of the N-terminus of CXCR4, in form a peptide bound to its natural ligand, the chemokine SDF-1, has been solved in 2008 [157]. From that structure it has been hypothesized that N-terminal tyrosines are sulfated, due to their interaction with positively charged clefts on SDF-1. However, a real breakthrough, which will further clarify the binding mode of V3 to host cell co-receptors was published just prior to this work. In late 2010, Wu *et al.* published five independent crystal structures of CXCR4 with bound small molecules and a cyclic peptide at high resolutions [167]. Structural data of the complete CCR5 co-receptor remains unavailable. However, template structures for homology modeling experiments of CCR5 are available, like the GPCR structures of bovine rhodopsin [108] or human $\beta 2$ adrenergic receptor [116]. Especially the crystal structure of bovine rhodopsin has been used as template for molecular modelling of the chemokine co-receptors CCR5 and CXCR4 [155, 101, 138].

1.4.3 Small-Molecule CCR5 Antagonists

Several companies have ventured to develop antagonists targeted against CCR5. However, most of the drugs that went to clinical trials have been discontinued, due to limited efficacy of adverse effects. The first small-molecule antagonist TAK-779, developed by Takeda Pharmaceutical was dropped due to poor oral availability [70]. A structural successor (TAK-652), now developed as Cenicriviroc by Tobira Therapeutics, is currently in clinical trials showing good viral potency and encouraging efficacy [69]. Aplaviroc was developed by GlaxoSmithKline until 2005, when cases of drug-related liver toxicity [104] and limited efficacy [22] lead to discontinuation of development. Until 2010 two small-molecule co-receptor antagonists, maraviroc and vicriviroc, showed promising results in various clinical trials and will be discussed in the following.

Maraviroc

The most successful CCR5 small molecule antagonist, maraviroc (MVC) [30] most likely binds to a hydrophobic cavity within the transmembrane domain of the CCR5 co-receptor [70], blocking interaction of HIV gp120. The hypothesized binding mode is shown in Figure 1.11. In 2007, maraviroc has been approved by the FDA and was recommended for treatment experienced patients. The drug was discovered through cell-based screening assays and was found to be an effective CCR5 antagonist, not affecting CCR5 expression levels or interfering with intracellular signaling. *In vitro* studies showed a good inhibitory activity of maraviroc against 43 different primary virus isolates as well as 200 clinically derived HIV-1 envelope-recombinant pseudoviruses. Additionally, maraviroc proved effective against primary isolates with known resistances against NRTIs, NNRTIs and PIs [90].

Prior to the approving of maraviroc by the FDA, two identical, parallel, double-blind, placebo-controlled phase III studies were carried out to determine the efficacy of maraviroc in treatment experienced adult patients with CCR5 HIV-1 infections [46]. The so-called MOTIVATE (Maraviroc plus Optimized Therapy In Viremic Antiretroviral Treatment Experienced patients) studies 1 and 2 assessed the efficacy and safety of

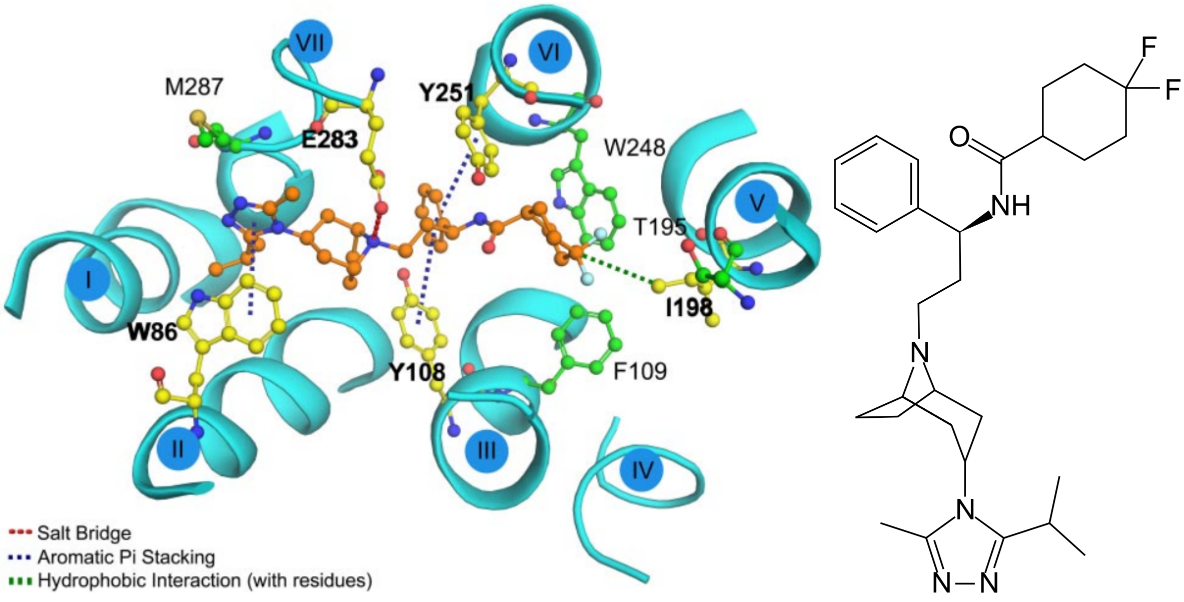

Figure 1.11: **Skeletal formula and putative binding mode of maraviroc to CCR5.** Illustration taken from Kondru *et al.* [70]

maraviroc over a time of 48 weeks, followed by an extension to period to 96 weeks. MOTIVATE 1 was carried out in the US and Canada, while MOTIVATE 2 took place in Europe, Australia and the US. Patients eligible to enroll had to be infected with R5 virus only, and have a viral load of more than 5000 copies/mL. In total, 601 and 474 patients were included into MOTIVATE 1 and 2, respectively. The study populations were randomized and split in to three arms receiving placebo or maraviroc once or twice daily. In addition, each patient received an optimized background therapy (OBT) adjusted to patient-specific resistances, consisting of three to six agents of NRTI, NNRTI, PI, and fusion inhibitors. The efficacy of maraviroc was analyzed at the primary end-point at week 48. In both studies MOTIVATE 1 and 2, patients receiving maraviroc plus OBT showed a significantly stronger reduction of viral load compared to patients receiving placebo plus OBT. The efficacy of the twice-daily arm was also significantly higher than the once-daily arm. The pooled results of both studies showed a mean change of viral load of -0.79, -1.68, and -1.84 $\log_{10}$ copies/mL for the placebo, once-, and twice-daily arms. The frequency of adverse effects was described as being equal between the different arms.

In the MOTIVATE studies maraviroc was administered to treatment experienced patients, however the highest frequency of R5 viruses, target of maraviroc is witnessed in treatment naïve patients [96]. The MERIT (Maraviroc versus Efavirenz in Treatment-

Naive Patients) phase III study compared the efficacy and safety of maraviroc to efavirenz in treatment-naïve patients [19]. The study was started in 2004 and aimed at showing maraviroc to be non-inferior to the NNRTI efavirenz in combination with the regular drug regimen at that time (NRTIs zidovudine and lamivudine). The study setup was similar to that of the MOTIVATE studies and carried out in Australia, Europe, South Africa, North and South America. The three study arms consisted of patients receiving maraviroc once-daily, twice-daily or efavirenz once-daily, each in combination with lamivudine and zidovudine. The maraviroc twice-daily arm established non-inferiority to efavirenz after 48 weeks with 70.6% of patients (maraviroc, twice-daily) reaching a viral load of $<$400 copies/mL, while 73.1% of patients receiving efavirenz reached that goal. A reanalysis of the results was performed after it had become clear that the assay for determining exclusive R5 virus in patients lacked desired sensitivity. A *post hoc* re-screening of patients, using an enhanced version of the same assay, revealed X4-using virus in an additional around 15% (106 of 721) patients. A reanalysis, excluding these patients, showed an improvement of therapy outcome on the maraviroc twice-daily arm. A viral load of $<$400 copies/mL was achieved by 73.3% of patients receiving maraviroc, compared to 72.3% of efavirenz-receiving patients.

Vicriviroc

Vicriviroc (VVC) is a small-molecule CCR5 antagonist initially developed by Schering-Plough [73]. Both skeletal formula and binding mode are similar to maraviroc (compare Figure 1.12) [70]. The drug has been shown to have a good oral bio-availability and long half-life and thus can be administered once daily [62].

The AIDS Clinical Trials Group (ACTG) 5211 phase II study was carried out to determine the efficacy of vicriviroc in treatment-experienced patients over a period of initially 48 weeks [47] and a follow up until week 96 [162]. The study setup was similar to the MOTIVATE 1 and 2 studies: double-blinded, placebo-controlled, and randomized. Patients were then randomized to four groups receiving 5, 10, 15 mg of vicriviroc once daily or placebo, respectively, with a background regimen containing the protease inhibitor ritonavir. At 24 weeks the mean change of baseline viral load was -0.29 $\log_{10}$ copies/mL for the placebo arm, and -1.51, -1.86, and -1.68 $\log_{10}$ copies/mL for the 5, 10, and 15 mg VVC arms, respectively. The 5 mg VVC study arm was discontinued after 48 weeks due to suboptimal treatment results.

In the recent double-blinded, placebo-controlled, phase III VICTOR-E3 and E4 studies [44] vicriviroc surprisingly failed to establish superior efficacy over a placebo control

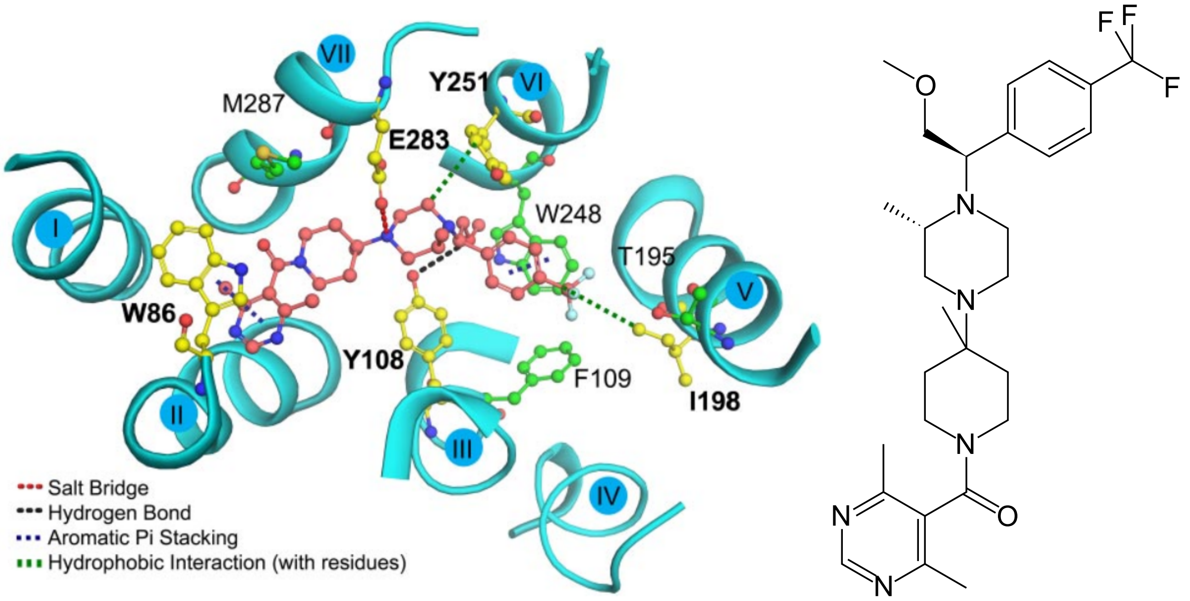

Figure 1.12: **Skeletal formula and putative binding mode of vicriviroc to CCR5.** Illustration taken from Kondru *et al.* [70]

group. The VICTOR-E3 study was carried out in South America, the VICTOR-E4 study took place in Australia, Europe, South Africa and the US. Study setup was identical, including treatment experienced patients with exclusive R5-virus. The two arms of both studies consisted of regimen of 30 mg VVC or placebo, both with OBT. Pooled results show that after 48 weeks 64% of VVC-receiving patients achieved viral loads of <50 copies/mL, compared to 62% of patients in the placebo arm. 72% of VVC-patients achieved viral loads of <400 copies/mL, the same was achieved by 71% of placebo patients. The authors attribute this disappointing result to the success of the OBT. Of all those patients receiving a background regimen with less than three drugs, 70% patients of the VVC arm achieved VL < 50 copies/mL, compared to only 55% of placebo-receiving patients. Using the same measure, no significant difference between the arms can be seen for patients receiving more than three or more drugs in their background regimen. Overall, 67% of patients received a strong background regimen (⩾3 drugs), compared to only 37% in the MOTIVATE studies. In summary, VVC was shown to be safe and well-tolerable but failed to meet efficacy end-points, which lead to the termination of all ongoing vicriviroc studies. Merck, developing vicriviroc after merging with Schering-Plough, has decided to not further pursue regulatory approval.

Effect of X4 Virus Presence on Therapy Outcome

The results from the MERIT study highlighted the strong influence of X4 viruses on therapy outcome for co-receptor antagonists. In the initial analysis of the MERIT study, around 15% of patients included had significant levels of X4 virus, reducing the efficacy of maraviroc for the study group. Detection of the X4 virus in these patients by a first generation phenotypic assay failed, and was only achieved with an enhanced, more sensitive version, during reanalysis. The *post hoc* removal of misclassified patients yielded better results, and maraviroc established non-inferiority to efavirenz under the condition of treating R5 virus patients only. The ineffectiveness of maraviroc on patients with X4 virus has been described in an earlier phase II placebo-controlled study where the efficacy of maraviroc on treatment experienced patients with X4 virus and a viral load of $>$5000 copies/mL was examined [93]. This study showed no significant difference changes in viral load between the maraviroc and placebo group. Taken together these results underline the importance of a sensitive and reliable detection of X4 viruses, prior to administrating co-receptor antagonists, like maraviroc to treatment experienced patients.

1.4.4 Phenotypic Testing of Co-receptor Tropism

Phenotypic test have long been the standard way of determining co-receptor tropism. Most cell-based assays, like Phenoscript®(VIRalliance) [120], PhenX-R®(inPheno) [49], Trofile®[160] and the Enhanced Sensitivity Trofile®Assay (ESTA) [149] (both Monogram Biosciences), rely on expressing patient-derived *env* genes in combination with replication-defective HIV virions. Over the last few years, the Trofile assay has established itself as the gold-standard in phenotypic HIV tropism testing and has been used for the characterization and clinical development of HIV-1 co-receptor antagonists [96, 47, 46, 44]. In the Trofile assay virus *env* genes are extracted from patient plasma, amplified by Reverse transcription-PCR and inserted into expression vectors. A replication-defective vector, carrying HIV genes of the NL4-3 strain of HIV-1, where the *env* gene was substituted for a luciferase cassette, is co-transfected with the patient-derived *env* library into human embryonic kidney 293 (HEK293) cells. Viral stocks of pseudo-virions are then harvested and used to infect cells expressing CD4 and either CCR5 or CXCR4. The successful infection of the respective cell lines is measured by luciferase activity and thus allows determination of co-receptor tropism. Possible outcomes are the detection of R5 or X4 viruses, that exclusively infect cells via CCR5 or via CXCR4, respectively, and of dual-tropic or virus mixtures (DM). The standard Trofile assay has recently been improved and was shown to be effective in detecting very small concentrations of X4 viruses. Reanalysis of patients enrolled in the MERIT study has demonstrated the improved performance and clinical value of the enhanced Trofile. However, phenotypic tests have important drawbacks: e.g. Trofile is expensive (1800 USD in 2008 [90]), has long turnaround times of about 16 days [118], is limited to specialized laboratories [54], and fails to produce results for viral loads of <1000 copies/mL (15% of cases [114]). Currently, German health insurance does not cover costs for phenotypic tropism testing [54]. Differences in sensitivity and outcome of various phenotypic tropism test have been reported repeatedly [148, 158].

1.4.5 Genotypic Testing of Co-receptor Tropism

The importance of accurate methods for determining the HIV co-receptor tropism has already been discussed. Phenotypic, cell-based assays, like Trofile have been applied in various clinical studies (see Section 1.4.4) and, with the recently improved sensitivity, are considered a useful indicator for the effectiveness of co-receptor antagonists. However, the potential progression of co-receptor antagonists towards first-line therapy, there is a growing need for cheaper and faster methods. Early on, it has been noted, that the V3 loop of gp120 is the main determinant of co-receptor tropism [61]. Simple rules like the 11/25 rule are capable of predicting tropism by evaluating the sidechain charge of positions 11 and 25 within the V3 loop sequence with fair success [39, 133]. In the past years different groups have devised a variety of mostly machine-learning-based methods to predict co-receptor tropism, considerably improving prediction accuracy [117, 111, 65, 136]. These methods rely on a set of V3 loop sequences with known associated co-receptor tropism, from which tropism-specific sequence features can be extracted. The generated models can then be used to predict the tropism of new V3-loop sequences, by evaluating the derived features. Different machine learning methods have been applied, like artificial neural networks [117], support vector machines [111, 136], or position-specific scoring matrices [65]. While most of these methods rely exclusively on V3 loop sequence information, it has been shown that additional information, like clinical parameters (CD4 cell counts, viral load) [136] or V3 loop structure models [126] can further improve prediction accuracy. Several groups have made their methods publicly available by offering web-services (see Table 1.2). While all of the services are free-of-charge and at least in principle able to generate predictions within seconds, there are severe differences in terms of usability. Here, wetcat serves as a negative example, requiring an initial alignment to be generated manually (accessed December 2010).

Table 1.2: **Publicly available genotypic co-receptor tropism prediction methods**

Name	URL	Method
Web-PSSM	http://indra.mullins.microbiol.washington.edu/webpssm/	PSSM
geno2pheno	http://coreceptor.bioinf.mpi-inf.mpg.de/	SVM
wetcat	http://genomiac2.ucsd.edu:8080/wetcat/v3.html	SVM

1.5 Research Motivation

In the past years there have been a number of publications pointing out that the performance of current genotypic prediction methods, compared to the results of state-of-the-art phenotypic tests was still inadequate. With the development and introduction of co-receptor antagonists like maraviroc and vicriviroc, there was, and still is, a need for reliable determination of co-receptor tropism. The possible introduction of maraviroc or other entry inhibitors as first-line drugs, also calls for cheaper, faster, and broadly available methods. The motivation for this work was to identify ways to improve genotypic co-receptor tropism prediction. This can be achieved for one, by developing a method with an improved prediction accuracy on data of the current clinical practice and will be elucidated in Chapter 2, or by application to more complex and reliable data, as will be the topic of Chapter 3.

2 Prediction of HIV-1 Co-receptor Tropism from Genotype

«No matter how good you are, you're going to lose one-third of your games. No matter how bad you are you're going to win one-third of your games. It's the other third that makes the difference.»

– Tommy Lasorda

2.1 Introduction

Computational models for predicting drug resistances from virus protein sequences have already become a standard procedure in HIV antiretroviral therapy and are explicitly included in the current German-Austrian Guidelines for ART in HIV-1 infections [27]. When the FDA approved maraviroc in 2007, computational methods for predicting patients co-receptor tropism were already available [117, 111, 65, 136, 126], but various studies found rather unsatisfying discordance between predictions by these genotypic methods and results obtained by state-of-the-art phenotypic assays [87, 139, 88]. Most of these applied machine-learning methods generate models able to distinguish R5 from X4 V3 loop sequences. Motivated by promising results of previous machine-learning-based prediction methods, the aim of this work, described in this chapter, was to explore ways of improving co-receptor tropism predictions. In the following, a new co-receptor tropism prediction model, relying on two different descriptions of V3 loop sequences will be presented. The model was trained using suitable sequences selected from the vast collection of the Los Alamos HIV Sequence Database and tested on an independent set of clinical sequences gathered in various laboratories in Germany, comparing it to current state-of-the-art methods.

2.1.1 Machine Learning in Biology

Machine learning, very generally, describes a field of artificial intelligence dealing with algorithms for pattern recognition, classification and prediction from models which were derived using existing data [146]. Machine learning algorithms are used to extract characteristic features from concrete example data to generate models which can in turn be applied to classify, recognize, or interpolate new data. The data consists of a set of observations (data points), where each observation contains a set features, or variables. In a biological context, observations consisting of features could be protein or DNA sequences consisting of aminoacids or nucleotides. The features depend on the description of the observation. So-called descriptors are used to created different sets of features. A protein, for instance, could be described as a sequence of aminoacids, secondary structure elements, or simply by properties such as volume, charge or mass. Some descriptors are more powerful than others, because they capture more details, or more important details. The field of machine learning is divided into *supervised* and *unsupervised* learning. In unsupervised learning, observations are grouped, or clustered, based on similarity. A prominent example is the discovery of the archae by Carl Woese and George Fox in 1977 [163]. Using the sequences of the 16s ribosomal RNA from a variety of microorganisms, Woese and his colleague constructed a similarity matix, measuring the pairwise distances between all sequences. Surprisingly, next to clusters representing the known groups of prokaryotes and eukaryotes, a third cluster became evident – the archae. Unsupervised learning methods like clustering are popular in phylogenetic or microarray analyses. In supervised learning observations are linked to classes. Machine learning methods will try to generate classifiers which relate the classes associated with the observations, based on rules generated from characteristic features found. The trained model, or classifier, can then be used to classify new observations. Popular classification problems in biology are secondary structure prediction from protein sequences or prediction of drug resistance from patient data.

2.1.2 Random Forests

Ensemble classifiers are sets of classifiers, applied to the same classification problem. The final step is the joint prediction of all single classifiers through fusion. The variety of fusion methods is vast: majority vote [15], minimum or maximum prediction, multiplication, or even the use of another machine-learning model [165]. The single classifiers can also result from different machine learning approaches, e.g. artificial neural networks, support vector machines or k-nearest neighbors algorithms. Optimally, these methods use different descriptors, trying to separate the observations in different feature spaces [53]. Nanni and Lumini applied classifier ensembles for HIV-1 protease cleavage prediction [100]. They used linear support vector machines trained on different feature subsets and fused them by summation. It has been shown in theoretical and empirical studies that combining classifiers trained using different descriptors can improve class prediction accuracy compared to a single classifier with combined feature space [169, 123, 32]. On the other hand, it may be unwise to combine too many classifiers as this can lead to suboptimal prediction performance [71, 123]. The descriptor selection can be crucial for the prediction performance [106]. Descriptor selection approaches, such as genetic algorithms or clustering, have been proposed to find optimal descriptor sets [99, 134, 72, 68].

The random forest algorithm was created by Leo Breiman in 2001 [10]. Random forests are ensemble classifiers consisting of decision trees. Each tree is unique and has been constructed, using a randomly selected subset of features found in randomly sampled subsets of the training data. When classifying new observations of unknown class, the final prediction is based on the majority vote of the single decision trees. Random forests have been shown to be very accurate and fast [10]. Notably, predictive power has been shown to converge with a growing number of generated trees, eliminating one possible source of overfitting [10]. In contrast to other machine learning algorithms, like neural networks or support vector machines, random forest estimates error rate and feature importance during construction. Decision trees and their construction comprises a number of concepts that are explained in the following.

Decision Trees

Schematically, a decision tree is a set of subsequent branching operations, based on the evaluation of predefined tests. A decision tree consist of three key elements (see Figure 2.1.2):

1. Nodes (Boxes): Test
2. Leaves (Circles): Decision
3. Branches (Lines): Consequence of test. Connecting nodes with leaves or nodes.

Figure 2.1.2 shows a binary decision tree. In a binary decision tree each inner node is only connected via three branches, every test having exactly two outcomes. The decision process always starts at the topmost node (here, Test I). When the first test evaluates to false, the corresponding branch leads to leaf with a negative decision. The overall result of the decision tree is 'No'. If the first test evaluates to true, the corresponding branch leads to a second node, representing an additional test. The result of this second test will lead to a final decision, as both resulting branches lead to a leaf.

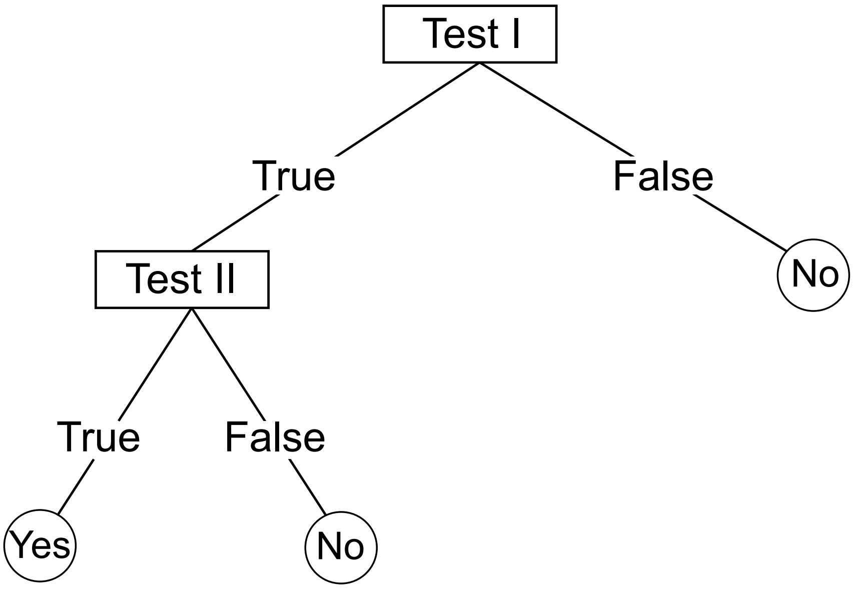

Figure 2.1: **Example of a binary decision tree.**

Binary decision trees are simple examples of trees. More complex trees can accommodate more than two test results. The decision tree depth describes the longest path, in units of branches, through the tree until reaching a decision. A practical example is given in the work of Beerenwinkel *et al.* [7] where decisions on drug resistance or susceptibility are made, based on the evaluation of the type aminoacids at different protein sequence positions. Figure 2.1.2 shows two trees taken from that work. The trees were grown to predict resistance or susceptibility for delavirdine (h) and efavirenz (i), two drugs targeting the reverse transcriptase of HIV-1. In this example the components nodes, branches and leafs represent:

1. Nodes: Evaluation of aminoacid at given sequence position
2. Leaf: Decision on susceptibility (green) or resistance (red)
3. Branch: Consequence of aminoacid found a given position

To predict the drug susceptibility or resistance of delavirdine for a new reverse transcriptase sequence, one again starts at the topmost vertex of the tree (h) and evaluates which aminoacid is present at position 103. Possible outcomes are decisions on susceptibility (L) or resistance (N,Q, or T), and a further test of the aminoacid at either positions 122 (E), 181 (K), 179 (R), or 211 (S). In that way, one proceeds down the tree until either a decision (leaf) is reached, or a test fails to produce a results. The latter is the case if the aminoacid found a given position during testing is not associated with a branch, and thus, no prediction can be made.

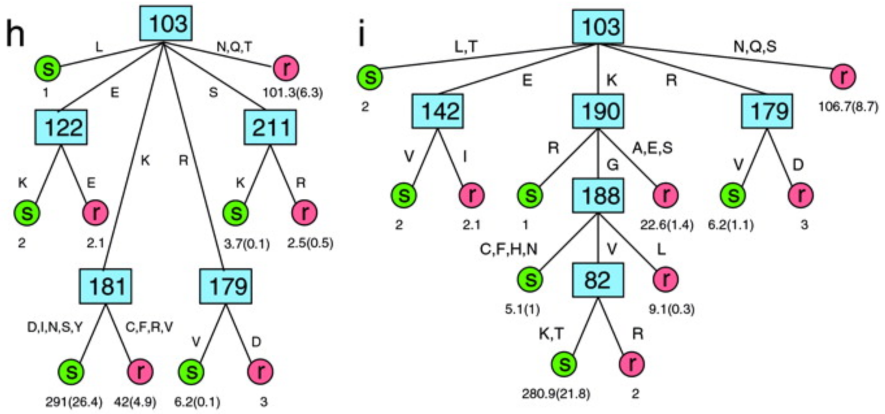

Figure 2.2: **Decision tree predicting drug resistance against RT inhibitors.** Decision tree taken from Beerenwinkel *et al* [7]. Here a decision on drug resistance is made, based on the aminoacids present at defined sequence positions. The trees predicts the resistance or susceptibility to delavirdine (left, h) and efavirenz (right, i).

Bootstrapping and Bootstrap Aggregating (Bagging)

In statistics, bootstrapping is a resampling method. It is used to approximate estimators, such as average or variance, of an existing dataset by determining the distribution of the estimator over a number of datasets resampled from the original. For generating resampled sets a random sampling with replacement is used. In machine learning, bootstrap aggregation (short: bagging), is closely related to bootstrapping and is used to improve prediction or regression ensemble classifier models. An ensemble of classifiers is created by aggregating multiple bootstrapped samples (in-bag data) of the training data and constructing one decision tree from each sample. During bagging the in-bag data consists of 63% of the training data [9]. For each sample, the data not used for training is referred to as out-of-bag (OOB) data and used for evaluation of performance, error, etc. Results of classification, regression or error estimation are averaged over the complete ensemble. By bagging, models become more stable in terms of prediction and overfitting is usually avoided.

Approximation Using OOB Data

The random forest algorithm performs an internal crossvalidation during training. The goal of crossvalidation is to estimate the performance of the model on a data set that is independent to that used for model training. For that purpose, the OOB data of every tree is used to calculate model estimators like error-rate, prediction performance, or variable (feature) importance. The model performance for each observation n in the training data is estimated by classifying n by each tree where it was out-of-bag, thus not used for tree construction. The final prediction for n is fraction of votes for one of the classes. By the same manner, the error rate of the forest can be estimated. The error estimate is then given by:

$$E_{forest} = \tfrac{1}{N} \sum_{n=1}^{N} E_{tree}(p(n), y_n)$$

with

$$E_{tree}(p, y) = \begin{cases} 0, & \text{if } p = y \\ 1, & \text{if } p \neq y \end{cases}$$

where N is the number of observations, $p(n)$ is the class predicted for observation n by majority vote of all models where n was OOB, and y_n is the true class of n.

Importance Analysis

Machine learning methods are often referred to as 'black boxes'. After model training it is complicated to understand what the models have learned. Random forests allow an in-depth view into the learned classification mechanism by examining single decision trees. A faster and more elegant approach is, however, directly implemented in the random forest algorithm: importance analysis. The importance of every variable for the successful classification is estimated by comparing the number of correct classifications of OOB data of each tree with those of a set of permuted OOB data. For every tree that is constructed, each variable of the OOB data is permuted and classified by the respective tree. The difference between the classification error of the original and permuted OOB data is recorded and averaged over all trees and normalized by the standard deviation of the differences. The importance of each variable is the estimated percent decrease in prediction accuracy [82].

Growing Decision Trees

Each decision tree as single classifier for the random forest is constructed as follows:

1. From N training data, a bootstrap sample of size N is generated.
2. If the N training data consists of $N \cdot M$ variables, choose a number $m \ll M$, representing the number of variable used at each tree node.
3. At each tree node, randomly choose m variables to base the test on.
4. Choose the best test to split the training data into the respective classes.
5. Repeat 3 and 4 until tree is fully grown.
6. Evaluate estimators (error rate, predictive performance, variable importance) using the OOB data.

Classifying New Data

Once a random forest has been trained, the model can be applied to new data for classification or regression, depending on the type of model. When classifying new data, the resulting prediction for an observation is a pseudo probability of belonging to a class. It is the arithmetic mean of the single tree decisions:

$$p = \tfrac{1}{N}\sum_{i=1}^{N} p_i$$

where N is the number of trees in the ensemble and p_i is the decision of a single tree. For this work, the random forest algorithm was chosen as the machine learning method of choice.

2.1.3 Structural Descriptor: The Electrostatic Potential

The 11/25 and 11/24/25 rules, as well as the V3 charge rule imply the importance of electrostatic interactions of V3 and the co-receptors CCR5 or CXCR4 for discriminating between R5 and X4 viruses. In order to analyze whether there exists a difference in the electrostatic potential of the co-receptors, homology models of both CCR5 and CXCR4 were kindly built by Dr. Steffen Wolf from the Biophysics Group at the Ruhr University in Bochum. The template selection and modeling process are described in detail in the Materials and Methods Section (2.2). The electrostatic potential at the solvent-accessible surface of the extracellular regions of co-receptors CCR5 and CXCR4 are shown in Figure 2.3. It can be seen that the electrostatic potential at this region is quite distinct between the models. While CCR5 (left) has a largely positive potential, shown in blue, the potential of CXCR4 (right) is found to be more negative, shown in red. This is accordance with the 11/25 and 11/24/25 rules, stating that X4 loops often have a positively charged aminoacid at positions 11 or 25.

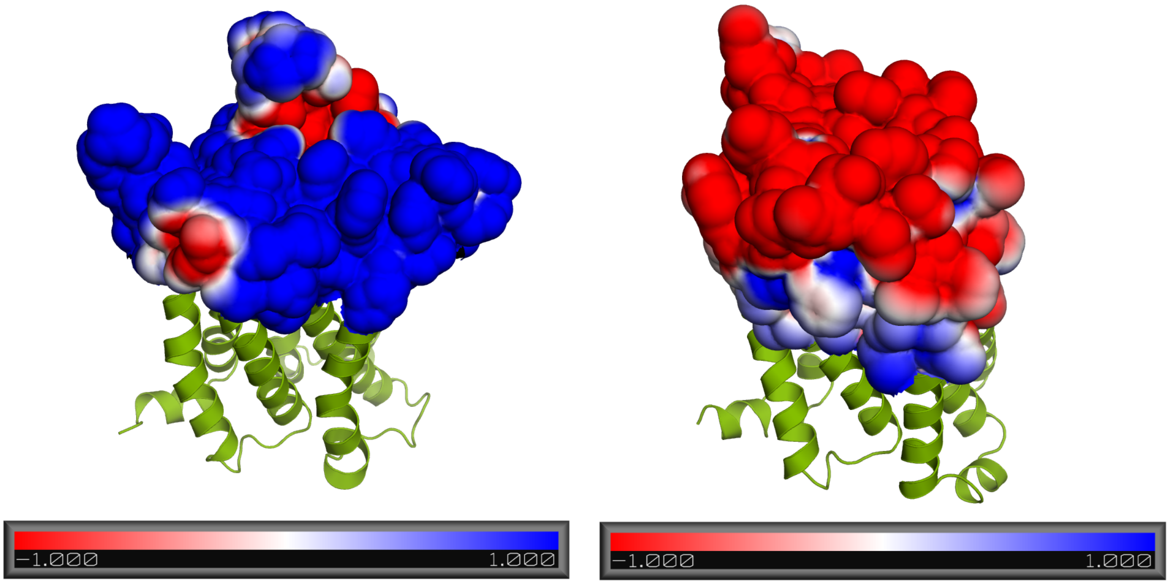

Figure 2.3: **Electrostatic potential of the extracellular region of Chemokine co-receptors CCR5 and CXCR4.** The computed electrostatic potential of the extracellular region of CCR5 (left) is largely positive as shown by the blue surface. In comparison the extracellular electrostatic potential of CXCR4 (right) is overall more negative. Unit of electrostatic potential in $k_B \cdot 300K/e$. Details on the computation are given in Section 2.2

The rule-based approaches of 11/25 and 11/24/25 are, however, biased by the underlying sequence alignment. Prior to prediction of co-receptor usage, most methods are aligning the target sequence to be predicted to a reference V3 loop sequence. In the case of the 11/25 and 11/24/25 rules the HXB2-V3 loop sequence is used as reference. The sequence numbering is relative to the reference and residues aligned to reference positions 11, 24, and 25 are thus dependent on the alignment procedure. The requirement of a positive charge at one of the alignment positions does not mean that a positive net charge is required for the natural binding mode of an X4 virus to the CXCR4 co-receptor. For instance, one could imagine a compensatory mutation at a different sequence position, neutralizing the net charge. This is in part covered by the charge rule, which considers the sum of all charges within the V3 loop and discriminates X4 from R5 viruses by the V3 net charge. However, while gaining information on possible compensatory charges at positions not included in the 11/25 or 11/24/25 rules, information on the sequence position of charges is lost. We hypothesize that it is more feasible to consider an aminoacid within the spatial ensemble of neighboring aminoacids, than judging it by its position within an alignment.

This spatial effect of charges can be measured by the electrostatic potential, which is directly related to the charge density as stated by Poisson's equation, and can be approximated by numerically solving the Poisson-Boltzmann equation (PBE). The electrostatic potential describes the interaction between atoms A and B as electrostatic interaction energy:

$$E_{elec} = \phi_A(\vec{r}_B) \cdot q_B$$

where $\phi_A(\vec{r}_B)$ is the electrostatic potential of atom A at position $\vec{r}_B$ of the charge q_B of atom B.

Using the electrostatic potential of the V3 loop to discriminate between R5 and X4 viruses, combines information of charges and spatial orientation by direct relation. Furthermore, the electrostatic potential already considers effects of local interaction of charges. A handful of computer programs are capable of numerically solving the Poisson-Boltzmann equation, which relates the electrostatic potential ϕ, the charge density ρ and the dielectric constant ε at position $\vec{r}$ as follows:

Poisson's equation directly relates the electrostatic potential ϕ to the charge density ρ, where ε_0 is the charge permittivity in vacuum:

$$\nabla^2\phi(\vec{r}) = -\tfrac{\rho(\vec{r})}{\varepsilon_0}$$

Here, ∇^2 combines the gradient (∇) and divergence operators ($\nabla\cdot$) and can be written as

$$\nabla \cdot (\varepsilon(\vec{r})\nabla\phi(\vec{r})) = -\tfrac{\rho(\vec{r})}{\varepsilon_0}$$

where the gradient operator (innermost ∇) represents the direction of propagation of the potential, damped by the permittivity of the medium $\varepsilon(\vec{r})$. The divergence operator (outer $\nabla\cdot$) determines whether the electric field is inbound or outbound.

Poisson's equation can thus be rewritten as

$$-\varepsilon_0 \nabla \cdot (\varepsilon(\vec{r}) \nabla \phi(\vec{r})) = \rho(\vec{r})$$

Poisson's equation only considers shielding of charges by dielectric constants (ε). To account for mobile ions in solution the Poisson-Boltzmann equation adds monovalent mobile ions (e_-/e_+) to determine charge density:

$$-\varepsilon_0 \nabla \cdot (\varepsilon(\vec{r}) \nabla \phi(\vec{r})) = \rho(\vec{r}) + ne_+ - ne_-$$

Based on the theory that positive ions are more probable to reside in regions of negative potential and vice versa, shielding of charges is modeled by a Boltzmann distribution of ions based on the local electrostatic potential

$$n_\pm(\vec{r}) = n_\infty e^{\frac{\pm e\phi(\vec{r})}{k_B T}}$$

where $n_\pm(\vec{r})$ is the ion number density at position $\vec{r}$ and n_∞ is the ion number density where $\phi(\vec{r}) = 0$. Finally Poisson's equation becomes the Poisson-Boltzmann equation

$$-\varepsilon_0 \nabla \cdot (\varepsilon(\vec{r}) \nabla \phi(\vec{r})) = \rho(\vec{r}) - 2en_\infty \sinh(\frac{e(\phi(\vec{r}))}{k_B T})$$

The electrostatic potential ϕ can be calculated for any given coordinate $\vec{r}$. Here, we are especially interested in a set of ϕ around different V3 loops, to serve as input for our machine learning efforts.

Rather than deriving a rule-based prediction from the electrostatic potential data, we applied machine learning to the problem. Machine learning algorithms are capable of learning complex interactions, like correlations between different variables, or here: the electrostatic potential at different coordinates and the property of being X4- or R5-tropic. This model can then be used to classify new V3 loop sequences with unknown tropism. The data vectors, describing the loop, used for training of a model, as well as future classification have to be of constant length and format. In the case of the electrostatic potential of V3 loops, which is a continuous quantity, the challenge was to define a representative set of ϕ-values. Two requirements had to be fulfilled. One, the set of values had to explain differences in tropism. Two, they had to be available for every existing V3 loop. For that case we devised the concept of the electrostatic hull. The electrostatic hull represents a set of N points in space surrounding the V3 loop structure. The common feature of the points is their equal minimum distance from the structures solvent-accessible surface (SAS). At each point a discrete value for the electrostatic potential ϕ can be calculated. The hull fulfills one important criterion, namely that the same set of points can be derived from every V3 loop structure. This is ensured by choosing the hull wide enough so that every atom of every natural V3 loop structure can be comfortably embedded (Figure 2.4).

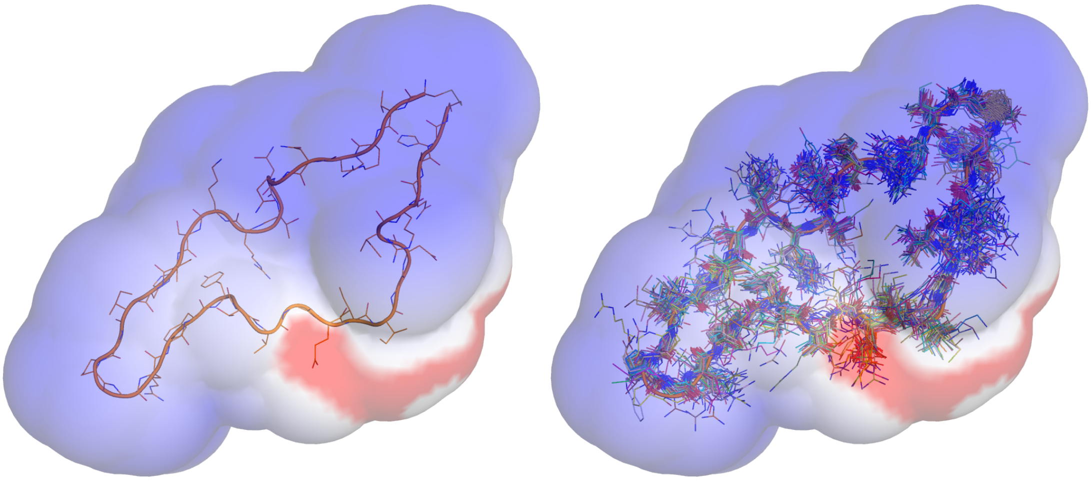

Figure 2.4: **The electrostatic hull.** The electrostatic hull consists of a set of points with an equal distance to the solvent-accessible surface of the V3 loop. At each point a discrete value for the electrostatic potential ϕ has been calculated. The hull is wide enough to embed every modeled V3 loop structure. Left: Hull surrounding the template V3 loop used for modeling of V3 sequences with unknown structure (PDB: 2B4C). Right: Electrostatic hull embedding various V3 loop model structures. No atom penetrates the hull.

The fact that the hull is merely a set of coordinates above the V3 loop surface, enables one to specifically determine the electrostatic potential at these coordinates for every modeled loop structure. This is done by superimposing the modeled structures to the template structure from which the hull was defined earlier. This results in a close to identical relative orientation of the coordinates of hull and V3 model backbone. The different electrostatic potential calculated at the hull, caused by the distinct side chain composition of V3 loop models, can then be used to learn the discriminating features of the two classes, R5 and X4, by means of machine learning. Because of the combination of spatial information from homology models and the electrostatic potential at a set of coordinates, the electrostatic hull is a structural descriptor.

2.1.4 Sequence Descriptor: The Hydropathy Scale

In previous attempts to predict co-receptor usage, various groups have had considerable success by using sequence information only, applying different machine learning or scoring methods [117, 65, 111]. To our knowledge random forests have not been used for co-receptor usage prediction based on V3 loop sequence as a descriptor. In addition to the structural, electrostatic potential descriptor, we were eager to test the performance when using a more basic descriptor, based solely on the V3 loop sequence. Machine learning algorithms often work with sequences of numerical values, thus, protein sequences have to be translated prior to model training. A simple translation scheme would be the use of one unique value for every aminoacid. For most cases, one could imagine 20 values, representing the 20 crucial aminoacids. Similarity between properties is modeled by the numerical difference of values. Thus, properties coded by values with small differences are more similar to each other than those with larger differences. Using the previously suggested coding scheme of unique values based on the aminoacid name will quickly produce problems. Performance of trained models would vary considerably depending on chosen pairs of $aminoacid \leftrightarrow n$. This problem can be solved by relating not the aminoacids per se to numerical values, but rather coding their physicochemical properties. There may be no similarity between lysine and arginine considering their names, but both carry positively charged sidechains. The positive net charge makes both aminoacids similar by the charge property. The same can be said for aspartate and glutamate, both side chains are negatively charged. When assigning numerical values to these aminoacids, the sum of numerical differences within each group should be minimized, whereas the difference between the groups should be maximized. This is only a simple example of how aminoacids can be reasonably coded into numerical values, so called numerical descriptors. In fact, a huge number of descriptor sets have already been published [66], based on properties such as molecular weight, charge, solubility, secondary structure preferences, or hydropathy. The hydropathy scale by Kyte and Doolittle [76] assigns a numerical value to each aminoacid, reflecting its hydrophilic or hydrophobic character, as determined by the authors. In general, more hydrophilic aminoacids are assigned more negative values, while more positive values are assigned to hydrophobic aminoacids. This scale has been successfully applied in protein classification problems. Heider *et al.* have been able to predict small GTPases and the susceptibility to the HIV maturation inhibitor berivimat, based on the aminoacids sequences translated using the hydropathy index [50, 52]. Additionally to the translation of aminoacids sequence into the Kyte-Doolittle hydropathy index, we used a normalization procedure to ensure a constant length of translated sequences, a prerequisite of the random forest algorithm. Alternatively, sequences could be aligned to a reference sequence, however, the problems arising from sequence alignments have already been discussed (see Section 2.1.3). The normalization procedure is described in detail in the materials and methods section. The normalized V3 loop sequences are referred to as sequence descriptors. The resulting machine-learning model will be referred to as the hydrophobicity model.

2.2 Materials and Methods

V3 sequence training data For the training of the tropism prediction models, V3 loop sequences with experimentally determined co-receptor usage were taken from the Los Alamos HIV Sequence Database (`http://www.hiv.lanl.gov/`). The web site search interface allows filtering based on co-receptor usage, thus, three searches were performed: one for each of the classes R5, X4, and R5X4. Additional filtering criteria offered by the search interface were used as follows. Virus type was set to HIV-1, while no subtypes were excluded. V3 was set as the genomic region. Problematic sequences, classified as those by the HIV Sequence Database were excluded. For the R5X4 class, only sequences of clonal nature were included, thus removing sequences with a dual/mixed tropism annotation. The tropism of a sequence was considered doubtful if the same sequence occurred in more than one tropism class. As a consequence, these sequences were removed from all classes. Sequences occurring more than once in the same class, were reduced to one single copy. Sequences shorter than 30 aminoacids were removed to reduce the noise in the dataset. The longest sequences consisted of 38 aminoacids. Furthermore, sequences with obvious errors, such as frameshifts or ambiguous residues were also excluded from the set. After application of the stated filtering criteria, the training set consisted of 1151 R5 sequences, 166 X4 sequences and 34 R5X4 sequences. 284 of these sequences contained insertions or deletions (17% of R5, 51% of X4, 10% of R5X4). We published the datasets were published as supplementary material with the here and can be accessed at `http://dx.doi.org/10.1371/journal.pcbi.1000743`.

Clinical Sequences Different sets of V3 loop nucleotide sequences were kindly provided by Saleta Sierra-Aragon and Rolf Kaiser from the Institute for Virology in Cologne, Germany, as well as Alexander Thielen from the Max Planck Institute for Informatics in Saarbrücken. Additional V3 loop sequences were collected from the literature [43, 67, 98, 58]. From this collective set, sequences containing ambiguous nucleotides were removed and the remaining sequences were translated into aminoacid code and subsequently filtered, using the same criteria as for the V3 sequence training data, described above. The final set of clinical sequences consisted of 381 sequences in total, with 132 X4 and 249 R5 sequences.

Homology modeling of V3 loops Homology modeling of all V3 loop sequences was done using the Modeller software [125], version 9.6. The V3 loop of the X-ray structure of gp120 bound to CD4 and an antibody (PDB: 2B4C) by Huang *et al.* [57] was used as template. Each sequence to be modeled was first pairwisely aligned to the template sequence. The initial model was then built allowing only for optimization at regions of insertions or deletions. In the following optimization step sidechain positions and rotations were energy minimized to resolve possible sterical problems. The root mean square deviation over all 1351 V3 models used for training was 0.85 Å with a standard deviation of 0.18 Å.

Homology Modeling and Electrostatic Potential of Co-receptors Models of both, CCR5 and CXCR4, were built using the crystal structure of bovine rhodopsin (PDB: 1GZM), chain A [80], identified by PSI-Blast [2]. Initial sequence alignments were taken from a PSI-Blast and manually corrected to accommodate transmembrane areas. Ligands and water molecules were removed from the template. During the modeling process, post-translational modifications, like sulfated tyrosines were not considered. The following steps were equally applied to the modeling of both co-receptors. Sidechain replacement, according to the alignments generated was performed by SCWRL [14]. Sterical clashes were removed with MOBY [55]. Loop regions were modified by manually resolving gaps or inserts through addition or removal of aminoacids. Next, modified regions were subject to simulated annealing (5 ps) and energy minimization using MOBY. Finally, the complete models were fully energy minimized. The modeling protocol can be found in Wolf *et al.* 2008 [164]. The electrostatic potential of the co-receptor models were calculated using APBS version 1.0 [5] with default settings.

Electrostatics hull Calculation of the electrostatic potential around a given V3 loop structural model was performed by solving the Poisson-Boltzmann equation with the Adaptive Poisson-Boltzmann Solver – APBS (Version 1.0.0) [5]. Charges and radii of V3 loop atoms were assigned by PDB2PQR (Version 1.3.0) [28], according to the AMBER forcefield specifications. Parameters of the Poisson-Boltzmann equation were used as follows: ionic strength was set to zero, dielectric constants of protein and solvent were altered to optimize prediction accuracy resulting in equal values of $\epsilon_{protein} = \epsilon_{solvent} = 5.0$. A temperature of 310 K was used and a single Debye-Hückel sphere was considered as boundary condition to solve the non-linear Poisson-Boltzmann equation. A cubic grid consisting of 33^3 grid points, with a grid spacing of 3 Å was centered over each V3 loop structure. For the construction of the electrostatic hull, the solvent-accessible surface of the template V3 loop structure was used. In addition to the grid containing the electrostatic potential, APBS produces a second grid with information on the protein solvent accessible surface. This surface is approximated using a solvent probe of $r = 1.4$ Å. Both grids are congruent and thus allow conclusions on the electrostatic potential on the SAS. In order to ensure complete embedding of every possible V3 loop structure in the electrostatic hull, the calculated surface was extended further into the solvent. The amount of extension was varied by $n = 1, 2, 3$ times the grid spacing distance, resulting in three hulls, equidistant from the V3 loop SAS at 3 Å, 6 Å, and 9 Å. All hulls were capable of completely embedding V3 structure models, but the hull at 6 Å yielded the best results.

Random Forests Random forest analyses were performed with the package randomForest [82] of R [115]. The out-of-bag error was used to estimate errors; a ten-fold external crossvalidation performed to test the robustness of error estimation yielded equal results. For ROC analyses the ROCr package [137] of R was used. True positive rates of the form $a \pm \delta$ given in the text are averages a over 10 random forest trainings with $\pm\delta$ marking a 95% confidence interval estimated with a t-distribution.

Patient-wise Crossvalidation Scheme Sequences were first grouped by the internal patient identification numbers, so that each group contained every sequence of one single patient. Model performance was then evaluated by a leave-one-patient-out crossvalidation scheme. One round of model training was performed per patient, using every sequence not associated with the specific patient. Tropism of the sequences left out were subsequently predicted and prediction results were used for evaluation.

Normalization Procedure In order to cope with the varying sequence length and provide a fixed length of input data for the machine learning methods applied, we used a length normalization procedure, previously described by Heider *et al.* [50, 51]. To normalize a signal (i.e. a sequence of descriptor values) of length N to a new target length M, one has to first interpolate the signal and then choose M equally distanced values from the interpolated signal. Any common method like simple linear, cubic spline, periodic spline, or natural spline interpolation can be used. The linear interpolation method has been shown to yield the best results in terms of predictive performance for various protein classification problems (Heider *et al.*, unpublished) and is exclusively used throughout this work. For the linear interpolation, the y-value for a given x-value in the interval $[x_n, x_{n+1}]$ is given by

$$y = y_n + (x - x_n)\tfrac{y_{n+1}-y_n}{x_{n+1}-x_n}.$$

Here the x-value and y-values describe the newly calculated sequence position and feature value, respectively. The signal of target length M can now be acquired by calculating and concatenating the y-values of M equidistant x-values along the interpolated signal from x_0 to x_N, with $\frac{[x_m, x_{m+1}]}{[x_n, x_{n+1}]} = \frac{M}{N}$.

Model Performance Analysis Performance measures of sensitivity, specificity, accuracy, true positive rate (TPR), and false positive rate (FPR) were calculated by

$$\text{TPR} = \text{sensitivity} = \tfrac{\text{TP}}{\text{TP+FN}}$$

$$\text{specificity} = 1 - \text{FPR} = \tfrac{\text{TN}}{\text{TN+FP}}$$

$$\text{accuracy} = \tfrac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$$

with TP: true positives, FP: false positives, FN: false negatives and TN: true negatives. Receiver Operating Characteristics (ROC) curves and the corresponding AUCs were calculated with ROCr [137]. All performance values of the form $a \pm \delta$ given in the text are averages over ten random forest trainings with $\pm\delta$ marking a 95% confidence interval estimated with a t-distribution. Crossvalidation schemes applied during training varies and is mentioned for each case.

Comparisons With Other Methods The performance of the 11/25 and 11/24/25 rules were calculated manually after alignment to the HXB2 reference V3 loop sequence. Predictions for tropism prediction by geno2pheno and PSSM the respective web-interfaces were used (see Table 1.2) with standard settings.

2.3 Results

2.3.1 Prediction of Tropism Based on the Electrostatic Potential

To train a machine learning model based on the electrostatic potential information of V3 loops, V3 sequences with known tropism were acquired from the Los Alamos HIV

Sequence Database (http://www.hiv.lanl.gov/), as described in detail in the Materials and Methods section. After filtering, the remaining set contained 1351 unique sequences. The majority (85%) of these originated from Europe (37%), Africa (34%) and North America (14%). 55% of sequences are of subtype B, most common in Europe and North America, followed by subtypes C (17%) and D (10%) the common subtypes in Southern and Eastern/Central Africa, respectively (see Figure 2.5). The majority of European sequences have been collected in the Netherlands, Belgium, Sweden, and Switzerland. The set contains no sequences collected in Germany.

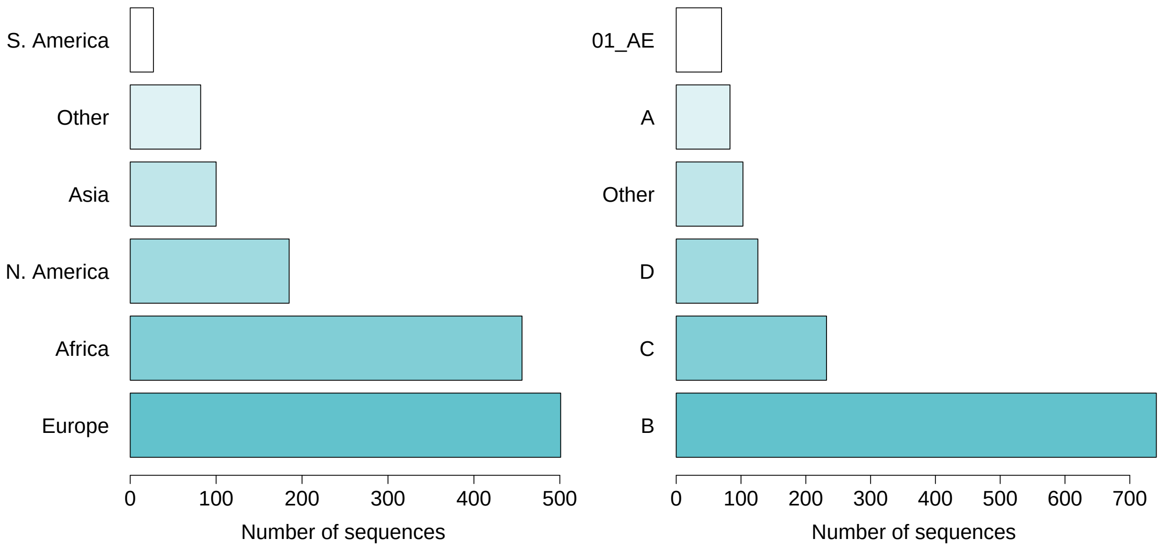

Figure 2.5: **Origin (left) and subtypes (right) of training sequences.**

All V3 loop sequences were subject to homology modeling based on the V3 crystal structure loop by Huang *et al.* [57]. The structural data on V3 is very limited, and from the experimentally determined structures, as well as in-house molecular dynamics simulations (data not shown) a great flexibility of the loop can be deduced. These facts combined leave no doubt as to the probable errors made by modeling over one thousand loops with varying sequences based on a single template. The central assumption justifying this approach, however, is that mapping of electrostatic potential values to one possible structure is sufficient to provide a basis for a separation of X4- and R5-tropic variants. The modeling procedure and minimization process is described in Materials and Methods.

The electrostatic hull was then created, based on the template structure. The solvent accessible surface was used as initial hull and extended into the solvent. Each extension step increased the distance of the hull surface from the solvent-accessible surface by 3 Å. To assess the impact of different hull volumes on the model performance multiple hulls were generated by varying the amount of extensions steps (see Figure 2.6). The best performance was achieved using a hull based on two extension steps, accounting for the electrostatic potential 6 Å above the solvent-accessible surface. In the following, the performance of hull sizes other than 6 Å are not shown.

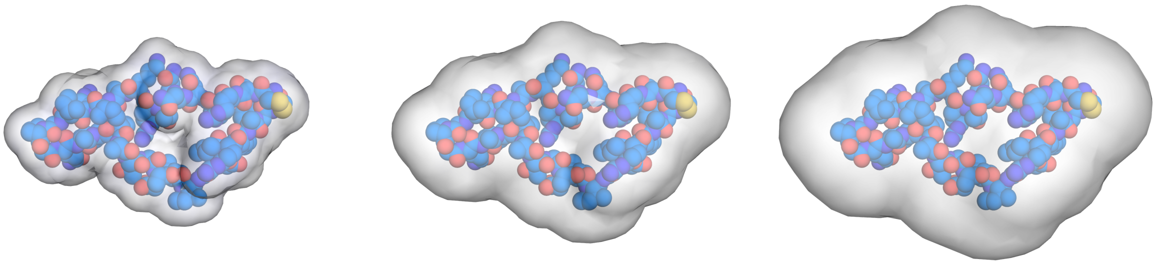

Figure 2.6: **The electrostatic hull with varying volume.** Here, the electrostatic hull is shown with varying extensions steps. The distance from the V3 loop solvent-accessible surface to the hull ranges from 3 Å, over 6 Å to 9 Å (from left to right).

The electrostatic hulls were used to construct a random forest model using the randomForest Package [82] for R. The training methods takes a table as input, relating a response variable (the tropism) to a set of feature variables (the electrostatic potential). The structure of training data table is shown in Table 2.1.

The clinical interest of co-receptor usage prediction is the identification of CXCR4-using viruses. In that regard, both X4 and R5X4 viruses belong to the same class. Consequently, all R5X4 viruses were labeled as X4 viruses. In total, 10 random forest models were built, each consisting of 500 decision trees. At each tree node $m = \lfloor\sqrt{642}\rfloor = 25$ variables were used for splitting. The OOB estimate of error averaged over all 10 models was 0.0472 ± 0.0002. The error rates for the respective classes were 0.0174 ± 0.0002 for R5 and 0.2198 ± 0.0007 for X4. Thus, the estimated error rate for the X4 sequences is over 10 times higher than that of the R5 rate.

The internal estimation of the error rate considers the results of the OOB data evaluated by the trees. A tree can only make a binary decision. A class, predicted by a decision tree as a single classifier, is either R5 or X4. The result of the complete forest as an ensemble classifier, is the fraction of votes for a specific class, thus a numeric value between 0 and 1. Usually, the class with the most votes, as in decisions by single trees, is accepted as the predicted class. In the case of two classes, a majority of at least 50% of the votes seems reasonable. A cutoff between the classes is drawn at 50% of votes,

Table 2.1: **Training data structure for the electrostatic potential model.** Each line holds the electrostatic potential values of the electrostatics hull in linearized form. In total 1351 V3 loop models were used, with 642 ϕ-values.

V3 Sequence	**Tropism**	ϕ_1	ϕ_2	ϕ_3	...	ϕ_{642}
$V3_1$	X4	-0.0019	12.7167	8.5402	...	15.7409
$V3_2$	X4	0.3707	17.1011	13.5197	...	12.6159
$V3_3$	X4	-0.0933	17.2641	10.3534	...	0.1495
...	...	...	...	...	...	...
$V3_{201}$	R5	0.0539	0.8321	0.4180	...	10.6525
$V3_{202}$	R5	0.6183	14.6741	8.5445	...	-0.0833
$V3_{203}$	R5	0.0307	0.7533	0.4082	...	1.5988
...	...	...	...	...	...	...
$V3_{1351}$	R5	0.1543	0.5784	0.0183	...	1.7093

or a fraction of 0.5; this cutoff can be altered. Hence, when setting the cutoff to above or below 0.5 one shifts the outcome in favor of one class, over the other. The cutoff has a clinical relevance, as it can be shifted to tune the prediction to be more conservative or more liberal, depending on the risk or value of treatment. Figure 2.7 shows the predictions made for the X4 sequences (orange) and the R5 sequences (blue) on the x-axis. The prediction value represents the fraction of X4 positive votes. The class predictions range from 0 (no X4 votes → R5) to 1 (all X4 votes → X4). Clearly, X4 sequences are clustering around high fractions of X4 votes. The opposite is true for the R5 sequences. The dotted line represents the standard cutoff of 0.5. With this cutoff a sequence with less than 50% of X4 votes is classified as an R5 virus. Notably, the X4 predictions of X4 sequences are more spread out than for the R5 sequences. This trend is visualized by the density plot on the top right of Figure 2.7. Interestingly, no X4 sequence has a prediction probability of 1.0. Thus, no X4 sequence was unanimously classified as such by all decision trees of the forests. In contrast, many R5 sequences have an X4-probability of zero, indicating recognition without errors.

Prediction accuracy, in this case, is not a fair measure as the accuracy measures the fraction of positive predictions over all sequences. In our sequence we have around six R5 sequences for every X4 sequence An accuracy of 83% can be achieved by simply setting the cutoff to 0, thus predicting every sequence to be R5. A better measure of performance is to observe the false positive rate against the true positive rate, varying the cutoff. Both estimators are rates, and do not take in account the absolute, but the relative number of predictions for the classes. Figure 2.8 shows a receiver operating

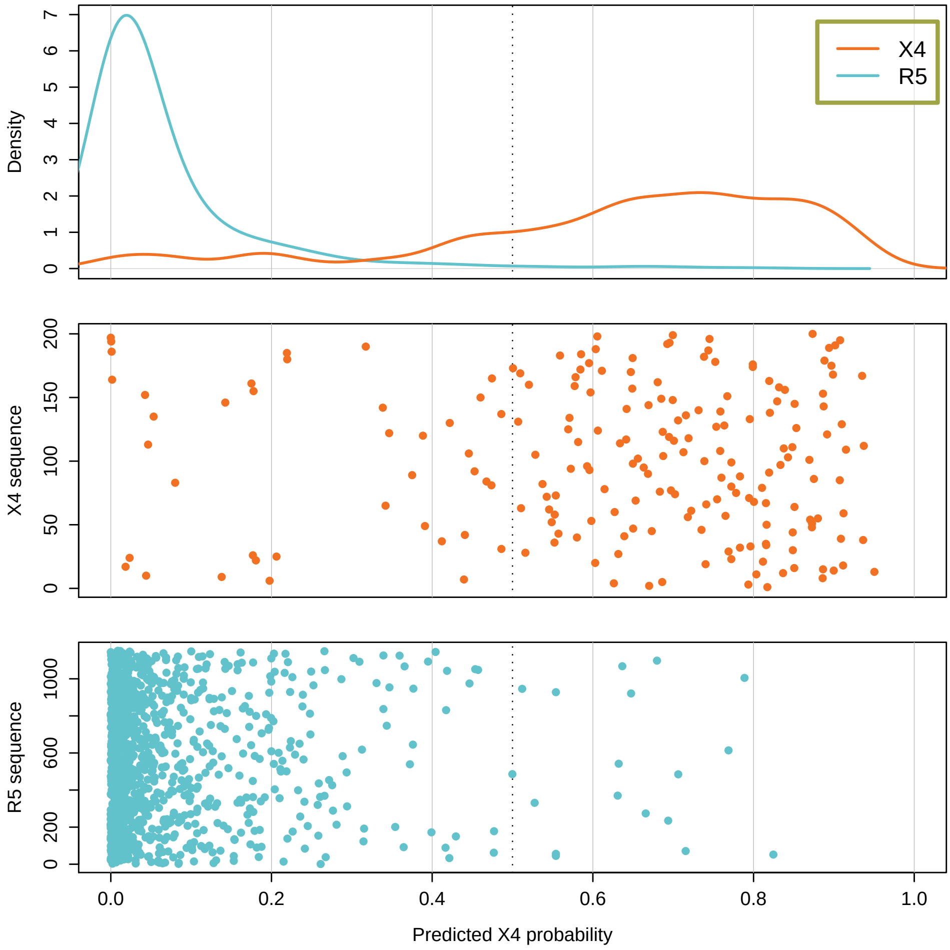

Figure 2.7: **Distribution of predictions from the electrostatic potential model.** The plot shows the distribution of prediction for R5 (blue) and X4 (orange) sequences made by the electrostatic potential model. Predictions were subject to an internal crossvalidation as described in Section 2.1.2.

characteristic (ROC) curve. This curve plots the true positive rate (TPR) against the false positive rate (FPR), for a cutoff ranging from 0 to 1. Given a perfect classifier and an error-free training and test set, the ROC curve would reach a combination of TPR = 1.0 and FPR = 0.0. A common measure in combination with ROC curves is to integrate the area under the ROC curve (AUC) [38]. The AUC can vary between 0 and 1, with a random classification resulting in an AUC of around 0.5. Performance measures calculated in the following were averaged over ten independently trained models. The electrostatic potential classifier exceeds a TPR of 0.9 at a FPR of 0.1, for a cutoff of 0.19. This means that while correctly predicting over 90% of the X4 sequences, less than 10% of R5 sequences are misclassified. The electrostatic potential model has an AUC of 0.957±0.001. For comparison reasons, the performance of the 11/25 and 11/24/25 rules on the same sequence set are shown. Figure 2.8 supports our initial theory that a classifier trained by machine learning and inspired by the 11/25 and 11/24/25 rules could outperform these simple models. This was possible by taking into account a finer resolution and more detailed representation of electrostatics by the electrostatic hull.

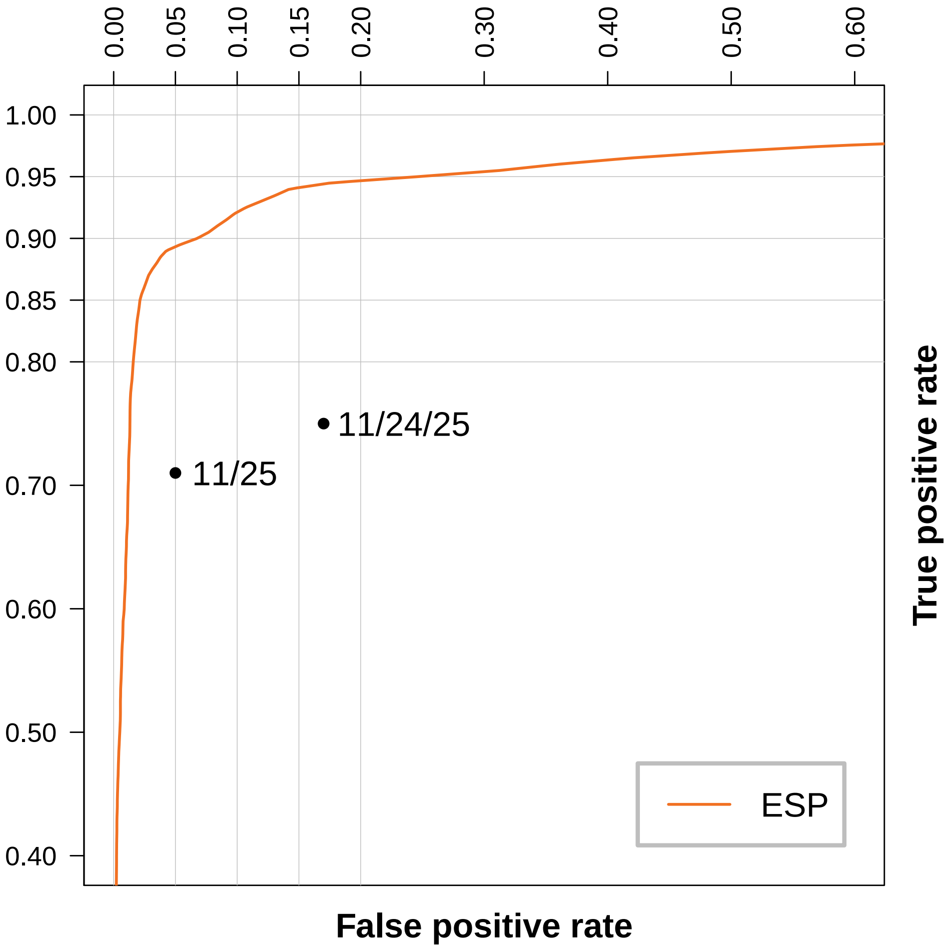

Figure 2.8: **ROC curve of the electrostatic potential classifier.** The plot shows a receiver operating characteristic (ROC) curve. The curve is drawn by calculating the true positive (y-axis) and false positive (x-axis) rates of the model predictions for a varying cutoff between the X4 and R5 classes. Predictions were subject to the internal crossvalidation of the random forest algorithm as described in Section 2.1.2.

The dataset of V3 sequences used for model training contains sequences that are originating from the same patient. It is plausible that these sequences are likely to have a higher similarity to each other than to other sequences within the dataset. Considering this similarity of sequences, one can deduce a source of bias, introduced by using a leave-one-sequence-out crossvalidation. To measure the effect of sequences originating from the same patient on the overall structure of the training data, pairwise alignment scores were calculated between all sequence pairs. For each sequence the highest alignment score to any other sequence within the set was recorded. The distribution of distances to the nearest neighbor show how well sequences within a set are connected to other sequences. A set of very similar sequences will result in many high alignment scores, a very diverse set will probably include lower alignment scores, due to sequences with smaller similarities to other sequences. Figure 2.9 shows the degree of similarity between all sequences used for model training in orange. The same similarity measure for a reduced sequence set, only containing a single sequence per patient, is shown in blue. There is a clear difference between the two distributions. While most sequences within the complete training set have at least one very similar sequence neighbor (high alignment score), the distribution is significantly shifted towards lower similarity in the reduced training set. This can be explained by the removal of similar sequences originating from the same patients. A higher inner similarity of training sets probably affects the performance in a simple crossvalidation. For every sequence left out during a crossvalidation iteration, a very similar sequence, possibly from the same patient, is used for model training, raising the chance of a successful classification. Under these conditions, this simple crossvalidation scheme should not be used, because its result tells little about the performance for genuinely new sequences.

To test this hypothesis and to quantify the effect, we developed a leave-one-patient-out, or patientwise crossvalidation scheme. Details are found in the Materials and Methods section of this chapter. Using this scheme, the whole set of sequences originating from a single patient are left out of the training set and predicted by the trained model.

Figure 2.10 shows a direct comparison of the performance of the electrostatic potential model with a sequencewise crossvalidation and a patientwise crossvalidation. The results clearly show a difference in performance, both in terms of true positive rate and false positive rate. The performance estimation based on the internal crossvalidation scheme of the random forest algorithm (orange) thus seems to be biased by sequences originating from the same patient. At a false positive rate of 0.05, the difference in sensitivity is around 0.1. For smaller FPR, this difference is even larger. Using the patientwise crossvalidation, the AUC drops to 0.934 ±0.001. However, despite the more conservative performance estimation, the ESP classifier still performs better than the

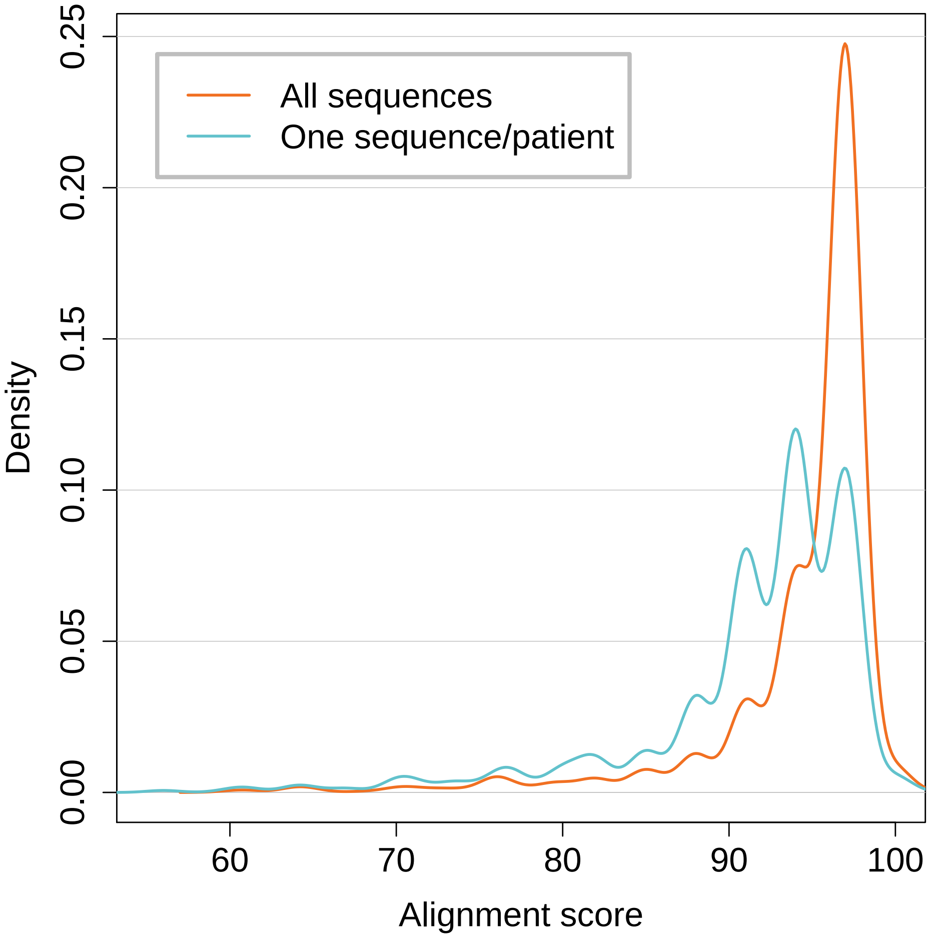

Figure 2.9: **Sequence similarities as alignment distances.** The density plots show the distribution of maximum alignment scores of all sequences within the complete sequence set against each other (orange), and for a reduced set containing only one sequence per patient (blue). The distributions show that in the complete set, the majority of sequences have at least one very similar sequence within the same set, as can be seen from the high alignment scores of 95 to 100. In the reduced set, the distribution is shifted towards lower alignment scores, meaning that, on average, sequences are less similar to their nearest sequence neighbor.

11/25 and 11/24/25 rules with an increased true positive rate of around 0.09 and 0.16 at the respective false positive rates, indicating that more sophisticated methods, like machine learning are able to extract more information from the charges of the V3 loop than alignment- or rule-based models.

Figure 2.11 shows the error rate of the model for the two classes X4 (blue) and R5 (orange) for different cutoff levels. On the left side, the error rates are estimated from the

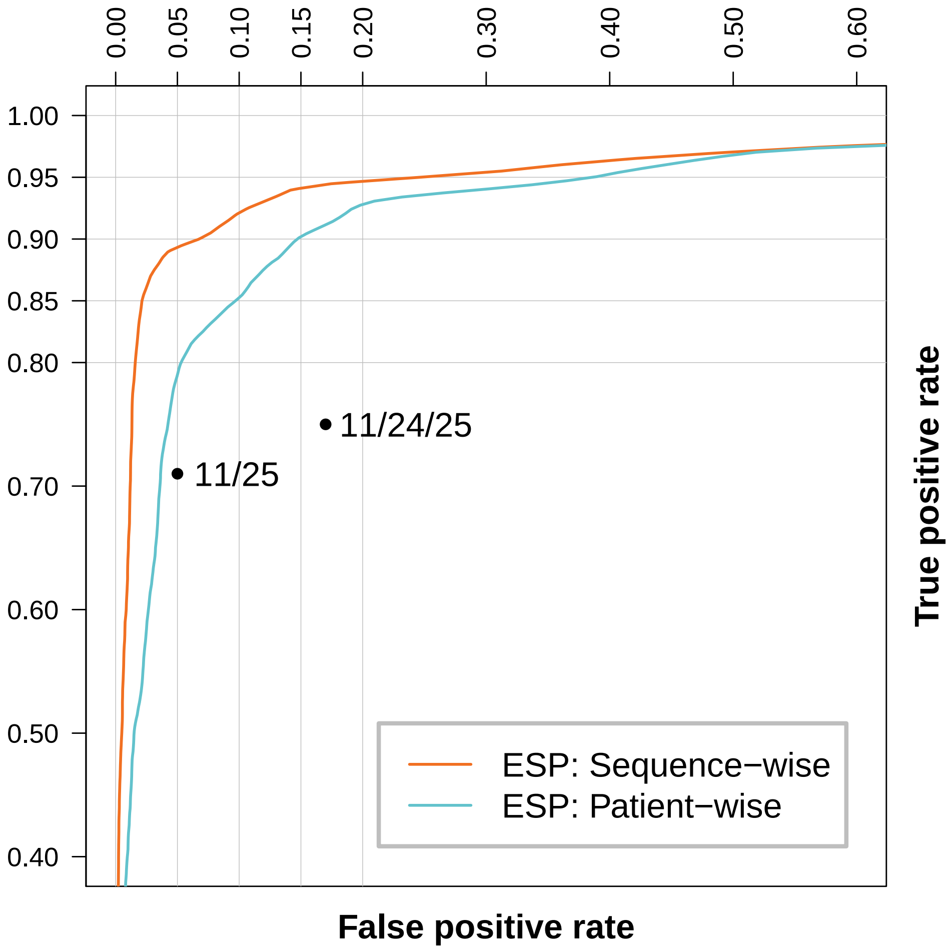

Figure 2.10: **Comparison of performance based on the crossvalidation scheme.** The plot compares the performance of the electrostatic potential model using the sequencewise (orange) and patientwise (blue) crossvalidation schemes. The sequencewise crossvalidation seems to overestimate the model performance, due to multiple sequences originating from the same patients.

internal sequencewise crossvalidation performed by the random forest algorithm. With the previously reported bias of that crossvalidation scheme due to sequences originating from the same patient, the error rates are higher when using patientwise crossvalidation (right side). The optimal cutoff depends on the importance one gives to either of the classes. Currently, the only co-receptor antagonist approved for treatment in both the United States and Europe is maraviroc. The MOTIVATE 1 and MOTIVATE 2 studies performed to asses the virologic and immunological efficacy of maraviroc have shown that maraviroc in combination with other drugs (NRTI, NNRTI, PI) was able to reduce significantly reduce the viral load of patients with non-detectable CXCR4 virus

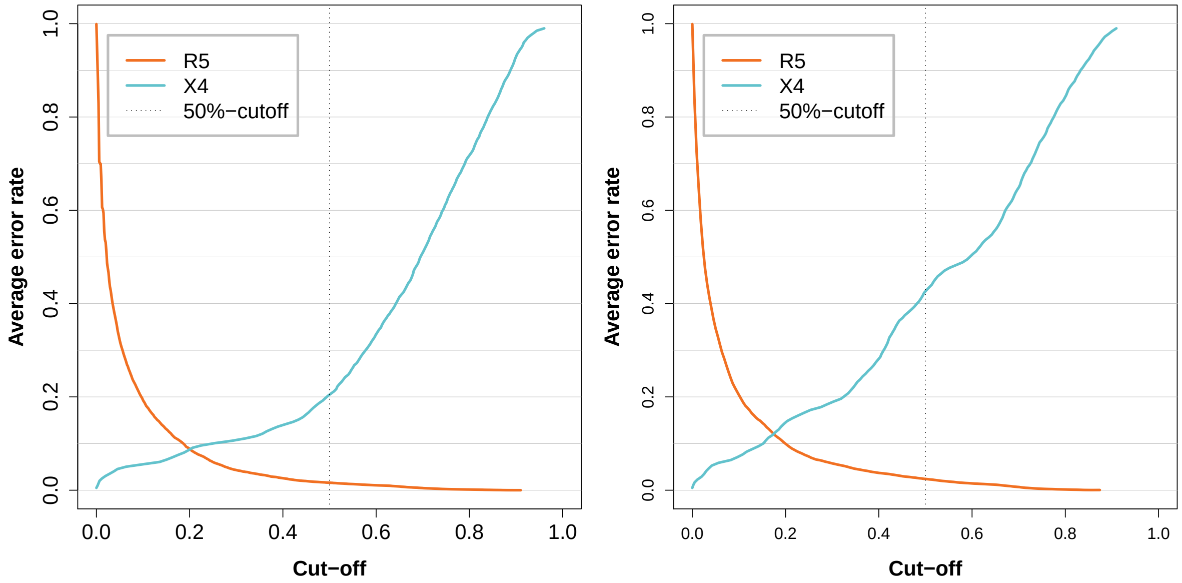

Figure 2.11: **Comparison of cutoffs based on the crossvalidation scheme.** The plots shows the rate of errors made for the respective classes X4 and R5 using a sequencewise (left) and patientwise crossvalidation (right).

(see Section 1.4.3). Patients with detectable CXCR4 virus were offered to participate in a parallel study and receive the same treatment. Although these patients did not benefit from the treatment of maraviroc, no rapid decline in CD4 counts of faster disease progression were witnessed [124]. In a recent follow-up study conducted by Pfizer, 15 patients from the MERIT study who experiences therapy failure due to co-receptor switches were analyzed. All patients were reported to respond well to subsequent therapy, showing no signs of adverse clinical consequences [113]. These findings suggest a more liberal CXCR4 detection strategy applying higher cutoffs to decrease the false positive rate, thus including more patients for treatment with maraviroc. On the other hand, increasing the cutoff will result in an increased rate of effective maraviroc treatments. A widely accepted method of choosing the prediction cutoff is to look at the results in false positive rates. The *German Recommendations for the Determination of Co-receptor Usage* [26], for example, recommend choosing a cutoff resulting in a FPR of 20% for patients with many therapy options, and a more liberal cutoff resulting in 12.5% FPR for patients with limited treatment options (Figure 2.11 shows that FPR is the R5 prediction error rate). When evaluating the performance with the patientwise crossvalidation scheme (right curves), an FPR of 12.5% and 20% would approximately be achieved for the cutoffs 0.17 and 0.1.

2.3.2 Prediction of Tropism Based on the Hydropathy Scale

Prediction of co-receptor usage based on the electrostatic potential of V3 loops relied on homology modeling of the V3 loop structure. The electrostatic potential above the solvent-accessible surface is directly related to the loop sequence, but is also largely dependent on the orientation, size, and charge of aminoacid sidechains. These attributes are of structural nature and considered by the electrostatic hull model. The dimensionality of a sequence descriptor is much lower, compared to that of a structural descriptor. For instance, changes in sidechain interaction due to their conformation are not modeled. To test the performance of a reasonable sequence descriptor, a second random forest model was built, based on the same set of V3 loop sequences, translated according to the Kyte-Doolittle hydropathy scale and normalized to a constant length of 38. The input table for random forest generation relates the known classes of tropism to the translated sequences. The data structure used for training is schematically shown in Table 2.2.

Table 2.2: Data structure of training data used for the hydrophobicity model.

V3 Sequence	**Tropism**	h_1	h_2	h_3	...	h_{38}
$V3_1$	X4	0.7777	0.4510	0.0684	...	0.7777
$V3_2$	X4	0.7777	0.4798	0.1369	...	0.7777
$V3_3$	X4	0.7777	0.7021	0.0378	...	0.7777
...	...	...	...	...	...	...
$V3_{201}$	R5	0.7777	0.4606	0.0912	...	0.7777
$V3_{202}$	R5	0.7777	0.4510	0.0684	...	0.7777
$V3_{203}$	R5	0.7777	0.9759	0.2162	...	0.7777
...	...	...	...	...	...	...
$V3_{1351}$	R5	0.7777	0.4510	0.0684	...	0.7777

The prediction model based on hydropathy scale consisted of 500 decision trees, each tree using $m = \lfloor\sqrt{38}\rfloor = 6$ variables at each node for splitting. In total, ten random forests were generated to determine the average performance based on a patientwise crossvalidation. Figure 2.12 shows the cutoff-dependent error rates for the classes R5 and X4. In comparison to the error-curves of the electrostatic model (Figure 2.13), the R5 curve is slightly shifted towards lower cutoffs. The cutoffs achieving false positive rates of 12.5% and 20% are 0.12 and 0.06, respectively. The error-rate of the X4 class reaches a lower level earlier than for the electrostatic potential model, but seems to maintain a higher level at low cutoffs.

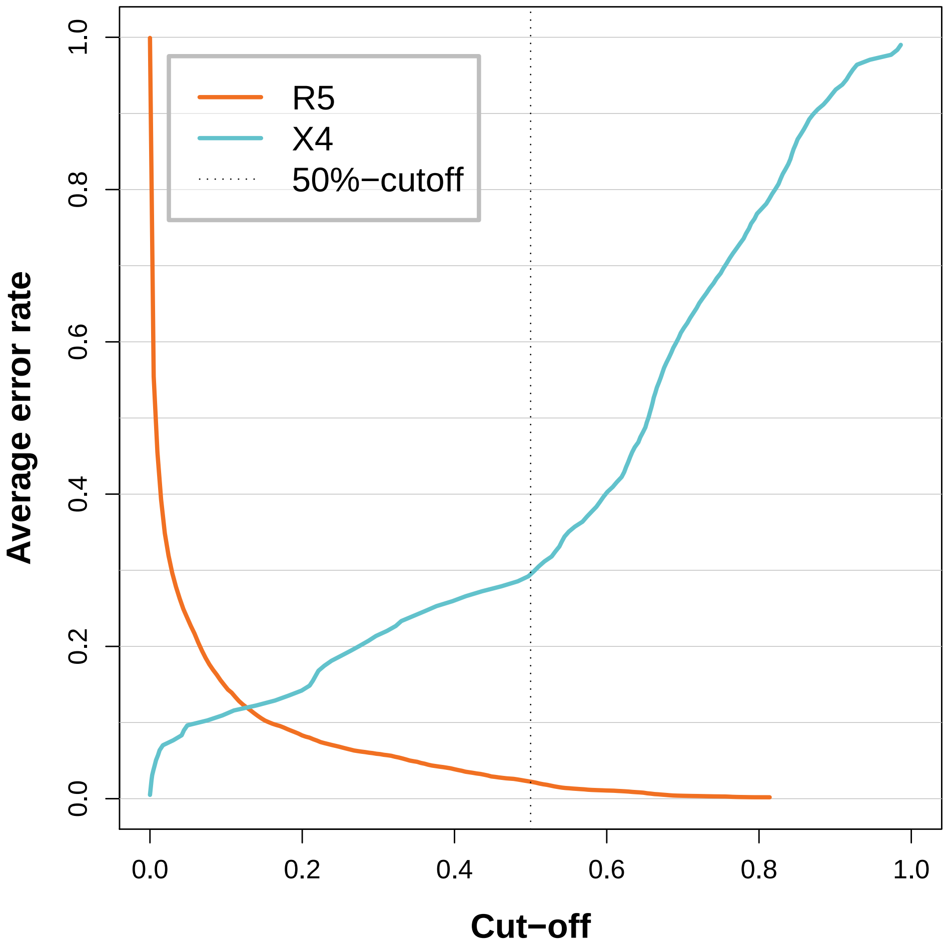

Figure 2.12: **Cutoff-dependent error rate of the hydrophobicity model.** The plots shows the rate of errors made by the hydrophobicity model for the respective classes X4 and R5 using a patientwise crossvalidation.

The ROC curves in Figure 2.13 allow comparison of the performance of the electrostatic and hydrophobicity models. None of the models is superior to the other, with the electrostatic model having an advantage at higher false positive rates and the hydrophobicity model dominating at lower FPRs. This insight leads to the conclusion that both models could be combined to improve the overall performance. When comparing the two false positive rates recommended by the German Guidelines, the electrostatic potential model seems to perform better. While both models reach the same TPR at a FPR of 12.5%, the TPR of the hydrophobicity model is approximately 3% lower at the 20% FPR than the ESP model. The AUC of the hydrophobicity model is 0.930 ±0.001.

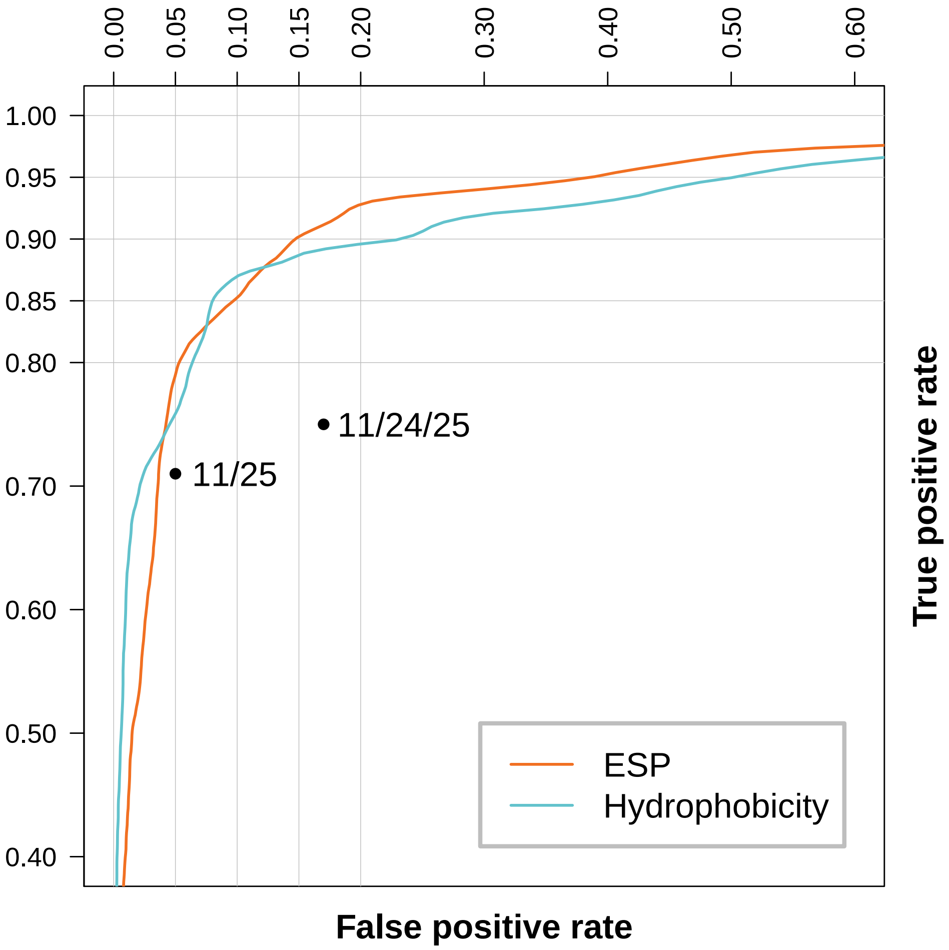

Figure 2.13: **Performance comparison of ESP and hydrophobicity classifiers.** ROC curves of the electrostatic potential and hydrophobicity models, based an the patientwise crossvalidation scheme. For this sequence set, neither model seems to be superior to the other, the electrostatics model performing better at higher, the hydrophobicity better at lower FPR.

2.3.3 Fusion of Electrostatics and Hydrophobicity Models

The previous results have shown that machine learning models based on different descriptors are equally able to classify HIV-1 viruses into R5 and X4 variants simply by evaluating their V3 loop sequences. Two models were trained. The first, based on the electrostatic potential on a virtual hull surrounding the loop, can be regarded as a structure-based descriptor. The second model, using the hydropathy scale of Kyte and Doolittle relies on the discretized aminoacid sequence of a V3 loop, excluding any structural information. Random forests are ensemble classifiers, a collection of independent decision trees voting on the class affiliation of unknown data, reaching a final decision by majority vote. An ensemble classifier is defined as a collection of single classifiers. These classifiers can be of any kind or complexity. Decision trees can be used, as seen in random forests, but even quite complex classifiers can be combined into an ensemble classifier. In theory, one can iteratively create ensemble classifiers, based on the votes of underlying classifier ensembles. This is known as stacking [165]. The performance gain of combining an ensemble of classifiers greatly depends on the underlying descriptors and combination method used. Combining classifiers by means of machine-learning, one would expect, as a worst case, to see no improvement in performance, as the resulting model should be able to learn to imitate the best single classifier in the ensemble. A simple example of a case where no performance improvement is expected is the combination of multiple classifiers of the same set up, based on the same descriptor. In this naïve case, every single classifier evaluates the same data under the same circumstances, thus there is nothing that can be learned from their combination. This changes, once one combines classifiers evaluating different aspects on the same data, which is true for our two previously built models.

The structural classifier, based on the electrostatic hull has a different perception of the data than the hydrophobicity-based sequence classifier. As a consequence, it is possible for the two classifiers make discordant predictions for the same V3 sequence. This was already indicated by the comparison the ROC curves of the two models. None of the models was superior to the other, each dominating the other at different cutoff levels (Figure 2.13). The assumption is further verified by correlating predictions of the two classifiers. Figure 2.14 shows a scatterplot of all sequences in a prediction plane. Each point in the plot represents a single sequence of the dataset placed according to the predictions made by the electrostatic potential classifier (x-axis) and the hydrophobicity classifier (y-axis). The prediction can be understood as a probability of a sequence being X4. Accordingly, R5 sequences (orange circles) are clustered in the lower left corner, while CXCR4 using virus variants (X4 and R5X4) cluster in the upper right corner. The majority of sequences have differing predictions from the two classifiers. The remaining corners, upper-left and lower-right are regions, where predictions from the two classifiers are contradictory. Although no sequences are found there, some sequences are closer to these conflicting corners than to either of the R5 or X4 corner. For these sequences it is hard to make a decision. Intuitively, one could couple the predictions using simple mathematical operators like minimum or maximum. Choosing the smaller, or larger of the values as final decision. Another possible solution would be to find a linear two-

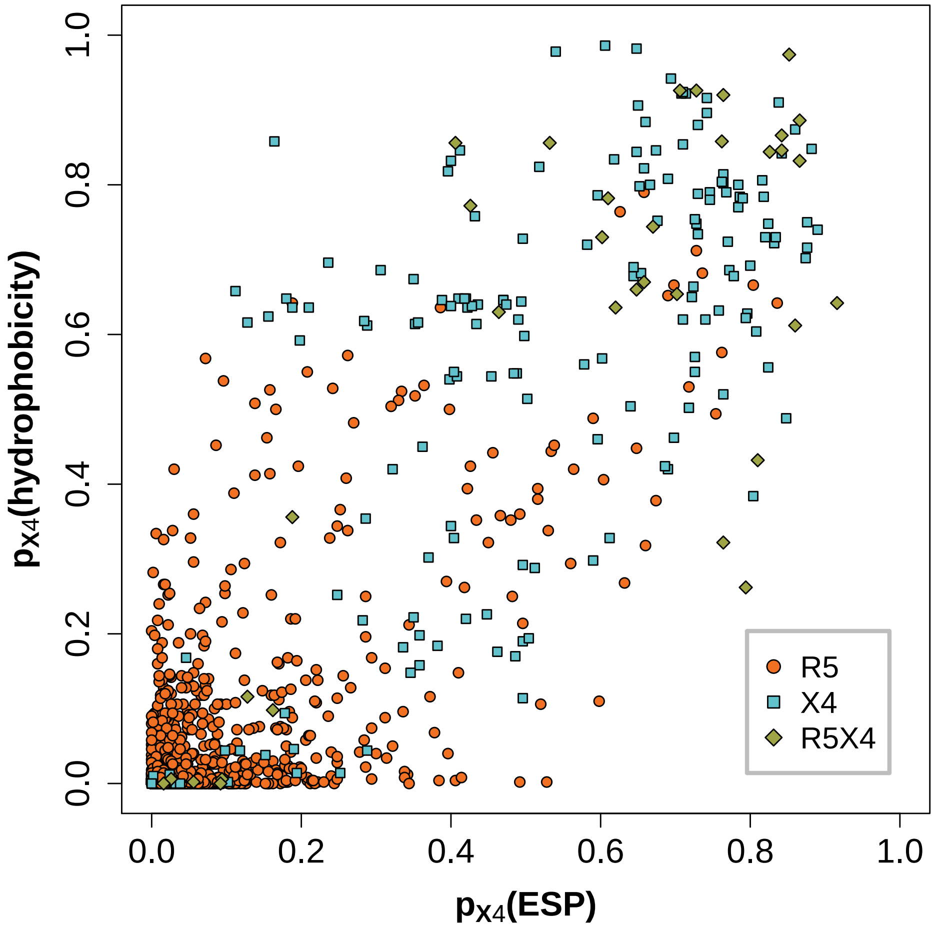

Figure 2.14: **Scatterplot of electrostatic model vs. hydrophobicity model votes.** The plot shows the results of both prediction models on each of the V3 loop sequences of the training data set. The votes are results of the patient-wise crossvalidation scheme. Concordant votes of both models are found in the lower left and upper right corners, while the sequences with discordant predictions spread towards the lower right and upper left corners.

dimensional separator function, dividing the prediction plane into an R5 and an X4 region. An even more sophisticated and possibly more powerful approach is to learn a non-linear separator function.

In order to determine the effect of a sophisticated separator function, another random forest was trained and compared to the simple separator functions minimum (min), maximum (max) and multiplication ($mult$). In the spirit of stacking the prediction results of the two independent classifiers were used as input for the new second-level model. The goal was to find an optimal second-level classifier to interpret the individual predic-

Table 2.3: Data structure of training data used for the combined model.

V3 Sequence	**Tropism**	$p_{X4}(ESP)$	$p_{X4}(Hydrophobicity)$
$V3_1$	X4	0.78	0.30
$V3_2$	X4	0.66	0.96
$V3_3$	X4	0.80	0.79
...	...	...	...
$V3_{201}$	R5	0.28	0.11
$V3_{202}$	R5	0.10	0.00
$V3_{203}$	R5	0.00	0.00
...	...	...	...
$V3_{1351}$	R5	0.01	0.00

tions and to learn how to classify sequences with discording results. In order to have a good data basis to the train the second-level random forest, the scatterplot of the two independent first-level predictions was closely examined. The training of ten individual first-level models revealed a standard deviation of predictions of 0.007 and 0.009 for the hydrophobicity and electrostatic potential classifiers, respectively. This variation can be explained by the the random selection of features during forest generation. The prediction landscape (shown for one set of first-level models in Figure 2.14) will, thus, look slightly different when re-training electrostatic or hydrophobicity models.

The simple classifiers *min* (2.1), *max* (2.2) and *mult* (2.3) considered the smaller, larger or multiplied values of the first-level predictions, respectively.

$$p_{min} = min(p_{ESP}, p_{hydrophobicity}) \tag{2.1}$$

$$p_{max} = max(p_{ESP}, p_{hydrophobicity}) \tag{2.2}$$

$$p_{mult} = p_{ESP} \cdot p_{hydrophobicity} \tag{2.3}$$

The data structure used for training of the second-level random forest is shown in Table 2.3.

Figure 2.15 shows the ROC curves of the second-level models random forest (RF), minimum (min), maximum (max) colored in green, red and black, respectively. The ROC curve of the random forest model (green) shows the superior performance of the new model over the two first-level models (orange and blue). The AUC of 0.95 ±0.001

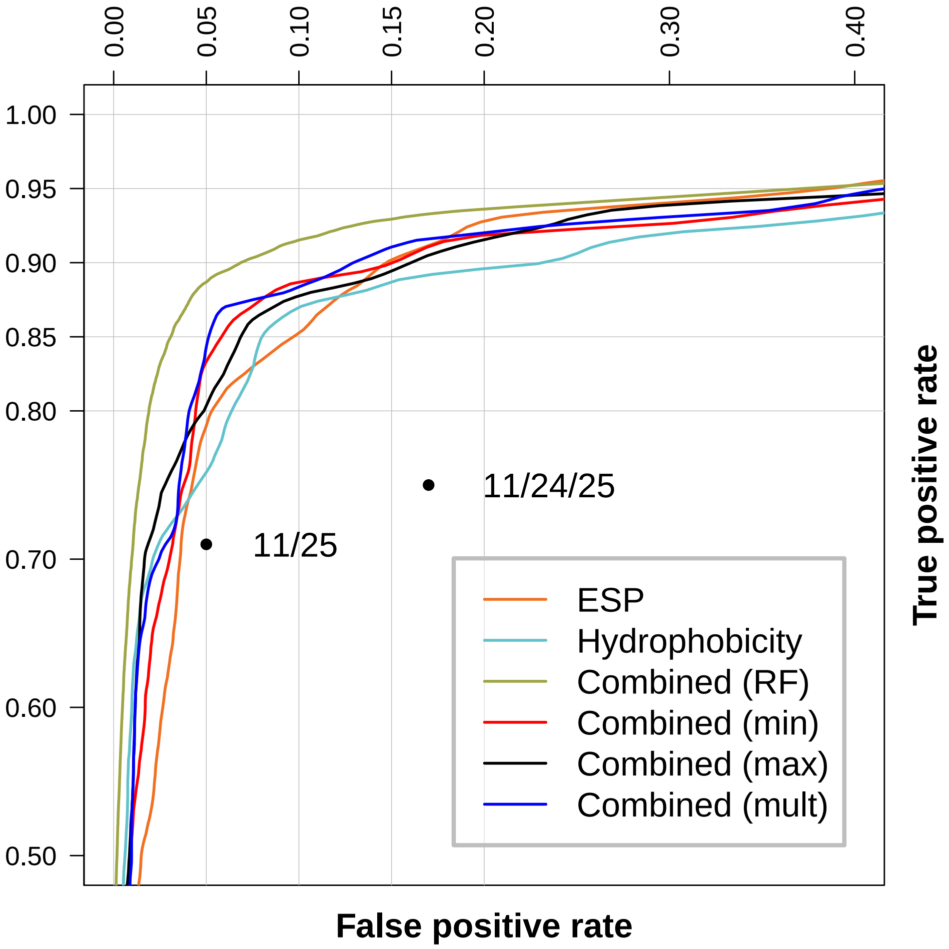

Figure 2.15: **ROC curves of the different second-level classifiers compared to the first-level classifiers.** The second-level classifiers seem to generally perform better than the first-level classifiers alone, with the exception of the ESP model at high FPR and the hydrophobicity model at low FPR. The random forest second-level classifier, however is superior to every other model for this dataset, using a patientwise crossvalidation.

is significantly higher than those of the other models. Although the combined model seems to slightly outperform the ESP model at the FPR of 12.5% it holds an increase of TPR of around 5% over both first-level models at the FPR of 20%. Remarkably, it seems that the FPR can be decreased by more than 14% (20% to 6%) while sacrificing only about 3% of TPR (from 93% to 90%).

The following analyses will examine how the second-level classifiers are evaluating the independent votes of the electrostatic and hydrophobicity model. In comparison to the simple classifiers *min*, *max*, and *mult*, the evaluation scheme of a sophisticated model such as the random forest is not intuitive. In order to visualize the prediction pattern learned by the random forest, a set consisting of all possible pairs of first-level predictions was generated. Each pair was then evaluated by the respective second-level classifiers *min*, *max*, *mult*, and the random forest. The results are prediction landscapes as shown in Figure 2.16. Each coordinate of a landscape represents a pair of first-level votes. The color at that coordinate encodes the outcome of the second-level classifier for the specific first-level votes according to the right handed color-keys. The colors are chosen to visualize the clinically relevant false positive rates (12.5% and 20%) of the models. Enclosed in the green areas lie 80% of all R5 sequences of the patientwise crossvalidated training data, thus choosing this rigid cutoff to distinguish between X4 and R5 sequences yields a FPR of 20%. The cutoff enclosing 87.5% of all R5 sequences (12.5% FPR) combines the green and orange areas. Extending the cutoff into the blue area will result in more sequences being interpreted as R5. In a clinical setting this would mean, that more patients would be considered for treatment with co-receptor antagonists. In turn, one would increase the risk an ineffective treatment, due to misclassified X4 viruses. Looking at the prediction landscapes of the simple second-level classifiers (min, max, product) one notices the uniform shape of different relevant FPR-classes (green and orange). All of these classifiers have been shown to work reasonably well, none being inferior to any of the first-level classifiers. On the other hand, none of the simple classifiers could outperform the first-level classifiers. In contrast, the second-level random forest is clearly superior to the single classifiers. At almost every false positive rate, the random forest had a higher true positive rate than any of the first-level classifiers (Figure 2.15). The superior performance can be explained by looking at the prediction landscape of the random forest (Figure 2.16, bottom right). In strong contrast to the other classifiers, the FPR classes are non-uniformly distributed within the lower left corner. In addition to a rather large landmass there are numerous islands, interrupted by other FPR-classes. One can find two very notable features. One, a large, rugged island of the 12.5% and 20% FPR classes (orange and green), completely surrounded by area of low FPR (blue), almost in the center of the landscape. This is somewhat surprising, because intuitively one might expect that a steady increase of one of the first-level predictions, will result in a steady increase of the second-level outcome. The opposite phenomenon can be seen in the lower left corner. A lake of low FPR-class (blue), resulting in an X4 prediction by the second-level model, is found in an area surrounded by high FPR classes. A similar, smaller lake is located right next to the larger one. This surprisingly shaped landscape of the second-level random forest can be understood when taking a look at the basis of training. Figure 2.14 shows the sequences located according to the respective first-level votes used for training the second-level model. In essence, the model has learned the shape and locations of the islands, both blue surrounded by green and orange, and vice versa, because they were present in the training data. Both areas show a significant local amount of sequences of one class, in the absence of sequences from the other class. Although the training landscape consisted of the votes of ten first-level forests to include the variation in outcomes due to the randomness during forest generation, it could be that the landscape was not populated enough to define a clear uniform shape or general distinction between the R5 and X4 classes. One can argue that due to possibly misclassified training sequences (X4 in the lower left corner, R5 in the upper right corner), or coincidental local accumulation of sequences of one class, the forest was confused, resulting in overfitting. The method of predicting co-receptor tropism based on evaluating independent classifications on a second-level has been named T-CUP (Two-level Co-receptor Usage Prediction).

In order to measure the degree to which the second-level random forest was prone to overfitting based on the training data, a blind and independent test set of V3 sequences was compiled. The set included sequences gathered from literature, as well as two set of sequences from the Institute of Virology in Cologne and the Computational Biology group at the Max Planck Institute in Saarbrücken (both in Germany). All sequences were filtered as described in Material and Methods of this chapter. The final validation set contained 381 sequences in total, with 132 X4 and 249 R5 sequences. First-level predictions by the electrostatic and hydrophobicity models were performed for each sequence. The first-level votes were then used as input for the second-level prediction by the four classifiers *min*, *max*, *product*, and the random forest. To complete the performance comparison, the test set was also submitted to prediction by geno2pheno$_{[corecpetor]}$ (`http://coreceptor.bioinf.mpi-inf.mpg.de/`) using the recommended FPR (12.5%) and Web-PSSM (`http://indra.mullins.microbiol.washington.edu/webpssm/`) using the x4r5 subtype B matrix. The performance of the 11/25 and 11/24/25 rules was also calculated. The results of the complete performance comparison in form of a ROC curve, is shown in Figure 2.17. The 11/25, 11/24/25 rules and geno2pheno (g2p) only gave a binary decision for each sequence, either X4 or R5. For these methods only single pairs of FPR and TPR could be generated, and are visualized by black dots in the ROC curve. PSSM calculates a score for each sequence which can be interpreted by a relative cutoff, thus calculation of a ROC curve was possible (black line). Immediately one no-

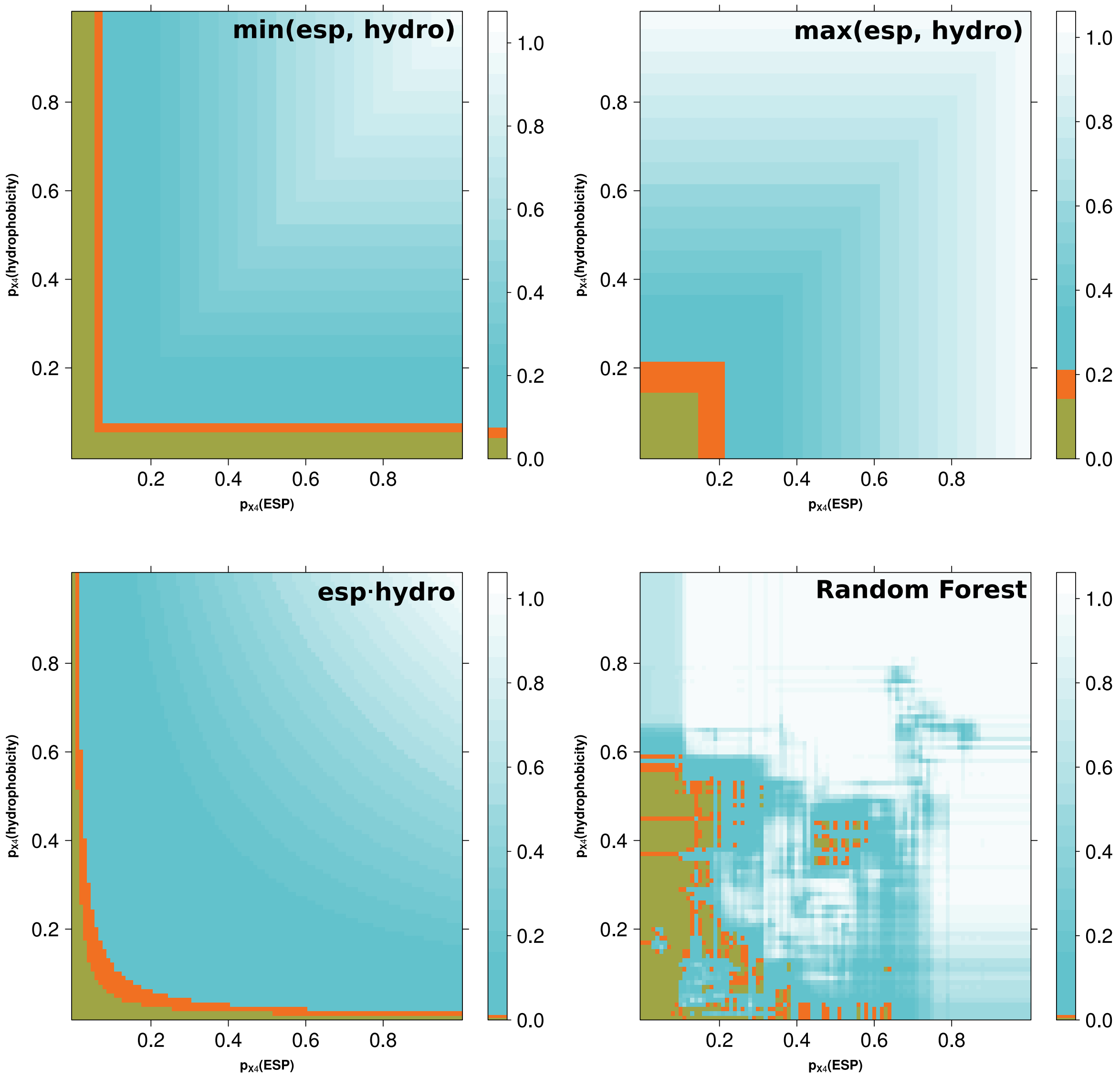

Figure 2.16: **Prediction landscapes of the different second-level models.** The landscapes show the predictions made by the second-level classifiers coded by colors. The respective first-level votes are shown on the axes. While the simple mathematical classifiers (min, max, multiplication) have a monotonous development from low to high prediction results, the random forest is not clearly structured, the islands could indicate local overfitting.

tices the overall reduced performance of T-CUP on the blind test set. This point will be addressed in the discussion (Section 2.4). All second-level classifiers are performing equally in the low FPR regions, with about 70% and 75% TPR at the relevant FPR or 12.5% and 20%. Astonishingly the random forest, frontrunner in the patientwise crossvalidation of training data, is even slightly outperformed by every of the simple classifiers. The superior performance of the random forest as second-level classifier on the crossvalidated training data was thus probably caused by local overfitting. The best performing classifier in the low FPR regions (<20%) seems to be *mult*. When looking at higher FPR (beyond 20%) the *max* classifier stands out. Thus, there is not one superior classifier, in fact, the right one should is based on the desired FPR. In essence, when the goal is to include ensure a larger number of effective treatments, the *max* classifier seems better suited than any other. On the other hand, when accepting a higher number of ineffective treatments in order to include more patients into therapy with co-receptor blockers, the more liberal classifiers *mult* or *min* could be chosen. All T-CUP classifiers performed well in comparison with current state-of-the-art methods geno2pheno and PSSM, the latter was outperformed by all of them. In direct comparison with g2p, the *mult* classifier performed equally well, while the *max* classifier showed an improvement of about 5% in TPR at the same FPR of 26%. The 11/25 and 11/24/25 rule based methods were inferior to PSSM, and all T-CUP classifiers.

2.4 Discussion

The results have shown that the combination of more than one prediction model can improve the prediction of co-receptor tropism based on V3 loop sequences. This has been shown both for the crossvalidated training data and, more important, on a blind-validation test set. T-CUP generates independent random forest machine learning models on a first level and combines these using different classifiers on a second level to reach a final result. Different second-level classifiers were tested. It has been hypothesized that a more sophisticated and non-linear second-level classifier, like another random forest, would outperform more simple second-level models. While this was true for the performance based on the crossvalidation of the training data, the random forest did not fare better than any of the simple classifiers during a validation using 381 clinical sequences.. In fact, the random forest, on average, was inferior to any of the simple classifiers, *min*, *max*, and *mult*. There are three possible explanations for this phenomenon. First, the training data did not contain any sequences originating from Germany, while the sequences of the blind test set were exclusively collected in Germany. In that case, the training data is not a globally representative ensemble of V3 sequences. Second, at least one of the data sets contained errors. Third, the second-level random forest was overfit-

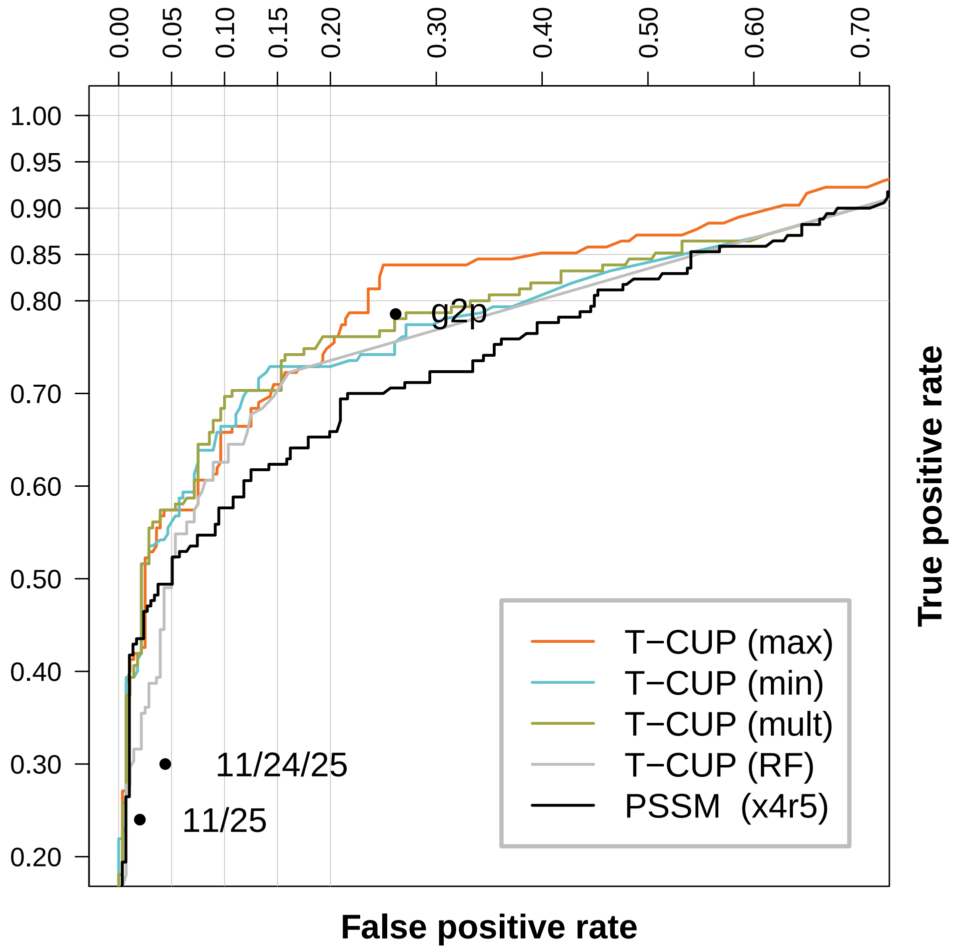

Figure 2.17: **Performance comparison of the different genotypic tropism prediction methods.** The overall performance is reduced in comparison to the patientwise crossvalidation on the training data. Still, the two-level models fare significantly better than PSSM and the rule-based approaches. The random forest second-level classifier is not superior to any of the simple models. The best models seem to be *mult* at lower and *max* at higher FPR.

ted to the training data. In either case, no definitive conclusion can be made about the performance of T-CUP. In an attempt to quantify the difference or similarity between the initial training data and the sequences in the blind test data set, the training set was compared to the test sequences. To that end, pairwise alignment scores were calculated between each pair of training sequences, while limiting the set to a single sequence per patient (see Section 2.3.1 and Figure 2.9). For each sequence the best score to any of the other sequences was selected. The distribution of these scores reflects how well sequences within a set are connected to each other. Then, for each sequence of the blind test set, the maximum alignment score to any of the sequences from the training set was calculated. The distribution of scores is shown in Figure 2.18. The orange curve shows the distribution of maximum alignment scores of training sequences to other sequences within the training set. Training sequences seem to be connected rather well with each other, with most sequences having a high alignment score to at least one other sequence. Many sequences of the blind test set have a quite considerable distance to the nearest sequence neighbor in the training set. Obviously the training set used was not a representative set for the V3 loop sequence space. It can be assumed, that the performance for these sequences is significantly lower than for sequences with close sequence relation to the training set.

The improved predictive performance of the models is an important result, but can the models be used to draw biologically relevant conclusions on the difference between V3 loops associated with X4 or R5 tropic viruses? The random forest algorithm comprises a metric by which the relative importance of single model variables can be described, a so called importance analysis (see Section 2.1.2). In the case of the hydrophobicity model, each V3 aminoacid sequence was encoded into a numeric sequence, representing the hydrophobicity of the present residue. Thus, the model learned the tropism, based on sequences of 38 numerical variables. The importance value calculated for each variable gives the mean decrease in prediction accuracy, when randomly permuting this variable in the training sequences. In essence, the higher the importance value, the more important a variable is for an accurate prediction. Figure 2.19 shows the importance of the 38 variables of the hydrophobicity model, as computed by the random forest algorithm. A direct interpretation of variables in terms of aminoacid sequence position is impossible, because the initial hydrophobicity sequences were normalized to a constant length of 38 to cope with the heterogeneity of V3 loop length. Still, the normalized positions closely reflect the initial residue numbering of the V3 loops. The most important variable is at position 12. This probably coincides with the 11/25 rule, highlighting the importance of that residue position for distinguishing X4 from R5 sequences. Although the 11/25 and 11/24/25 rules base their prediction on the presence, or absence, of a charged residue at the respective positions, the hydrophobicity model pays great attention to this position as well. This can be explained by the hydrophilic nature of charged aminoacids. Charges are implicitly included in the hydrophobicity scale. Surprisingly, position 25 is not of elevated importance. The second most important variables for predicting co-receptor usage are found at positions seven and eight. In the normalized hydrophobicity sequence, these positions correspond to a region where asparagines are frequently discovered, and are

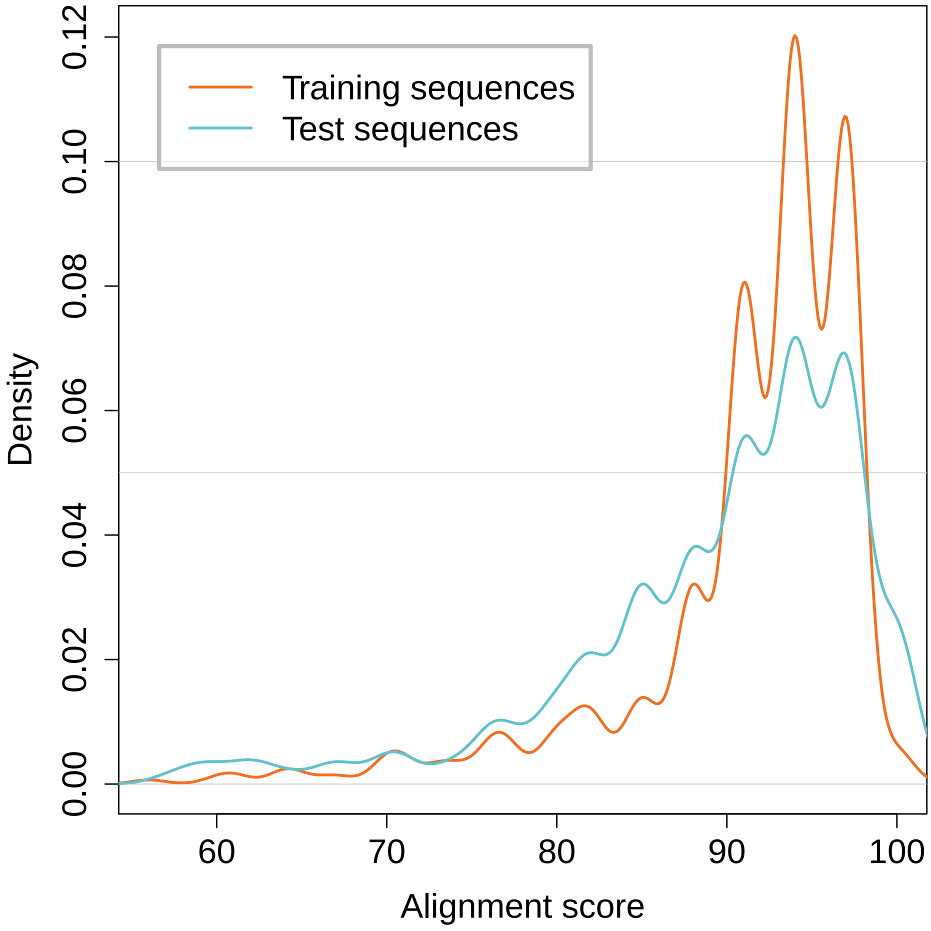

Figure 2.18: **Sequence similarity of blind test data to training data.** Density of maximum alignment distances of each sequence to other sequences within in the training data is shown in orange. The distribution of maximum alignment scores of each sequence of the blind test data set to any sequence of the training set is shown in blue.

hypothesized to undergo N-linked glycosylation. The post-translational modification at this multi-asparagine motif has been reported to be important for gp120 interaction with host co-receptors. Ogert *et al.* [105] found that the mutation of one of the asparagines at the glycosylation site of the V3 loop into glutamine inhibited the CCR5-mediated cell fusion, while the same mutation retained about 50% of the wild type fusion activity with CXCR4. These results were also found in another study, where mutations to the glycosylation site resulted in a complete reduction of CCR5-usage and only a slight decrease in CXCR4-usage [81]. Recent findings have hypothesized that asparagines in the region are involved in the binding to sulfated tyrosines at the N-terminus of CCR5

[56]. The importance analysis of the hydrophobicity model supports the relevance of positions 6-8 (glycosylation) and 11 (11/25 rule) of the V3 loop for co-receptor tropism.

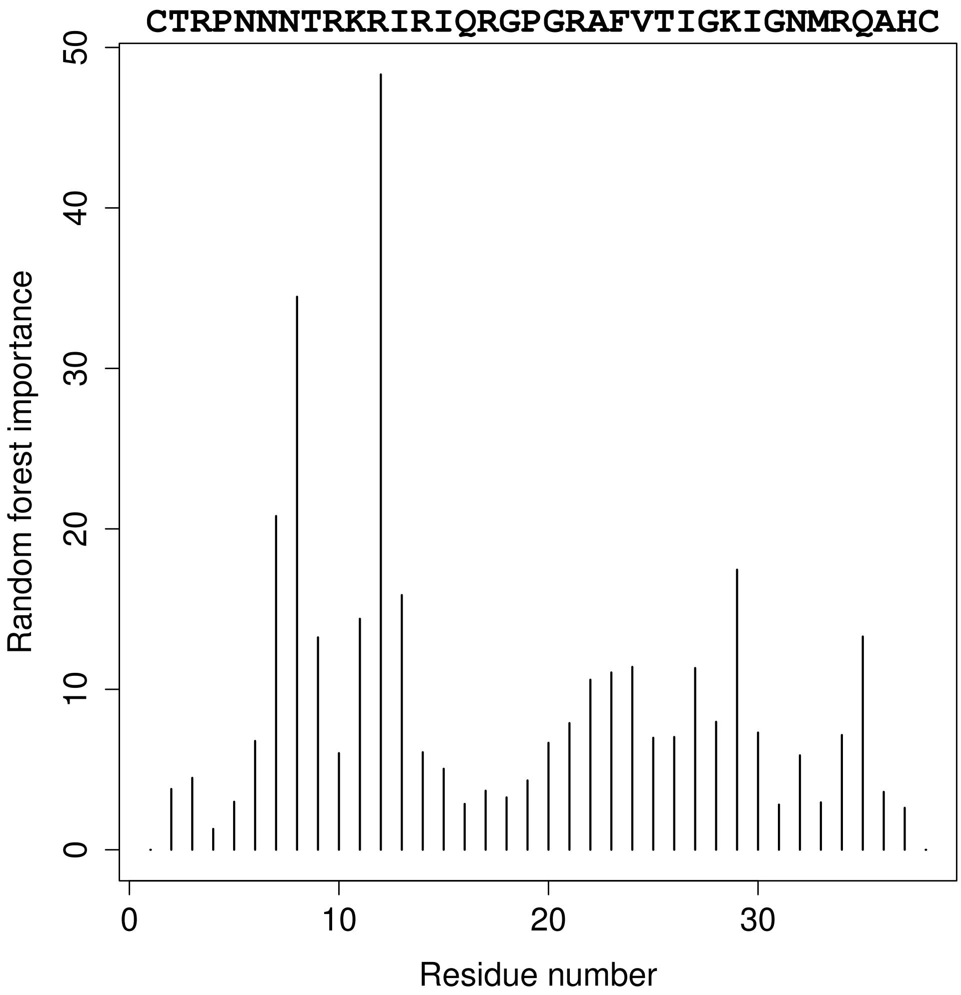

Figure 2.19: **Random forest variable importance analysis of the hydrophobicity model.** Importance of positions of normalized V3 sequence in random forest classification with Kyte-Doolittle descriptor [76]. The higher the peak at the respective position, the more important this position for correct classification of sequences with respect to co-receptor tropism. For orientation, the HXB2 V3 loop reference sequence is shown above. The most important region is around normalized sequence position 12, in agreement with the 11/25 rule. The second most important region around position 8 could be involved in binding of sulfated tyrosine on CCR5 [56].

The importance analysis of the electrostatic model consisted of 642 variables in total. Each variable represents a value of the discrete electrostatic potential, located on the electrostatic hull around a V3 loop homology model (see Section 2.1.3). To visualize the region of important electrostatic potential, the 5% (n=32) most important variables were selected. For each of the classes, R5, R5X4, and X4, the mean electrostatic potential for selected variables was calculated. The results are shown in Figure 2.20, where the V3 loop is represented as a coil with small beads for the aminoacids (both in gray). Each loop is surrounded by colored spheres representing the important variables of the electrostatic hull. The spheres are colored according to the mean electrostatic potential ϕ for the respective classes R5, R5X4, and X4. The unit of the electrostatic potential is $k_B \cdot 300K/e$. The color classes are: red, $\langle\phi\rangle \leq -2.5$; light red, $-2.5 < \langle\phi\rangle \leq -0.5$; white, $-0.5 < \langle\phi\rangle \leq 0.5$; light blue, $0.5 < \langle\phi\rangle \leq 2.5$; blue, $2.5 < \langle\phi\rangle$. The selected coordinates of descriptive electrostatics are basically divided into three separate regions. One regions is located around residue 11, in concordance with the 11/25 and 11/24/25 rules, as well as the importance analysis of the hydrophobicity model. However, in contrast to the hydrophobicity model, the electrostatic model finds discriminative information at a second region around position 25. The third region again coincides with the hydrophobicity importance analysis, as the potential close to residue six is found to be relevant. At this third region, one can see the advantage of a structural classifier over a simple sequence-based model. While residues six and 30 are far apart in primary sequence, they are in close distance in the tertiary structure. The third region, thus, encloses coordinates right in the middle of the two residues. Interestingly both have been implicated to interact with sulfated tyrosines of the co-receptor N-terminus by forming hydrogen bonds [56].

Overall, there are clear differences in the electrostatic potential at the selected points between the R5 and X4 classes. The overall mean potential of the R5 loops is more negative than that of the X4 loops. In fact, on average there is no negative electrostatic potential found in the marked regions of the X4 class. Both findings are in good agreement with the electrostatic potential calculated for the extracellular regions of the co-receptor models. The CCR5 receptor has a strong positive electrostatic potential, while the CXCR4 model was found to generate a negative potential. These findings for CXCR4 have recently been confirmed by calculations on the newly available crystal structure of the co-receptor [167]. The distribution of ϕ of dual-tropic V3 loops (R5X4) is especially interesting. It seems, that the dual-tropic class is a chimera between the R5 and X4 classes. The mean electrostatic potential of the region around residue 11 is highly positive, resembling the X4 loop, while the potential around residue 25 is more similar to the R5 class. Around the hypothesized region of N-terminal co-receptor interaction the potential is weakly positive, resembling neither of the other classes in detail.

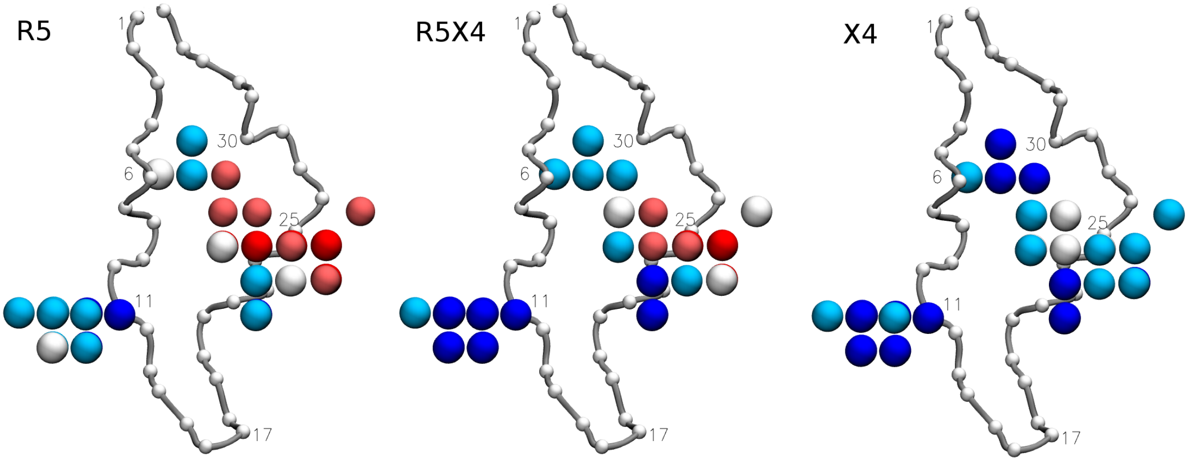

Figure 2.20: **Random forest variable importance analysis of the electrostatics model.** 5% most important positions on electrostatics hull for tropism classification by electrostatics based random forest. The backbone of the template V3 conformation [57] is shown as tube with C_α atoms marked by small beads and some residues numbered for orientation, starting with the N-terminal cysteine as residue 1. Important electrostatic features of the hull are colored according to the mean electrostatic potential $\langle \phi \rangle$ (unit $k_B \cdot 300K/e$) in the respective tropism class (red, $\langle \phi \rangle \leq -2.5$; light red, $-2.5 < \langle \phi \rangle \leq -0.5$; white, $-0.5 < \langle \phi \rangle \leq 0.5$; light blue, $0.5 < \langle \phi \rangle \leq 2.5$; blue, $2.5 < \langle \phi \rangle$).

The limited number of independent test sequences and the fact that different training data was used for the training of T-CUP, geno2pheno or PSSM make it impossible declare a front-runner. Future analyses with new sets of V3 sequences will be required to draw a definitive result, however, in conclusion, it seems that T-CUP performs at least equally well, if not better than current state-of-the art methods, like geno2pheno or PSSM.

3 Prediction of HIV-1 Co-receptor Tropism from Quasispecies

«The art of science isn't necessarily to avoid mistakes; rather, progress is often made by making mistakes as fast as possible, while avoiding making the same mistakes twice.»

– Robert Hazen

3.1 Introduction

Soon after the initial infection by HIV-1 the virus evolves by error-prone replication processes, evading the immune system, adapting to the host. The result is a diverse swarm of related, but differing virus variants, the viral quasispecies. Phenotypic tests are able to detect X4 minority variants, while conventional genotypic tests are limited by relying on single majority sequences to predict co-receptor usage. Discordance between both methods is common in a clinical setting. Next-generation sequencing methods are able to sequence representative subsets of the viral quasispecies, enabling genotypic methods to make predictions considering minor variants. The two-level co-receptor usage prediction method T-CUP was applied to test concordance between phenotypic results for twelve samples from a longitudinal phase II vicriviroc study. Moreover, the approach was found to offer deeper insight into the dynamics of co-receptor usage switching, yielding a possible description of patients with an elevated risk of switching viral co-receptor tropism.

3.1.1 The Quasispecies

The theory of the quasispecies has first been described by Manfred Eigen and Peter Schuster in the introduction of the hypercycle [35, 36, 37]. It is based on the concept of low-fidelity replication of RNA in a precellular world as formulated by Manfred Eigen in 1971 [34]. The quasispecies is defined as a dynamics distribution of different but closely related replicons [29]. The swarm-like quasispecies is scattered around a master sequence, which is usually best adapted to the environment. The theoretical quasispecies extends the classical view of population biology where one constant wild-type sequence is representative of a population. The wild-type sequence is replaced by a steady-state ensemble of closely related sequences, with varying fitness. A quasispecies develops by constant production of novel sequences, through replication errors and selection of variants. The mutant spectrum describes the composition of the quasispecies. The mutant spectrum complexity describes the coverage of sequence space and is reciprocal to the copy fidelity. It thus increases with a decreasing copy quality.

RNA virus population dynamics have been shown to be closely related to theoretical quasispecies behavior. Sources of erroneous replication in most RNA viruses are the reverse transcriptase, translating viral RNA into DNA, and the subsequent integration into the host genome by the integrase. The combined error rate of RNA viruses is around 10^{-4} per base and replication cycle. Thus, with a viral genome of around 10^4 bases, approximately one error is generated in each new virus [17]. In addition to the intrinsic error source of viral replication, recombination events can take place in the case of infection of a host cell my multiple virus variants. Driven by the constant pressure induced by changing environmental factors, such as the host immune system, a diverse viral quasispecies can quickly emerge. The error rates of RNA viruses are close to the threshold of a so-called error catastrophe. With a rising error rate, the frequency of the master sequence decreases until the rate reaches a threshold of around 10^{-3}, at which point the mutant spectrum becomes so diverse, that the conservation of single variants is very unlikely. This insight has been exploited to successfully eradicate viral populations by raising the error rate to critical levels through mutagenic nucleosides [97]. It has been proposed that viral quasispecies have a memory, in form of minority variants, for genomes that once dominated the populations [122]. *In vivo*, this is facilitated by an archive of integrated provirus reservoirs, unaffected by antiretroviral therapy [12]. Indeed, the reemergence of variants that have been successfully controlled by ART can be witnessed, once the drug is omitted [63, 8].

3.1.2 Next-Generation Sequencing

For decades, the automated Sanger sequencing was the standard method for determining DNA sequences. It is based on the original sequencing method by random chain termination events proposed by Frederick Sanger [127, 128] (Figure 3.1). To determine the sequence of a single stranded DNA, the template is distributed into four tubes, together with a DNA primer, DNA polymerase and fluorescently labeled dNTPs (dA, dG, dC, dT). Finally, ddNTPs, necessary for chain termination, are added, one type (ddA, ddG, ddC, ddT) for each tube. The ddNTPs are dNTPs, lacking the 3'-OH-group, required for bond formation with the subsequent nucleotide. During the elongation reactions in each tube ddNTPs will randomly be used, terminating the chain. Thus, a more or less equal distribution of resulting chain lengths can be expected. By means of gel-electrophoresis, the chain fragments of each tube can be separated by size, one lane for each tube.

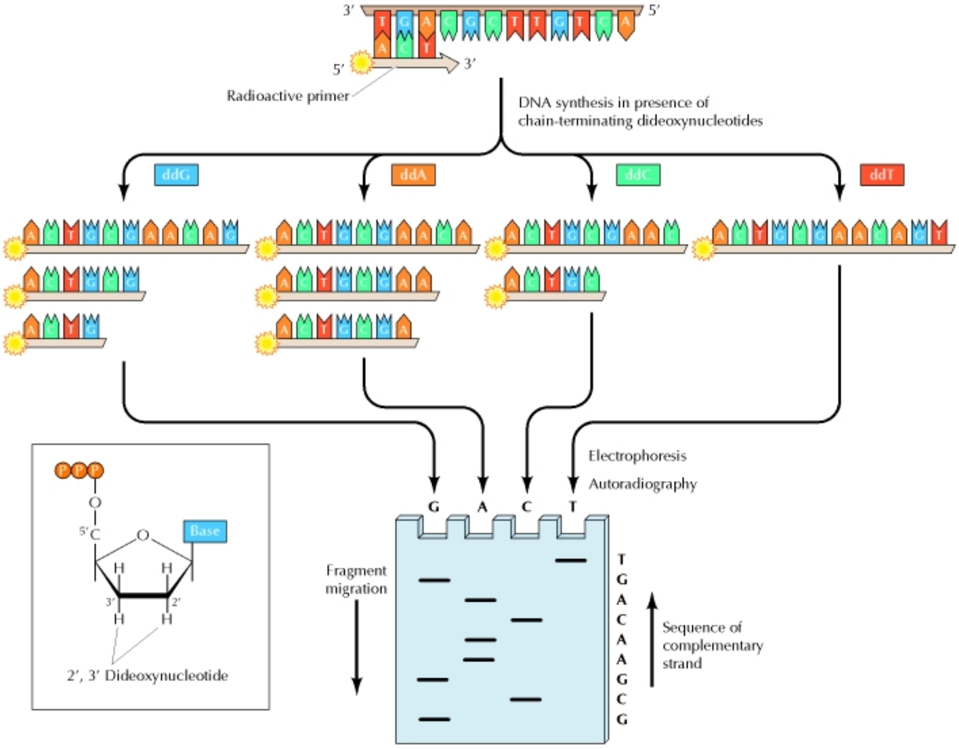

Figure 3.1: **Sanger sequencing.** Illustrations taken from Cooper and Hausmann [20].

The automated Sanger sequencing applies fluorescent dyes at the ddNTPs, one color for each type of ddNTP, reducing the number of parallel reactions required [141, 140]. Instead of four parallel reactions chain fragment are generated in a single reaction. The type of ddNTP terminating the fragments can be visualized in an automated manner. Sanger sequencing has been a key to many discoveries and scientific breakthroughs such as the first sequencing of the complete human genome [64]. However, a major limitation of the method lies in the length of sequences that can be determined. Clear separation of fragments after approximately 700 bases is difficult and results in decreasing signal quality. For the determination of whole-genome sequences this means splitting the genome DNA into small fragments, using each as a template for sequencing. This approach is referred to as shot-gun sequencing. A tedious and expensive task using Sanger sequencing.

The Sanger method was regarded as a 'first-generation-sequencing' technology [94]. Recently the emphasis has shifted towards automated, highly parallel sequencing methods, able to determine sequence of large genomes in less time and at lower price. Next-generation sequencing (NGS), technologies combine methods of sequence template preparation, sequencing and subsequent data analysis. The aim is the automated, cost-efficient generation of huge amounts of sequence data, to be used for whole-genome sequencing, transcriptome and sequence variation analysis. Already, NGS has made a huge impact on various fields of biology, especially on the field of whole-genome sequencing. The sequencing of J. Craig Venters personal genome by automated Sanger cost around 70 million US dollars. Sequencing of the genome of James D. Watson by Roche/454 cost a comparably small amount of one million US dollars. Other genomes have been determined by as little as 250.000 to 48.000 US dollars [94]. A handful of technologies are currently commercially available: GS FLX Titanium by Roche/454, SOLiD 3 by Life/APG, and the Genome Analyzer II by Illumina/Solexa. The technologies all rely on the generation of many short DNA fragments, so called reads. They differ in aspects such as the number or length of reads produced, run-time, or error-rate. The methods of how reads are generated and sequences are determined also varies between the different platforms. In this work, sequences generated by Roche/454 have been used and, thus, only the methods of this particular technology will be discussed here.

The sequencing technology initially developed by 454 Life Sciences [91] consists of three phases: preparation of template sequences, sequencing and imaging, evaluation of imaging data. A schematic overview is shown in Figure 3.2. Thousands of DNA fragments can be sequenced in parallel. The source of DNA fragments to be sequenced depends on the experiment. For whole-genome sequencing the genomic DNA is randomly sheared into single DNA fragments. In the case of variant analysis, where the variability of one or more genomic regions is to be determined, regions of interest are extracted by specific primers targeting the flanking regions. The extracted DNA regions can then be used for sequencing and amplification. Central to sequencing process is pyrosequencing. Pyrosequencing is a bioluminescence method, measuring light that is emitted during each extension step. The amount of light emitted is very small, and thus, it is necessary to produce many extension events occurring at once. This is ensured by amplifying the

DNA fragments to be sequenced by emulsion PCR (emPCR) [31]. For this, adapter sequences containing universal primer sites for amplification and sequencing, are ligated to the ends of the single stranded templates. Via these adapters the templates bind to beads under conditions favoring one template per bead. Following the attachment of templates, each bead is captured within a water droplet in an oil phase, so called microreactors, additionally containing primers, dNTP, and DNA polymerase. In these microreactors, parallel DNA amplification is performed, producing around 10^7 copies of the template attached to a single bead. After successful amplification, the emulsion is broken and the beads placed within small wells along a fibre optic glass slide, a PicoTiterPlate (PTP). The well radius restricts the occupation of beads to one per well. The preparation process is illustrated in Figure 3.2A. The beads placed in the wells, covered with millions of identical DNA templates are now subjected to pyrosequencing. For pyrosequencing, a large number of considerably smaller beads, covered with luciferase and sulphurylase as well as polymerase are added to the wells. In an iterative process dNTPs of the same type are now added to the wells. Upon incorporation, a light is emitted and recorded by a CCD camera, attached beneath the fibre-optic slide (Figure 3.2B). The high resolution camera is able to specifically record signals for each well. The signals of each well are translated in a flowgram (Figure 3.2C). This flowgram signal is then normalized to remove background noise. The x-axis shows the nucleotides that were incorporated, based on the color of the light emitted. The number of times each nucleotide was subsequently incorporated during dNTP flow can be read from the signal intensity, shown on the y-axis.

3.1.3 Tropism Prediction and the Viral Quasispecies

Over the years, several groups have reported lacking accuracy of genotypic tropism prediction methods in a clinical setting [87, 139, 88, 129], or rather an unsatisfying rate of discordance between the results of genotypic and phenotypic methods. This phenomenon can readily be explained by the viral quasispecies. By means of NGS methods, groups have described the existence of hundreds of unique virus variants within single patients [151, 121, 1]. While tropism prediction is usually performed on a single sequence derived from the patient by amplification and determination of the most prominent virus variant in a patients quasispecies, phenotypic methods, such as Trofile, will consider the whole quasispecies. For the same patient genotypic and phenotypic methods can produce discordant results if the most prominent variant is indeed an R5 virus, but the mutant spectrum includes X4 variants in sufficient concentration to be picked up by

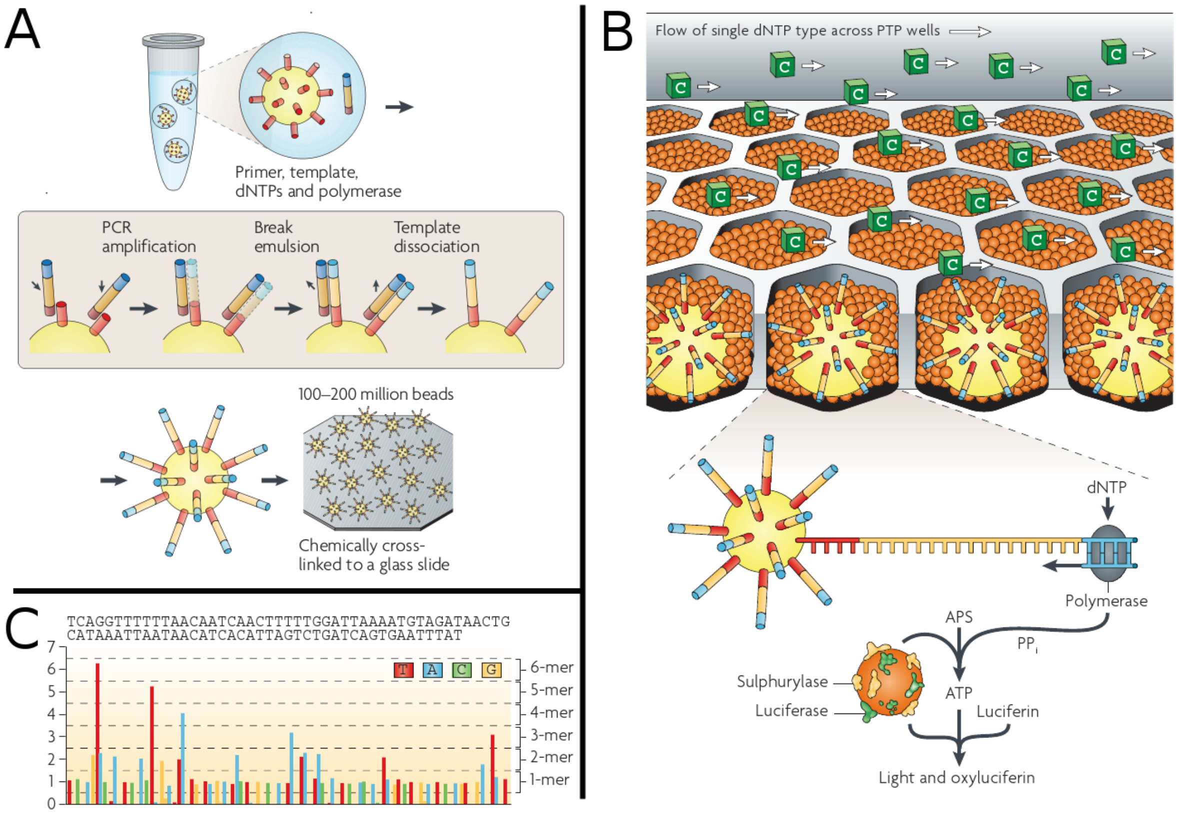

Figure 3.2: **Parallel sequencing method of Roche/454 GS FLX Titanium.** Illustrations taken from Metzker 2010 [94].

the phenotypic assay. A straightforward strategy to improve detection rate of minority strains is to perform multiple sequencing rounds. Triplicate sequencing offers an improved detection rate but increases laboratory work and sequencing costs. This strategy succeeded in improving the concordance of genotypic and phenotypic methods [145]. In the same study, the authors reported that independent triplicate amplification from the same sample produced differing sequences and tropism prediction results in 19% of the cases. This especially applied to sequences classified as X4 by the Trofile test. In that case, almost half of the samples yielded at least one sequence that was classified as R5 by the phenotypic test. The overall sensitivity could be improved by nearly 20%.

Triplicate sequencing improves the prediction accuracy of genotypic test by an increased detection rate of minor variants. Next-generation sequencing methods offer an even more robust and reliable way to pick up minor virus variants from patient samples by so-called ultra-deep sequencing (UDS). UDS allows the detection of mutations expressed at very low levels, by sequencing many variants of specific fragments of DNA or RNA found in samples. In 2010 Vandenbroucke *et al.* [156] have proposed a genotypic tropism prediction scheme based on UDS data acquired from patients failing phenotypic tests. Phenotypic tests have been known to fail, i.e. being unable to produce any results, in more than 15% of the cases [114], often caused by low viral load. The authors proposed to sequence the V3 loop region of the patients viral quasispecies to be used for genotypic prediction. To exclude ghost variants, caused by infrequent sequencing errors, they only considered variants with a frequency of more than 1% of the sequences. For their prediction, the authors used geno2pheno [136] and PSSM [65] via the respective web services. In total, 14 samples were sequenced and predicted, finding concordance with the phenotypic results in 12 of the cases. In a similar study, Swenson *et al.* analyzed the potential of UDS data gathered by 454 sequencing and found concordance with phenotypic tests in all twelve samples [145]. A year earlier, Tsibris *et al.* had sequenced the V3 loop area of the quasispecies of four patients under treatment with the co-receptor antagonist vicriviroc [151]. They focused on the analysis of the quasispecies composition without performing extensive genotypic tropism analysis. Using ultra-deep sequencing, they found experimental evidence that minor variants responsible for co-receptor switching were not generated *de novo* during one replication cycle, but had already existed at treatment start. Furthermore, they could highlight the clinical relevance of minority variants present at frequencies of less than 1%, a concentration at which these variants were not detected by the standard Trofile assay. The standard Trofile assay was able to detect X4 viruses down to a concentration of 5-10% [160] and has recently been replaced by the Enhanced Sensitivity Trofile Assay (ESTA), detecting X4 viruses at concentrations as low as 0.3% [149]. The two-level tropism prediction method described in chapter 2, was shown to perform better than standard current state-of-the-art methods on an independent testset, consisting of clinical samples containing discordant calls by phenotypic and genotypic tests. In order to test the accuracy of our method when evaluating not a single sequence but a set of sequences, representative for the whole viral quasispecies of a patient, we applied it to the sequence data deposited by Tsibris *et al.* This longitudinal study, provides an insight into the evolution of the viral quasispecies under drug pressure. The authors examined four patients during the course of treatment with the co-receptor antagonist vicriviroc (VVC). All patients were enrolled in a phase II clinical trial of VVC. In the initial clinical trial, phenotypic testing of patient tropism was performed at various times. These patients were chosen specifically, because they had a common feature: they all developed virologic failure, defined as failing to achieve and maintain a reduction by >1 log_{10} of viral load at or after week 16. For each of the patients, the authors determined the viral quasispecies by 454 ultra-deep sequencing, at three distinct times. The first sequencing was performed prior at treatment start (week 0) to establish a baseline. The second and third times of sequencing were chosen independently for each patient, based on the first finding of CXCR4 viruses by phenotypic testing and finally when virologic failure occurred. In case CXCR4 viruses could not be detected prior to virologic failure, the next point of available plasma was used for sequencing. To us, this study presented the opportunity of evaluating the performance of our method when using ensembles of V3 variants, also including minority variants, rather than using single sequences based on population majorities.

3.2 Materials and Methods

V3 Deep Sequencing Data V3 deep sequencing data was publicly available from the NCBI Short Read Archive (accession numbers SRS000811-SRS000829), were it was kindly deposited by Tsibris *et al.* [151]. The source of this longitudinal data were four patients under vicriviroc treatment. The V3 region of each patients quasispecies was sequenced on three occasions, resulting in twelve unique quasispecies compositions. Sequencing was done using the 454 GS FLX sequencing platform by Roche (Roche, Palo Alto, CA). Detailed information on primer design, preparation and further technical detail can be found in the original publication. To extract the V3 sequences from each sample, reads were aligned in forward and reverse manner against the reference HXB2-V3 strain by Smith-Waterman local alignment [142]. Reads shorter than 105 nucleotides were not considered. The resulting sequences were then translated into aminoacid code and grouped by sequence, counting the number of occurrences.

Read Frequency Cutoff In order to estimate the error rate of the PCR amplification carried out during deep sequencing, Tsibris *et al.* performed validation runs using an artificial quasispecies. This ensemble consisted of three unique clones from of the patients with frequencies of 0.89, 0.1, and 0.01. They report that this ratio was essentially preserved after amplification and post-processing. However, 4.5% of the newly amplified variants showed mutations with regards to the input sequences. The vast majority of errors were single-nucleotide mutations (99.8%). The per-nucleotide error-rate was estimated to be 0.0011 and 0.0016 in two independent experiments. This error-rate results from a combination of amplification and the applied deep sequencing protocol, after filtering of problematic sequences. We use this error-rate estimation to determine a cutoff for filtering out minority variants that might have emerged due to experimental errors, rather than as consequence genomic variation. The question we raise is: what is the minimal amount of identical sequences that has to be present in order to accept this variant as a true variant? This was answered by using binomial probabilities.

If mutations are assumed to occur independently and with the same probability p along the sequence, the probability of having exactly k errors on a nucleotide sequence of a given length n can be calculated as a binomial probability:

$$P(n,k) = \tbinom{n}{k} p^k (1-p)^{n-k}$$

Using a mean V3 nucleotide sequence length of 105 bases and the per-nucleotide error-rate of 0.0016 we calculate the probability of having exactly one error by

$$P(105,1) \approx 0.14$$

In turn, the probability of encountering the same error in a second sequence, originating form the same source sequence, can be calculated by

$$p = \tfrac{P(n,k)^2}{n} = \tfrac{P(105,1)^2}{105} \approx 0.0002$$

Note that this is a conservative estimate since different nucleotides are not considered. The calculation can be extended to estimate the probability of finding the same error in m sequences, originating from the same source by

$$p = \tfrac{P(n,k)^m}{n^{m-1}}$$

An erroneous sequence can have more than a single mutation, thus the probability is expressed as the sum of the probabilities of all possible parallel mutations ($n = 105$):

$$p = \sum_{k=1}^{n} \frac{P(n,k)^m}{n^{m-1}}$$

We calculated the probabilities of different variant frequencies $m = \{1, ..., 5\}$ to occur by chance. These probabilities multiplied by the maximum amount of reads produced by a single deep sequencing experiment (c) gives number of false-positives (variants resulting from sequencing error that are being accepted) to be expected per experiment. The results are visualized in Figure 3.3. Choosing a cutoff of at least four observations of a variant in order to be accepted as a true variants seems a reasonable choice, as it is estimated to produce less than one false positive in one of 10^4 UDS experiments, each generating 10^5 reads. In the rare case where a false positive is produced, it is likely to be an infrequent variant with very little impact on the overall quasispecies structure.

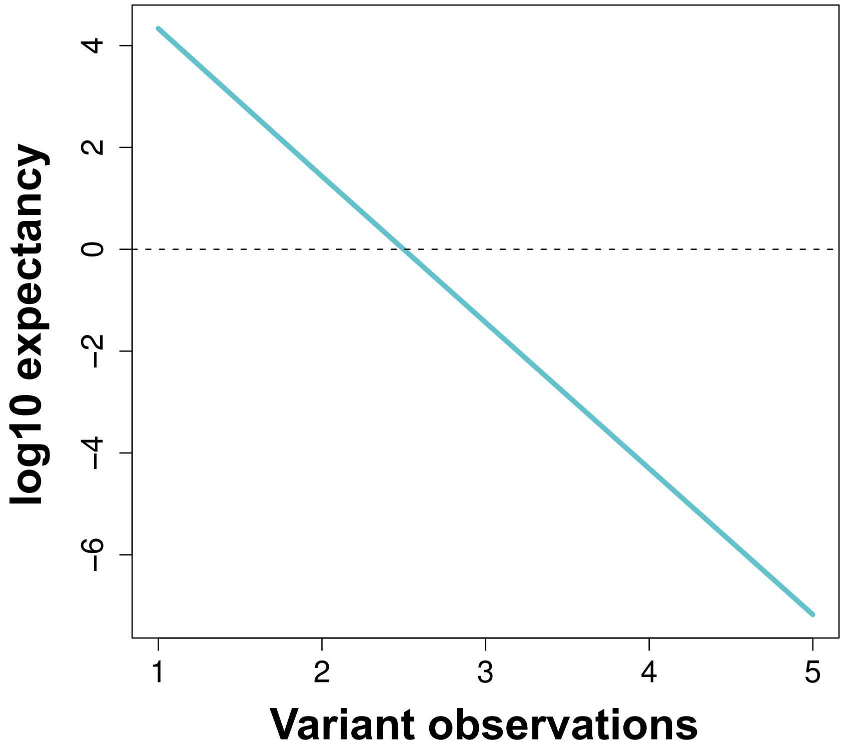

Figure 3.3: **Expected number of falsely sequenced variants per quasispecies.** The number of falsely accepted variants, resulting from sequencing error per UDS experiment decreases exponentially with an linearly increasing variant cutoff. Choosing a cutoff of at least four variants results in less than one falsely accepted variant in around 10^4 UDS experiments, each producing 10^5 reads.

Distance Calculation A prediction in the two-dimensional prediction plane consist of two elements: the single predictions made by the electrostatic and hydrophobicity models. It is of the form

$$p = \begin{pmatrix} p(ESP) \\ p(hydro) \end{pmatrix}$$

Distances between predictions in the prediction plane are treated as Euclidean distances, calculated by

$$d_{prediction}(p1, p2) = \sqrt{(p1_{p(ESP)} - p2_{p(ESP)})^2 + (p1_{p(hydro)} - p2_{p(hydro)})^2}$$

Alignment distance calculations are based on pairwise local alignment using the Smith-Waterman algorithm [142] and the BLOSUM62 matrix. Alignment distances represent a sequence similarity score normalized between 0 and 1 by

$$d_{alignment}(s_1, s_2) = 1 - \frac{A(s_1, s_2)}{A(s_1, s_1)}$$

. The minimum alignment distance of zero is achieved by aligning a sequence to itself.

Co-receptor Usage Prediction Prediction of co-receptor usage of V3 sequences was performed using T-CUP as described in Chapter 2, applying the random forest second-level classifier. The cutoff between X4 and R5 classes was set to 0.02.

3.3 Results

3.3.1 Prediction of Tropism Based on the Viral Quasispecies

After downloading the sets of reads, we extracted probable V3 sequences and discarded variants with a frequency of less than four reads per sample. Filtering scheme and read cutoff have been described in detail in the Materials and Methods section. For every of the twelve sets we predicted the tropism of each unique sequence present. The number of unique sequences analyzed from each sample is shown in Table 3.1. For each sample we first calculated the number of sequences predicted as X4, relative to the complete sequence ensemble. For each patient we thus received three values, which could be used to draw a trend of X4 population development over the course of treatment. These fractions were then related to the X4 fraction of the viral load, which was taken from the original publication by Tsibris *et al.* The progressions of the X4 fraction of the quasispecies and the X4 fraction of the viral load are shown in Figure 3.4. Overall, we were able to reproduce the results of Trofile testing in all of the cases where sequence data of the viral quasispecies was available together with a Trofile result. This was the case in 75% of the samples. In the remaining three cases, we extrapolated linear trends of the X4 fraction progression, which were in complete agreement with Trofile results prior and/or after the times of sampling. Interestingly, we could reproduce the X4 detection level of the standard Trofile assay, which detects X4 viruses at fractions of 0.1 of the total population [160]. This level was reproduced in all nine cases, when a Trofile result was provided with the quasispecies sequences.

At therapy start, all four patients were confirmed negative for majorities of X4 viruses (week 0) by the standard Trofile assay. We predicted a fraction of $\leqslant$0.01 of X4 viruses

Table 3.1: **Unique V3 sequences extracted from quasispecies samples.** Number of unique sequences used in the analysis for each patient and each of the three sample times. Based on the data provided by Tsibris *et al.*, a cutoff of a minimum of four reads per sequence was applied to limit the number of spurious sequences.

Patient	time 1	time 2	time 3
07	174	112	86
18	240	112	41
19	148	134	104
47	126	84	78

in each of the patient samples. From there, patient quasispecies diverged. Subject 07 does not develop a detectable X4 population until week 19, according to Trofile. A close examination of the quasispecies sequenced at weeks 12 and 19, reveals an increasing fraction of X4 viruses, of 0.097 and 0.76, respectively. It is noteworthy that Trofile missed to capture the X4 population at week 12. The upward trend, however, is clearly visible by genotypic prediction considering the quasispecies. Subject 18 quickly developed a major X4 population with more than 99% of viruses being predicted as X4. This high level of X4 viruses is maintained until the last time of sequencing at week 16, and X4 tropism is confirmed by Trofile testing at weeks 8, 12, 20, and 23. The development of Subject 19 is particularly interesting. Phenotypic testing detects X4 viruses at weeks 2 and 10 but not at week 22. The predicted X4 fractions for samples at weeks 2 and 17 indeed reproduce this trend at 0.49 and 0.12, respectively. Unfortunately, no sequencing has been performed at week 23, but a linear extrapolation of the X4 fraction progression, predicts the population of X4 viruses to fall below Trofile detection threshold of 0.1. According to Tsibris *et al.* this patient had a poor adherence prior to experiencing virologic failure. Subject 47 never developed any strong X4 population, neither Trofile nor the genotypic testing performed on our side revealed any major X4 strains. According to the authors, the reason for this development remains unclear, however a failing antiretroviral regimen is possible, even though a loss of VVC susceptibility could not be demonstrated.

The analyses so far have shown that results by standard Trofile testing can be reproduced by genotypic prediction of quantitative deep sequencing data. In fact, the case of subject 07 hinted that this method is even superior to standard Trofile in terms of detection of minor X4 populations. To further evaluate the potential and relevance of UDS in clinical practice, we performed more detailed analyses of the quasispecies structure. In a

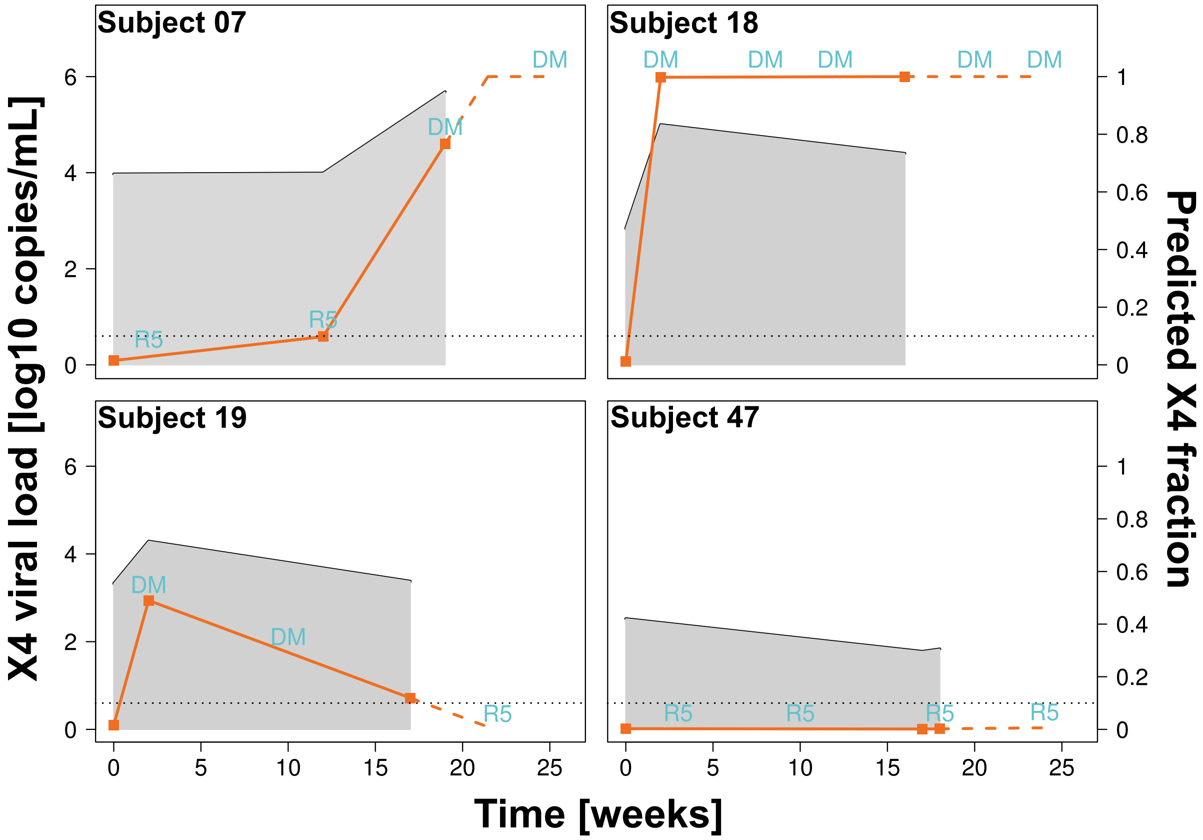

Figure 3.4: **Development of predicted fraction of X4-using viruses in four patients during vicriviroc treatment.** Labels "R5"(R5-using) and "DM"(dual/mixed or X4-using) are Trofile predictions of the patients' quasispecies. Predicted relative X4 fraction from patient samples are marked as orange squares and annotated by the right axis. The gray areas show the calculated viral load, based on the predicted X4 fraction and viral loads published by the Tsibris *et al* [151]. The dotted lines mark the 10% detection rate of standard Trofile assay.

first step, we took a closer look at patients developing a co-receptor switch and a major X4 population. Tsibris *et al.* had already reported that patients 07, 18, and 19 carried minor X4 variants at baseline (week 0), which later developed into major populations. The authors used the publicly available PSSM tool to predict co-receptor tropism from sequences. Using our prediction method, described in chapter 2, we were able to confirm these findings. Table 3.2 shows the progression of the single most prominent X4 variants of patients identified at week 0. All of these variants, with the exception of subject 47, develop into dominating variants under drug pressure, despite initially representing only fractions of 0.005 to 0.01 of the quasispecies. In subject 07 the most prominent variant alone accounted for a considerable viral load of over 6000 copies/mL. In subjects 18 and 19, the most often occurring X4 variant was not represented as frequently as that found in subject 07, which is even more alarming, considering their proliferation under drug pressure. The most frequent X4 variant at week 0 of subject 07 accounts for only a fraction 0.0009 of the total population, the viral load of that variant consists of around 100 copies/mL. In fact, it has completely vanished by week 17. The most frequent baseline X4 variants of subjects 18 and 19 are present at around 300 and 900 copies/mL. Figure 3.5 shows the correlation of baseline X4 variants and the overall X4 population. The progressions of the most frequent baseline X4 variant and the overall X4 population are highly correlated. Pearson correlation testing, omitting subject 47 due to vanishing of the variant, yields a correlation of $R^2 = 0.975$ with $p = 6.7 \cdot 10^{-7}$ for the null hypothesis of zero correlation. The close to linear correlation depicted by the dashed line in Figure 3.5 shows that minor baseline X4 variants are responsible for the long-term overall population development.

Table 3.2: **Development of most prominent baseline X4 virus strains.** Development of the X4 seed strains in subjects 07, 18, 19 with tropism switches and the largest initial X4 strain in patient 47 who does not show a tropism switch. The three times are the sampling points along the "Time" axis in Figure 3.4.

Patient	Variant	Fraction of population at		
		time 1	time 2	time 3
07	CTRPGNNTRRSIRIGPGQTFFAREDIIGDIRQAYC	0.01	0.07	0.73
18	CERPNNNTRQRLSIGPGRSFYTSRRIIGDVKKAHC	0.005	0.79	0.71
19	CTRPNNNTRKGIYLGPGRAFYTTDKIIGDIRQAHC	0.007	0.43	0.08
47	CTRPNNSTRKSINIGPGSAWYTTGDIIGDIRQAHC	0.0009	0.0	0.0

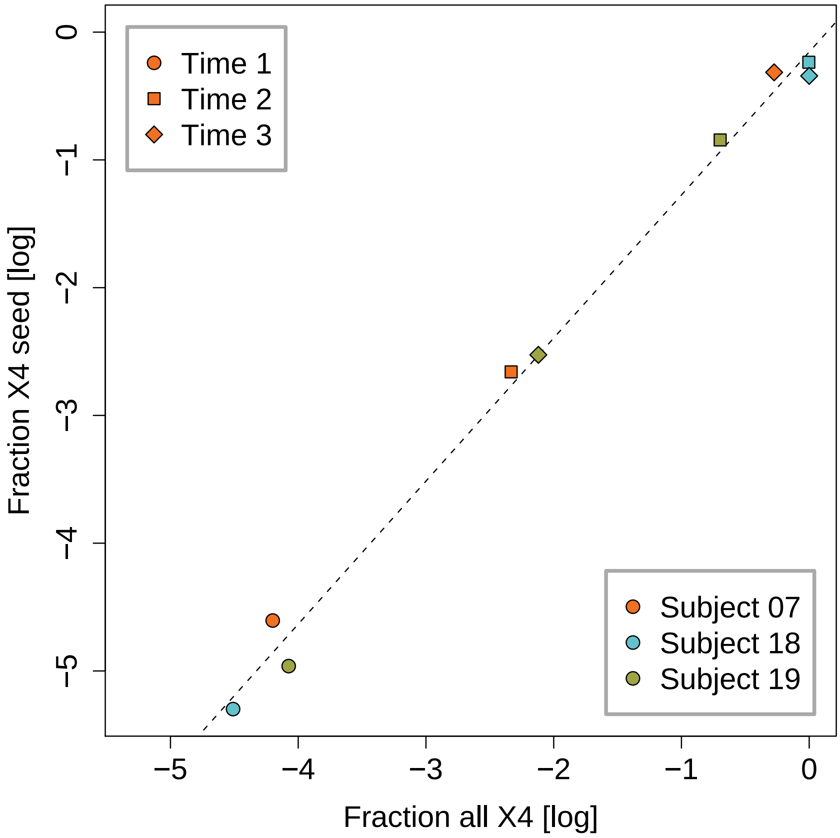

Figure 3.5: **Correlation of most prominent baseline X4 strain and total X4 population.** The fraction of the most prominent baseline X4 strain against the total X4 fraction is shown for all patients (coded by color) developing a tropism switch and time points (coded by symbols).

3.3.2 Structure of Viral Quasispecies as Indicator for Tropism Switches

T-CUP relies on two voting results made by independent models (see Section 2.3.3). The two outcomes are then combined by a second-level model to reach a final decision. In the above analyses, V3 sequences were classified by the final classification of this two-level approach. The occasional discordance between the independent first-level outcomes were discussed and introduced as justification for the second-level model. Figure 2.14 visualizes this phenomenon. In these “prediction planes“, V3 sequences are located according to the numerical outcome of the independent first-level outcomes. These represent a pseudo-probability of being an X4 virus. Thus, sequences with a clear X4 prediction by both models, hydrophobicity and electrostatic potential, are located in the upper right quadrant of the prediction plane. In the lower left corner, one finds sequences, with a concordant R5 prediction. The two remaining quadrants contain sequences where the two first-level models disagree. The quadrants result from an arbitrarily chosen cutoff below which sequences are interpreted as R5 variants. Using the cutoffs chosen here, each model had a specificity of 90%, or false positive rate of 0.1, correctly classifying 90% of the R5 sequences during crossvalidation of the training data.

Here, we analyzed the patients quasispecies structure by positioning each unique sequence of a patient into the prediction plane. Figure 3.6 (page 93) shows the prediction planes containing the quasispecies at subsequent time points (left to right) of patients 07, 18, 19, and 47 (top to bottom). The number of occurrences of each sequence is encoded by a color key. Sequences with more than 3, 10, 100, or 1000 occurrences are represented as dots, colored in red, orange, yellow, and white, respectively. The most frequent baseline X4 variants are marked by green arrows. At week 0 sequences of all patients are still largely clustered in the lower left corner, visualizing the predominance of R5 variants. Over the course of treatment, the quasispecies of subject 07 detaches from the R5 corner and slightly moves towards the cutoffs of the first-level predictions. One can witness the R5 population dividing into two distinct clusters, starting at week 12. This process is accompanied by a thinning-out of the initial R5 population directly attached to the lower left corner. By week 19, the newly developed cluster is clearly visible, while the initial R5 population has largely disappeared. Although many variants are located within the R5 quadrant, defined by the first-level predictions, the second-level model classifies most of these as X4 variants due to its relatively conservative cutoff of 0.05. At this strict cutoff level, the model tends to classify variants without a clear R5 vote from both the hydrophobicity and electrostatic potential models as X4 variants. The quasispecies progression of subject 18 is much clearer. The X4 population, already visible at week 0 in the upper right corner, grows rapidly within the first two weeks of treatment, while the R5 population quickly disappears. By week 16, the population consists of X4 viruses almost exclusively. Subject 19 clearly develops a major X4 population by week 2. Like in subject 07, this process is accompanied by the development of a separate X4 cluster. This quasispecies nicely demonstrates a reciprocal correlation between R5 and X4 populations. While the R5 population decreases at week 2, the X4 population increases. At week 17 this development is reversed, with a shrinking X4 population and a reemerging R5 population. This quasispecies actually undergoes two co-receptor switches within 20 weeks. Subject 47 never shows any substantially populated clear X4 variants.

The prediction planes helped in visualizing the clustering process of X4 and R5 populations during co-receptor switches and the reciprocal correlation between the two pop-

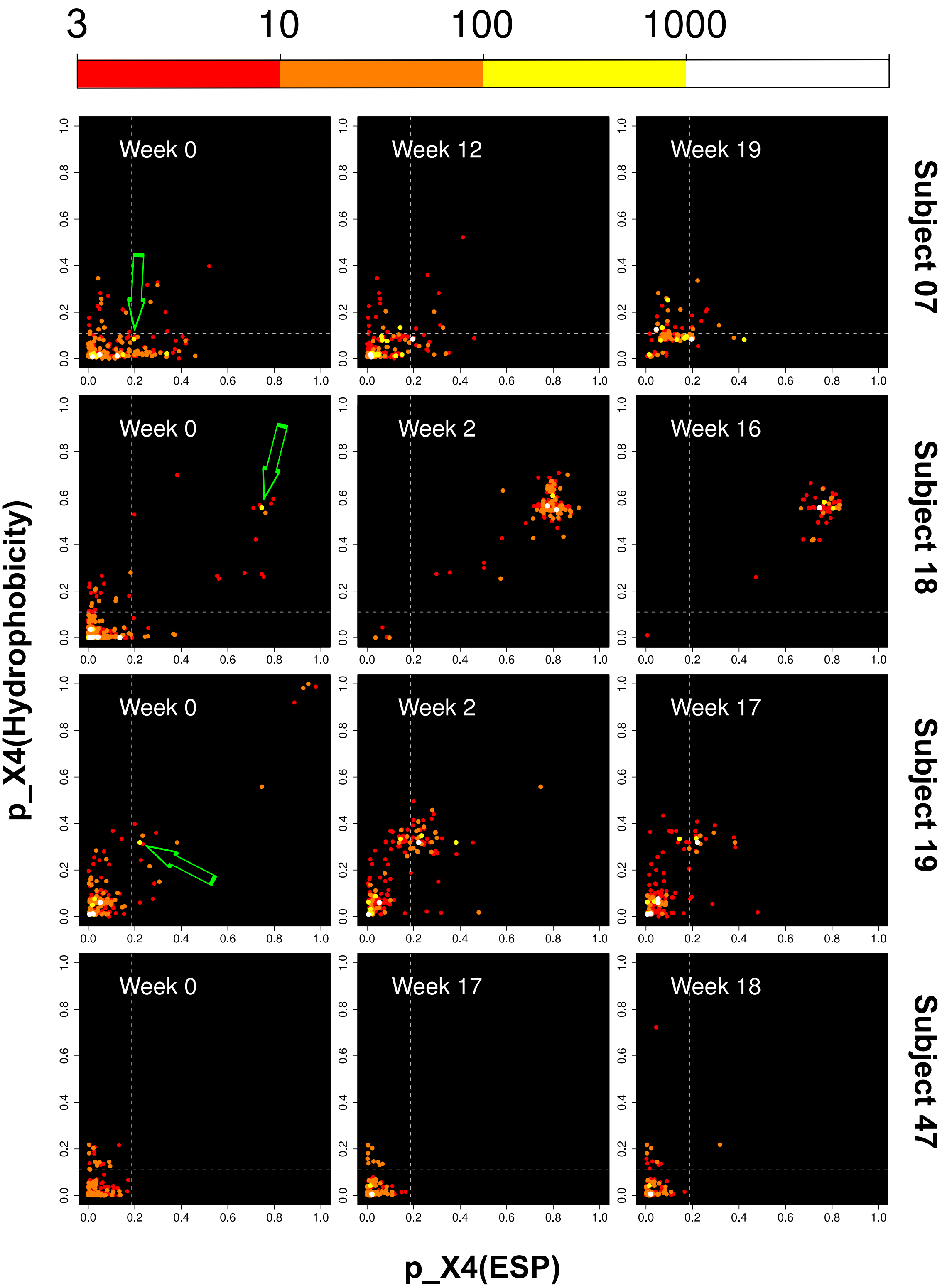

Figure 3.6: **Quasispecies development of patients over time.** Virus population development of patients (top to bottom) over time (left to right). The axes show independent predictions of the first level of T-CUP, based on electrostatic potential and hydrophobicity. Dashed lines represent cutoff levels between R5 and X4 classes. Cutoffs were chosen at 90% specificity in the training set.

ulations. The clustering process was observed in all of the cases where a co-receptor switch occurred. In fact, it might serve as an early indicator for a switch occurring in the near future. While at week 0 the quasispecies are still dominated by R5 viruses, baseline X4 variants responsible for co-receptor switches in patients 07, 18, and 19 are already visible and can be clearly distinguished from other minor variants (Figure 3.6, green arrows). Specifically, these variants share two important features: first, the have reached a critical number, as the color code clearly distinguished them from other minority variants; second, because of their number, these variants are serving as seed for new variants by producing offspring. Looking at the quasispecies in the prediction planes, one can see distinct variants close to the X4 seed variants. This can be seen for all three dominating baseline X4 variants of patients 07, 18 and 19. The assumption that a close distance on the prediction plane can be interpreted as a close sequence relation, however is not directly clear. In order to test this assumption, we correlated the Euclidean distance of variants in the prediction plane to the actual sequence distance. Specifically, for every quasispecies we calculated the distances of all variants to the X4 seed variant of the respective patient, referred to as ”prediction distance”. In addition, for every resulting pair we calculated a sequence distance measure, referred to as ”alignment distance“ (see Materials and Methods of this chapter). The correlation of prediction distance and alignment distance for each patient and quasispecies is shown in Figure 3.7. R5 variants are shown as blue dots, X4 variants in orange. Patient 47 was excluded from the analysis, as no co-receptor switch had occurred. In each of the quasispecies plots, the seed variant, to which all other variants are compared, is found in the lower left corner. Its prediction and alignment distance to itself is obviously 0. At treatment start, the X4 population of subject 07 shares a close sequence similarity to the seed strain, the R5 population is more spread out with higher sequence distances on average. In the prediction plane, however the X4 population is more distant to the seed strain, than the R5 population. This can be understood when looking at the prediction plane and the location of the seed strain (green arrow). The seed is located very close to the cutoffs. A higher cutoff would have predicted this variant to be an R5 virus. The situation does not change at week 12, except for a marginal decline of X4 diversity. The relative fraction of X4 viruses was predicted to be just under 10% at that time. By week 19, the X4 population accounted for over 70% of the population. Accordingly, it was more closely located to the seed strain, hypothesized to be responsible for the overall X4 progression. Both, prediction and alignment distances of the overall X4 population to the seed variant have decreased during the course of treatment. Unfortunately, due to the close vicinity of the seed strain to the R5 quadrant in the prediction plane and the overall population sequence similarity, a relation between closely related offspring and the outcome of prediction could not be established. The proposed relation could, however, be shown in subjects 18 and 19. The X4 population of subject 18 was relatively diverse at treatment start, with variants reaching alignment distances from close to zero up to over 0.6, compared to the seed X4 strain. In comparison, the R5 population is clearly separated from the seed strain, showing that the V3 loop of the seed already has a considerable genetic distance to the initial R5 population. Looking at the prediction distance, a correlation with sequence distance can be seen for the X4 population. Through weeks 2 and 16 the X4 population increases and later shares a close relation in sequence and prediction outcome to the proposed seed strain. The genetic diversity of subject 19 is not as great as that of subject 18, showing smaller extreme alignment distances. As in subject 07 the overall X4 population is closely related to the R5 population when looking at sequence distances. In the prediction plane, the X4 population is composed of three clusters, which become apparent when calculating the distances from the X4 seed variant. Some variants are closely coupled to the seed strain in both alignment and prediction distance, while another group overlaps with the R5 population. An outlying third group has quite a considerable prediction distance. The prediction plane of subject 19 at week 0 tells us that this outlier group consist of a small number of X4 variants with a clear X4 prediction, while the seed variant is located nearer, but clearly outside the R5 quadrant. While the composition and position of the R5 population remains unchanged throughout treatment, the first cluster around the seed strain grows in number of variants, reaching a peak at week 2, then decreasing again by week 17. Generally, one can see in this patient that a close distance to the seed strain in then prediction plane also means a close genetic relation.

3.4 Discussion

The high prediction accuracy of genotypic prediction methods based on UDS data coupled with the ability to detect X4 minority variants are encouraging. However, the rather limited amount of samples makes further in-depth analyses necessary. The use of UDS for the clinical practice is dependent on its cost-efficiency. Currently, a single 454 run costs in the order of 10.000 dollars and takes a few days, considering preliminary preparations, post-processing of reads, and genotypic testing. In comparison, phenotypic tests like the Trofile assay cost around 1500 USD and takes at least two weeks. The availabil-

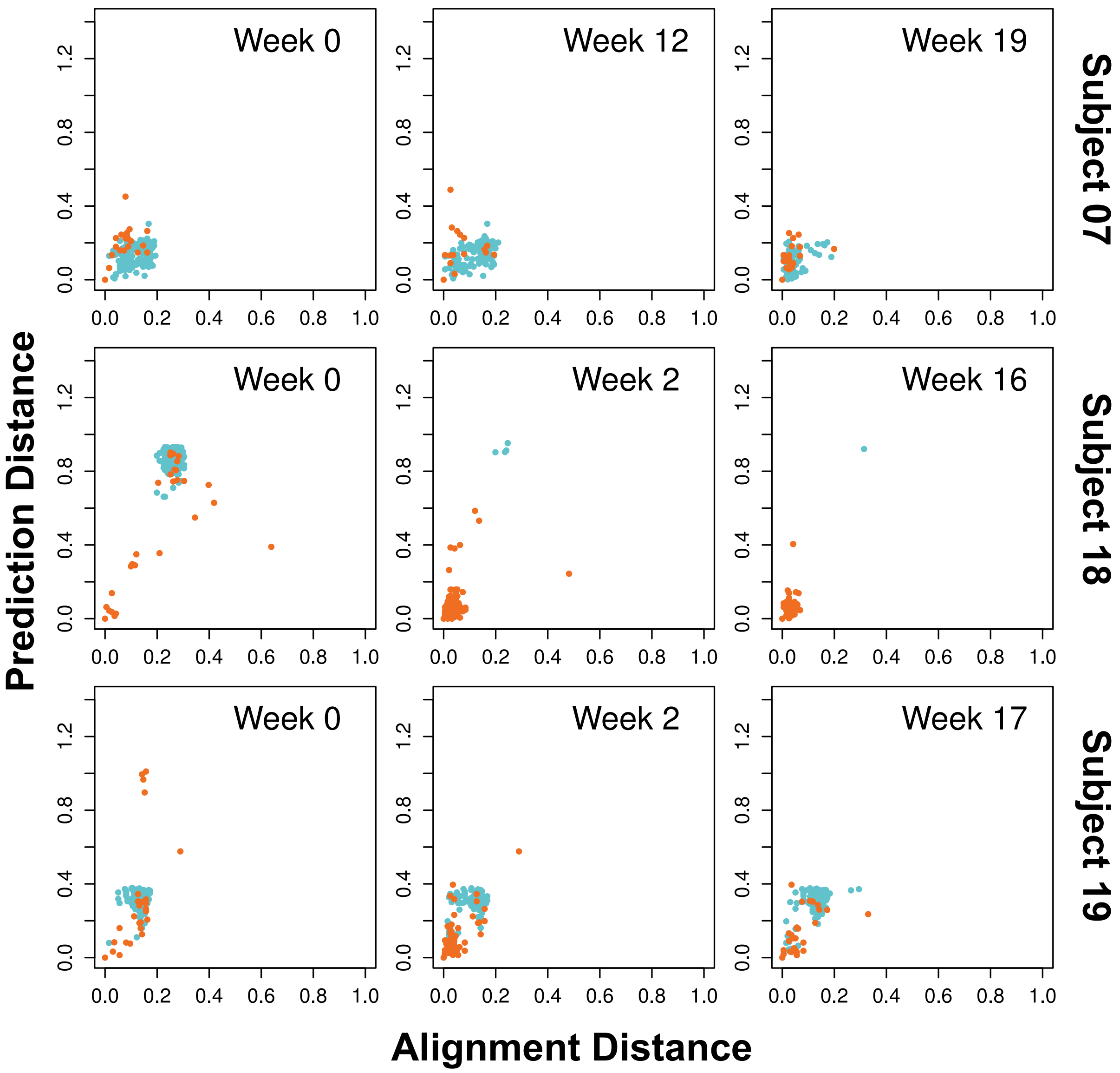

Figure 3.7: **Alignment distance vs. prediction distance.** The plots show the prediction and alignment distances of each sequence of the quasispecies to the X4 seed strain. The prediction distance is calculated as Euclidean distance between the points in the prediction plane (see Figure 3.6 on page 93). The alignment distance is calculated as described in the Materials and Methods section of this chapter.

ity of NGS technologies, required to perform ultra-deep sequencing, is still limited. Due to the broad experimental spectrum of the technology, however, a development towards a more broad availability is to be expected. The higher costs of UDS experiments, compared to phenotypic testing, does not necessarily have to limit the use of UDS in clinical practice. Most NGS technologies allow the parallel analysis of samples. Thus, samples from more than one patient can be sequenced at a time, resulting in less reads generated per sample. The genotypic prediction is then based on less reads, possibly missing minor variants and reducing the overall accuracy. To test the performance of our method based on the read coverage, we simulated sequencing runs by random sampling of reads from a quasispecies. In detail, sequencing of each of the twelve quasispecies was simulated the process by random sampling with replacement of reads, calculating the relative fraction of X4 viruses based on the sample. For each quasispecies the experiment was repeated, steadily increasing the number of randomly chosen reads by 1000, starting at 100 in the first round. For each sample size, the experiment was repeated 200 times, to calculate 95% confidence intervals for the relative X4 fraction derived from the sample of size n. The results are shown in Figure 3.8. As expected, the relative fraction of X4 viruses, calculated from the randomly chosen samples, converges with a growing sample size. It is, however, surprising how fast this convergence occurs. Our results suggest, that read coverages as low as 10.000 reads per sample, are sufficient to sample a representative subset of the viral quasispecies. In the case of 454 sequencing technology, the state of the art machine, the GS FLX Titanium Series, is able to produce around 1 million reads per run, while the more affordable GS Junior produces on the order of 100.000 reads per run. Even in the latter case, this would translate to a parallel sequencing of ten patient samples, reducing the cost per patient to around 1000 US dollars. It is to be expected that read coverages will be further increased in the future, reducing the cost per reads or patient.

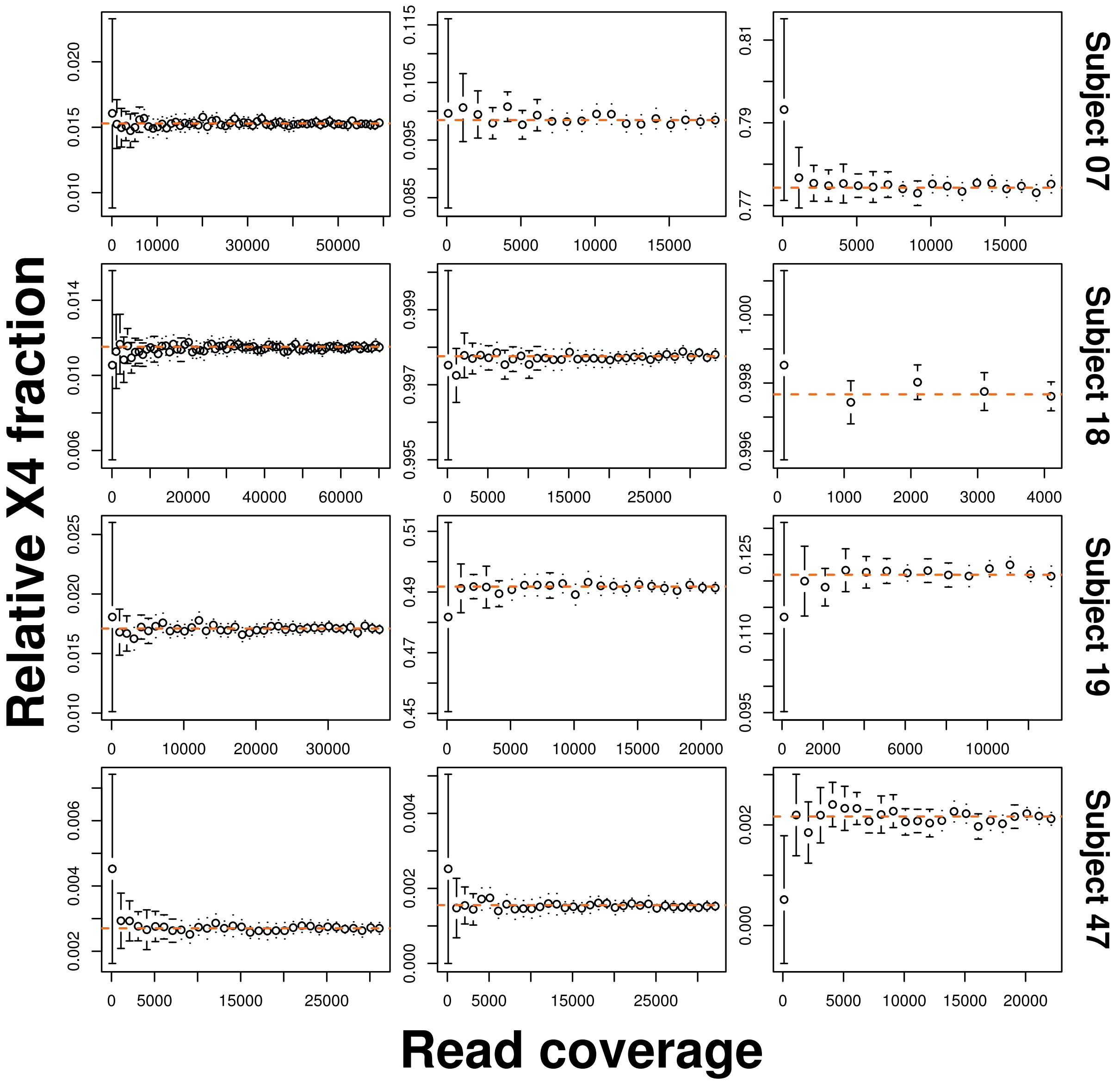

Figure 3.8: **Population reconstruction with varying read coverage** Quasispecies were resampled using an increasing amount of reads. The relative fraction of X4 viruses was then calculated from the generated samples. Results quickly converge to the predicted X4 fraction using all sequences, shown by the orange line.

4 Summary and Outlook

> *«You gotta have fun. Regardless of how you look at it, we're playing a game. It's a business, it's our job, but I don't think you can do well unless you're having fun.»*
>
> – Derek Jeter

4.1 Summary

In this work, a new method to predict the HIV co-receptor tropism from HIV genotype was developed. The method comprises a two-level random forest machine learning approach, generating two independent predictions based on the electrostatic potential and the hydrophobicity of the V3 loop of HIV envelope protein gp120. The final prediction is reached by combining the first-level votes. This was done using different combination methods: the simple mathematical functions *min* and *max*, multiplication, and training of an additional random forest model. First results showed superiority of the random forest over other methods, but a blind test set of clinical V3 loop sequences with known co-receptor tropism suggested otherwise. In a closer analysis, generated prediction landscapes showed indications of an overfitting effect by the second-level random forest, explaining its non-superiority over the other methods for the clinical test set. Still, the two-level approach showed an improvement over single prediction models ESP and hydrophobicity. However, prediction accuracy was not perfect. This effect of discording assay-determined phenotype and genotypic prediction could be explained by heterogeneous viral quasispecies, where minority X4 virus populations might be detected by phenotypic tests, while standard sequencing produces the V3 loop sequence of the majority R5 virus. To investigate this point, quantitative deep sequencing data sets, containing representative cross-sections of viral quasispecies were analyzed. These sets, produced by Tsibris *et al.* [151], resulted from a phase II clinical trial of patients receiving vicriviroc, a small molecule co-receptor antagonist. Plasma samples of four patients experiencing virologic failure were selected for ultra-deep sequencing at treatment start, after showing assay-detectable X4 viruses and finally after virologic failure. In this work, co-receptor usage of all V3 loop sequences of each quasispecies was predicted. A perfect concordance was found between the fraction of sequences predicted as X4 and phenotypic results, applying the detection level of the test used. Further, the analysis showed early signs of switches in co-receptor tropism, weeks before X4 viruses were detected by the phenotypic assay used throughout the trial.

The prediction method presented in this work is currently lacking a critical feature, which needs to be addressed in the future: The presence of a diverse quasispecies and resulting mix of viruses present in the plasma during sequencing, often leads to unclear base calls. A straightforward prediction on these ambiguous sequences is not possible. Other methods have introduced ways of dealing with these sequences. Both PSSM and geno2pheno are relying on a worst-case treatment, advising against treatment with CCR5 antagonists, should even a single ambiguous base call result in an X4 prediction. While this is certainly a feasible approach, the performance with different decision-making processes should be elucidated.

4.2 Outlook

The impact of an accurate determination of co-receptor tropism on therapy outcome has been highlighted on several occasions throughout this work. This emphasis will become even stronger should co-receptor antagonists prove to be valuable first-line drugs, administered as part of standard HAART. Phenotypic methods have become more sensitive. The enhanced sensitivity Trofile assay, for example, is able to detect minority variants down to 0.3% of the total virus population. Questions concerning the predictive power of genotypic methods have been addressed. Still, it is unclear whether predictions made, based on the V3 loop region alone, are already optimal. For instance, discordance between genotypic and phenotypic methods may results from bulk sequencing or features in other tropism determining regions, such as the variable loop 2 (V2) of gp120 or even regions in gp41. These are currently not sequenced in clinical practice but have been shown to improve accuracy when included into genotypic prediction [147]. Both, phenotypic assays and genotypic prediction methods based on bulk sequences are currently unable to provide information about virus population dynamics or structures. By combining next-generation sequencing with genotypic prediction methods, this feature has been addressed in recent publications [151, 33, 3]. Under the same setting, very satisfying concordance with phenotypic assay results has been described [145, 156], even offering additional aspects that cannot currently be provided by cell-based assays [33]. Whether UDS is able to make an impact on routine clinical practice in diagnostics is mostly a monetary question. Costs for technology, sequencing and infrastructure dealing with the massive amounts of data have to drop further in order to make this technology available to a broad spectrum of scientists and clinicians. The advantages for diagnostics are at hand: regular, parallel sequencing of important regions of the major HIV drug targets PR, RT, Env, IN, could reveal an advancing trend towards treatment failure. This would, for the first time, put clinicians in a position where a constant and effective suppression of viral load, even through phases of changes in therapy, seems possible. However, until then, more knowledge has to be gathered. Future research has to include determination of critical frequency thresholds of virus minorities, answering the question at which frequency or population structure of X4 viruses co-receptor antagonists should not be administered. For parallel drug resistance testing using next-generation sequencing of viral quasispecies, it will be helpful to effectively sequence only drug resistance hotspots of the respective HIV proteins, while still maintaining a high prediction accuracy. Recently, Steegen *et al.* have found good results in predicting drug resistances using only a short fragment of the HIV-1 reverse transcriptase [143]. With the introduction of integrase inhibitors, similar studies will have to performed for the integrase. In conclusion, it can be stated that the genotypic algorithms are not lacking accuracy, but are rather limited by the type of sequencing applied.

Bibliography

[1] I. Abbate, G. Rozera, C. Tommasi, A. Bruselles, B. Bartolini, G. Chillemi, E. Nicastri, P. Narciso, G. Ippolito, and M. R. Capobianchi. Analysis of co-receptor usage of circulating viral and proviral HIV genome quasispecies by ultra-deep pyrosequencing in patients who are candidates for CCR5 antagonist treatment. *Clin Microbiol Infect*, Aug 2010. 81

[2] S. Altschul, T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389, 1997. 48

[3] J. Archer, A. Rambaut, B. E. Taillon, P. R. Harrigan, M. Lewis, and D. L. Robertson. The evolutionary analysis of emerging low frequency HIV-1 CXCR4 using variants through time-an ultra-deep approach. *PLoS Comput Biol*, 6(12):e1001022, 2010. 100

[4] S. Attia, M. Egger, M. Müller, M. Zwahlen, and N. Low. Sexual transmission of HIV according to viral load and antiretroviral therapy: systematic review and meta-analysis. *AIDS*, 23(11):1397–1404, Jul 2009. 19

[5] N. A. Baker, D. Sept, S. Joseph, M. J. Holst, and J. A. McCammon. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc Natl Acad Sci U S A*, 98(18):10037–10041, Aug 2001. 48

[6] F. Barré-Sinoussi, J. C. Chermann, F. Rey, M. T. Nugeyre, S. Chamaret, J. Gruest, C. Dauguet, C. Axler-Blin, F. Vézinet-Brun, C. Rouzioux, W. Rozenbaum, and L. Montagnier. Isolation of a t-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science*, 220(4599):868–871, May 1983. 12

[7] N. Beerenwinkel, B. Schmidt, H. Walter, R. Kaiser, T. Lengauer, D. Hoffmann, K. Korn, and J. Selbig. Diversity and complexity of HIV-1 drug resistance: a

bioinformatics approach to predicting phenotype from genotype. *Proc Natl Acad Sci U S A*, 99(12):8271–8276, Jun 2002. 38, 39

[8] S. Boucher, P. Recordon-Pinson, D. Neau, J.-M. Ragnaud, K. Titier, M. Faure, H. Fleury, and B. Masquelier. Clonal analysis of hiv-1 variants in proviral dna during treatment interruption in patients with multiple therapy failures. *J Clin Virol*, 34(4):288–294, Dec 2005. 78

[9] L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996. 39

[10] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. 37

[11] J. A. G. Briggs, T. Wilk, R. Welker, H.-G. Kräusslich, and S. D. Fuller. Structural organization of authentic, mature HIV-1 virions and cores. *EMBO J*, 22(7):1707–1715, Apr 2003. 15

[12] C. Briones and E. Domingo. Minority report: hidden memory genomes in HIV-1 quasispecies and possible clinical implications. *AIDS reviews*, 10(2):93–109, 2008. 78

[13] L. C. Burkly, D. Olson, R. Shapiro, G. Winkler, J. J. Rosa, D. W. Thomas, C. Williams, and P. Chisholm. Inhibition of HIV infection by a novel CD4 domain 2-specific monoclonal antibody. dissecting the basis for its inhibitory effect on hiv-induced cell fusion. *J Immunol*, 149(5):1779–1787, Sep 1992. 20

[14] A. Canutescu, A. Shelenkov, and R. Dunbrack Jr. A graph-theory algorithm for rapid protein side-chain prediction. *Protein Science*, 12(9):2001–2014, 2003. 48

[15] D. Chen and X. Cheng. An asymptotic analysis of some expert fusion methods. *Pattern Recognition Letters*, 22:901–4, 2001. 36

[16] S. K. Choudhary and D. M. Margolis. Curing HIV: Pharmacologic approaches to target HIV-1 latency. *Annu Rev Pharmacol Toxicol*, 51:397–418, Feb 2011. 15

[17] J. Coffin. HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. *Science*, 267(5197):483, 1995. 78

[18] D. C. Committee. Delta: a randomised double-blind controlled trial comparing combinations of zidovudine plus didanosine or zalcitabine with zidovudine alone in HIV-infected individuals. *Lancet*, 348(9023):283–291, Aug 1996. 18

[19] D. A. Cooper, J. Heera, J. Goodrich, M. Tawadrous, M. Saag, E. Dejesus, N. Clumeck, S. Walmsley, N. Ting, E. Coakley, J. D. Reeves, G. Reyes-Teran, M. Westby, E. V. D. Ryst, P. Ive, L. Mohapi, H. Mingrone, A. Horban, F. Hackman, J. Sullivan, and H. Mayer. Maraviroc versus efavirenz, both in combination with zidovudine-lamivudine, for the treatment of antiretroviral-naive subjects with CCR5-tropic HIV-1 infection. *J Infect Dis*, 201(6):803–813, Mar 2010. 30

[20] G. M. Cooper and R. E. Hausman. *The Cell - A Molecular Approach*. ASM Press and Sinauer Associates, Inc., 2000. 79

[21] G. E. Crooks, G. Hon, J.-M. Chandonia, and S. E. Brenner. WebLogo: a sequence logo generator. *Genome Res*, 14(6):1188–1190, Jun 2004. 26

[22] J. Currier, A. Lazzarin, L. Sloan, N. Clumeck, J. Slims, D. McCarty, H. Steel, J.-P. Kleim, T. Bonny, J. Millard, and A. S. C. E. N. T. study team. Antiviral activity and safety of aplaviroc with lamivudine/zidovudine in hiv-infected, therapy-naive patients: the ascent (ccr102881) study. *Antivir Ther*, 13(2):297–306, 2008. 28

[23] E. S. Daar, X. L. Li, T. Moudgil, and D. D. Ho. High concentrations of recombinant soluble CD4 are required to neutralize primary human immunodeficiency virus type 1 isolates. *Proc Natl Acad Sci U S A*, 87(17):6574–6578, Sep 1990. 19

[24] S. G. Deeks. Determinants of virological response to antiretroviral therapy: implications for long-term strategies. *Clin Infect Dis*, 30 Suppl 2:S177–S184, Jun 2000. 18

[25] S. G. Deeks, J. D. Barbour, R. M. Grant, and J. N. Martin. Duration and predictors of CD4 T-cell gains in patients who continue combination therapy despite detectable plasma viremia. *AIDS*, 16(2):201–207, Jan 2002. 19

[26] Deutsche AIDS Gesellschaft e.V. und Österreichische AIDS Gesellschaft. *Empfehlung zur Bestimmung des HIV-1-Korezeptor-Gebrauchs*. Deutsche AIDS Gesellschaft e.V. und Österreichische AIDS Gesellschaft, 2009. 59

[27] Deutsche AIDS Gesellschaft e.V. und Österreichische AIDS Gesellschaft. *Deutsch-Österreichische Leitlinien zur antiretroviralen Therapie der HIV-1-Infektion*. Deutsche AIDS Gesellschaft e.V. und Österreichische AIDS Gesellschaft, 2010. 17, 18, 21, 35

[28] T. J. Dolinsky, J. E. Nielsen, J. A. McCammon, and N. A. Baker. PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res*, 32(Web Server issue):W665–W667, Jul 2004. 48

[29] Domingo, Menéndez-Arias, and Holland. RNA virus fitness. *Rev Med Virol*, 7(2):87–96, Jul 1997. 78

[30] P. Dorr, M. Westby, S. Dobbs, P. Griffin, B. Irvine, M. Macartney, J. Mori, G. Rickett, C. Smith-Burchnell, C. Napier, R. Webster, D. Armour, D. Price, B. Stammen, A. Wood, and M. Perros. Maraviroc (UK-427,857), a potent, orally bioavailable, and selective small-molecule inhibitor of chemokine receptor CCR5 with broad-spectrum anti-human immunodeficiency virus type 1 activity. *Antimicrob Agents Chemother*, 49(11):4721–4732, Nov 2005. 28

[31] D. Dressman, H. Yan, G. Traverso, K. W. Kinzler, and B. Vogelstein. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc Natl Acad Sci U S A*, 100(15):8817–8822, Jul 2003. 81

[32] J. N. Dybowski, D. Heider, and D. Hoffmann. Prediction of co-receptor usage of HIV-1 from genotype. *PLoS Comput Biol*, 6(4):e1000743, 2010. 37

[33] J. N. Dybowski, D. Heider, and D. Hoffmann. Structure of HIV-1 quasi-species as early indicator for switches of co-receptor tropism. *AIDS Res Ther*, 7(1):41, Nov 2010. 100, 101

[34] M. Eigen. Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften*, 58(10):465–523, Oct 1971. 78

[35] M. Eigen and P. Schuster. The hypercycle. a principle of natural self-organization. Part A: Emergence of the hypercycle. *Naturwissenschaften*, 64(11):541–565, Nov 1977. 77

[36] M. Eigen and P. Schuster. The hypercycle. a principle of natural self-organization. Part B: The abstract hypercycle. *Naturwissenschaften*, 64:7–41, 1978. 77

[37] M. Eigen and P. Schuster. The hypercycle. a principle of natural self-organization. Part C: The realistic hypercycle. *Naturwissenschaften*, 65(7):341–369, Jul 1978. 77

[38] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27:861–874, 2006. 54

[39] R. A. Fouchier, M. Groenink, N. A. Kootstra, M. Tersmette, H. G. Huisman, F. Miedema, and H. Schuitemaker. Phenotype-associated sequence variation in the third variable domain of the human immunodeficiency virus type 1 gp120 molecule. *J Virol*, 66(5):3183–3187, May 1992. 33

[40] A. E. Friedman-Kien. Disseminated Kaposi's sarcoma syndrome in young homosexual men. *J Am Acad Dermatol*, 5(4):468–471, Oct 1981. 12

[41] R. C. Gallo. A reflection on HIV/AIDS research after 25 years. *Retrovirology*, 3:72, 2006. 12

[42] R. C. Gallo, S. Z. Salahuddin, M. Popovic, G. M. Shearer, M. Kaplan, B. F. Haynes, T. J. Palker, R. Redfield, J. Oleske, and B. Safai. Frequent detection and isolation of cytopathic retroviruses (HTLV-III) from patients with AIDS and at risk for AIDS. *Science*, 224(4648):500–503, May 1984. 12

[43] C. Garrido, V. Roulet, N. Chueca, E. Poveda, A. Aguilera, K. Skrabal, N. Zahonero, S. Carlos, F. García, J. L. Faudon, V. Soriano, and C. de Mendoza. Evaluation of eight different bioinformatics tools to predict viral tropism in different human immunodeficiency virus type 1 subtypes. *J Clin Microbiol*, 46(3):887–891, Mar 2008. 47

[44] J. Gathe, R. Diaz, G. Faetkenheuer, J. Zeinecker, C. Mak, R. Vilchez, W. Greaves, S. Kumar, C. Onyebuchi, and L. Dunkle. Phase 3 trials of Vicriviroc in treatment-experienced subjects demonstrate safety but not significantly superior efficacy over potent background regimens alone. In *17th Conference on Retroviruses and Opportunistic Infections (CROI 2010), San Francisco, USA*, 2010. 30, 32

[45] M. S. Gottlieb, R. Schroff, H. M. Schanker, J. D. Weisman, P. T. Fan, R. A. Wolf, and A. Saxon. Pneumocystis carinii pneumonia and mucosal candidiasis in previously healthy homosexual men: evidence of a new acquired cellular immunodeficiency. *N Engl J Med*, 305(24):1425–1431, Dec 1981. 12

[46] R. M. Gulick, J. Lalezari, J. Goodrich, N. Clumeck, E. DeJesus, A. Horban, J. Nadler, B. Clotet, A. Karlsson, M. Wohlfeiler, J. B. Montana, M. McHale, J. Sullivan, C. Ridgway, S. Felstead, M. W. Dunne, E. van der Ryst, H. Mayer, and M. O. T. I. V. A. T. E. S. Teams. Maraviroc for previously treated patients with R5 HIV-1 infection. *N Engl J Med*, 359(14):1429–1441, Oct 2008. 28, 32

[47] R. M. Gulick, Z. Su, C. Flexner, M. D. Hughes, P. R. Skolnik, T. J. Wilkin, R. Gross, A. Krambrink, E. Coakley, W. L. Greaves, A. Zolopa, R. Reichman, C. Godfrey, M. Hirsch, D. R. Kuritzkes, and A. I. D. S. C. T. G. . Team. Phase 2 study of the safety and efficacy of vicriviroc, a CCR5 inhibitor, in HIV-1-infected, treatment-experienced patients: AIDS clinical trials group 5211. *J Infect Dis*, 196(2):304–312, Jul 2007. 30, 32

[48] S. M. Hammer, D. A. Katzenstein, M. D. Hughes, H. Gundacker, R. T. Schooley, R. H. Haubrich, W. K. Henry, M. M. Lederman, J. P. Phair, M. Niu, M. S. Hirsch, and T. C. Merigan. A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. AIDS clinical trials group study 175 study team. *N Engl J Med*, 335(15):1081–1090, Oct 1996. 18

[49] F. Hamy, V. Vidal, and e. a. Hubert, S. Development of a new hybridization-based 'genosorting' method for highly sensitive determination of hiv co-receptor tropism. In *3rd International Workshop on Targeting HIV Entry, Washington DC*, 2007. 32

[50] D. Heider, J. Appelmann, T. Bayro, W. Dreckmann, A. Held, J. Winkler, A. Barnekow, and M. Borschbach. A computational approach for the identification of small GTPases based on preprocessed amino acid sequences. *Technology in Cancer Research and Treatment*, 8(5):333–342, 2009. 46, 49

[51] D. Heider, S. Hauke, M. Pyka, and D. Kessler. Insights into the classification of small GTPases. *Advances and Applications in Bioinformatics and Chemistry*, in press, 2010. 49

[52] D. Heider, J. Verheyen, and D. Hoffmann. Predicting bevirimat resistance of HIV-1 from genotype. *BMC Bioinformatics*, 11:37, 2010. 46

[53] T. K. Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:832–844, 1998. 37

[54] C. Hoffmann and J. Rockstroh. *HIV 2010*. Medizin Fokus Verlag, 2010. 12, 13, 15, 18, 20, 33

[55] U. Höweler. *MAXIMOBY 8.1 and MOBY 3.0*. CHEOPS, Altenberge, Germany, 2007. 48

[56] C.-C. Huang, S. N. Lam, P. Acharya, M. Tang, S.-H. Xiang, S. S.-U. Hussan, R. L. Stanfield, J. Robinson, J. Sodroski, I. A. Wilson, R. Wyatt, C. A. Bewley, and P. D. Kwong. Structures of the CCR5 N terminus and of a tyrosine-sulfated antibody with HIV-1 gp120 and CD4. *Science*, 317(5846):1930–1934, Sep 2007. 22, 27, 74, 75

[57] C.-C. Huang, M. Tang, M.-Y. Zhang, S. Majeed, E. Montabana, R. L. Stanfield, D. S. Dimitrov, B. Korber, J. Sodroski, I. A. Wilson, R. Wyatt, and P. D. Kwong. Structure of a V3-containing hiv-1 gp120 core. *Science*, 310(5750):1025–1028, Nov 2005. 26, 48, 51, 76

[58] W. Huang, S. H. Eshleman, J. Toma, E. Stawiski, J. M. Whitcomb, J. B. Jackson, L. Guay, P. Musoke, N. Parkin, and C. J. Petropoulos. Vertical transmission of X4-tropic and dual-tropic HIV-1 in five Ugandan mother-infant pairs. *AIDS*, 23(14):1903–1908, Sep 2009. 47

[59] Y. Huang, W. A. Paxton, S. M. Wolinsky, A. U. Neumann, L. Zhang, T. He, S. Kang, D. Ceradini, Z. Jin, K. Yazdanbakhsh, K. Kunstman, D. Erickson, E. Dragon, N. R. Landau, J. Phair, D. D. Ho, and R. A. Koup. The role of a mutant CCR5 allele in HIV-1 transmission and disease progression. *Nat Med*, 2(11):1240–1243, Nov 1996. 20

[60] A. Hughes and M. Nelson. HIV entry: new insights and implications for patient management. *Curr Opin Infect Dis*, 22(1):35–42, Feb 2009. 23

[61] S. S. Hwang, T. J. Boyle, H. K. Lyerly, and B. R. Cullen. Identification of the envelope V3 loop as the primary determinant of cell tropism in HIV-1. *Science*, 253(5015):71–74, Jul 1991. 33

[62] V. Idemyor. Human immunodeficiency virus (HIV) entry inhibitors (CCR5 specific blockers) in development: are they the next novel therapies? *HIV Clin Trials*, 6(5):272–277, 2005. 30

[63] H. Imamichi, K. Crandall, V. Natarajan, M. Jiang, R. Dewar, S. Berg, A. Gaddam, M. Bosche, J. Metcalf, R. Davey Jr, et al. Human immunodeficiency virus type 1 quasi species that rebound after discontinuation of highly active antiretroviral therapy are similar to the viral quasi species present before initiation of therapy. *The Journal of infectious diseases*, 183(1):36–50, 2001. 78

[64] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, Oct 2004. 79

[65] M. A. Jensen, F.-S. Li, A. B. van 't Wout, D. C. Nickle, D. Shriner, H.-X. He, S. McLaughlin, R. Shankarappa, J. B. Margolick, and J. I. Mullins. Improved coreceptor usage prediction and genotypic monitoring of R5-to-X4 transition by motif analysis of human immunodeficiency virus type 1 env V3 loop sequences. *J Virol*, 77(24):13376–13388, Dec 2003. 33, 35, 46, 83

[66] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, and M. Kanehisa. AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res*, 36(Database issue):D202–D205, Jan 2008. 46

[67] B. F. Keele, E. E. Giorgi, J. F. Salazar-Gonzalez, J. M. Decker, K. T. Pham, M. G. Salazar, C. Sun, T. Grayson, S. Wang, H. Li, X. Wei, C. Jiang, J. L. Kirchherr, F. Gao, J. A. Anderson, L.-H. Ping, R. Swanstrom, G. D. Tomaras, W. A. Blattner, P. A. Goepfert, J. M. Kilby, M. S. Saag, E. L. Delwart, M. P. Busch, M. S. Cohen, D. C. Montefiori, B. F. Haynes, B. Gaschen, G. S. Athreya, H. Y. Lee, N. Wood, C. Seoighe, A. S. Perelson, T. Bhattacharya, B. T. Korber, B. H. Hahn, and G. M. Shaw. Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proc Natl Acad Sci U S A*, 105(21):7552–7557, May 2008. 47

[68] A. Kernytsky and B. Rost. Using genetic algorithms to select most predictive protein features. *Proteins*, 75:75–88, 2009. 37

[69] O. M. Klibanov, S. H. Williams, and C. A. Iler. Cenicriviroc, an orally active CCR5 antagonist for the potential treatment of HIV infection. *Curr Opin Investig Drugs*, 11(8):940–950, Aug 2010. 28

[70] R. Kondru, J. Zhang, C. Ji, T. Mirzadegan, D. Rotstein, S. Sankuratri, and M. Dioszegi. Molecular interactions of CCR5 with major classes of small-molecule anti-HIV CCR5 antagonists. *Mol Pharmacol*, 73(3):789–800, Mar 2008. 28, 29, 30, 31

[71] L. I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. John Wiley and Sons Inc., 2004. 37

[72] L. I. Kuncheva and C. J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51:181–207, 2003. 37

[73] D. R. Kuritzkes. HIV-1 entry inhibitors: an overview. *Curr Opin HIV AIDS*, 4(2):82–87, Mar 2010. 30

[74] D. R. Kuritzkes, J. Jacobson, W. G. Powderly, E. Godofsky, E. DeJesus, F. Haas, K. A. Reimann, J. L. Larson, P. O. Yarbough, V. Curt, and W. R. Shanahan. Antiretroviral activity of the anti-CD4 monoclonal antibody TNX-355 in patients infected with HIV type 1. *J Infect Dis*, 189(2):286–291, Jan 2004. 20

[75] P. D. Kwong, R. Wyatt, J. Robinson, R. W. Sweet, J. Sodroski, and W. A. Hendrickson. Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. *Nature*, 393(6686):648–659, Jun 1998. 26

[76] J. Kyte and R. Doolittle. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, 157:105–132, 1982. 46, 74

[77] F. C. Lampe, J. M. Gatell, S. Staszewski, M. A. Johnson, C. Pradier, M. J. Gill, E. de Lazzari, B. Dauer, M. Youle, E. Fontas, H. B. Krentz, and A. N. Phillips. Changes over time in risk of initial virological failure of combination antiretroviral therapy: a multicohort analysis, 1996 to 2002. *Arch Intern Med*, 166(5):521–528, Mar 2006. 21

[78] B. Ledergerber, J. D. Lundgren, A. S. Walker, C. Sabin, A. Justice, P. Reiss, C. Mussini, F. Wit, A. d'Arminio Monforte, R. Weber, G. Fusco, S. Staszewski, M. Law, R. Hogg, F. Lampe, M. J. Gill, F. Castelli, A. N. Phillips, and P. L. A. T. O. Collaboration. Predictors of trend in CD4-positive T-cell count and mortality among HIV-1-infected individuals with virological failure to all three antiretroviral-drug classes. *Lancet*, 364(9428):51–62, 2004. 19

[79] M. M. Lederman, A. Penn-Nicholson, M. Cho, and D. Mosier. Biology of CCR5 and its role in HIV infection and treatment. *JAMA*, 296(7):815–826, Aug 2006. 20

[80] J. Li, P. C. Edwards, M. Burghammer, C. Villa, and G. F. X. Schertler. Structure of bovine rhodopsin in a trigonal crystal form. *J Mol Biol*, 343(5):1409–1438, Nov 2004. 48

[81] Y. Li, M. A. Rey-Cuille, and S. L. Hu. N-linked glycosylation in the V3 region of HIV type 1 surface antigen modulates coreceptor usage in viral infection. *AIDS Res Hum Retroviruses*, 17(16):1473–1479, Nov 2001. 73

[82] A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002. 41, 49, 52

[83] J. Liu, A. Bartesaghi, M. J. Borgnia, G. Sapiro, and S. Subramaniam. Molecular architecture of native HIV-1 gp120 trimers. *Nature*, 455(7209):109–113, Sep 2008. 22, 26

[84] J. Liu, Y. Deng, Q. Li, A. K. Dey, J. P. Moore, and M. Lu. Role of a putative gp41 dimerization domain in human immunodeficiency virus type 1 membrane fusion. *J Virol*, 84(1):201–209, Jan 2010. 26

[85] N. Lohse, G. Kronborg, J. Gerstoft, C. S. Larsen, G. Pedersen, C. Pedersen, H. T. Sœrensen, and N. Obel. Virological control during the first 6-18 months after initiating highly active antiretroviral therapy as a predictor for outcome in HIV-infected patients: a danish, population-based, 6-year follow-up study. *Clin Infect Dis*, 42(1):136–144, Jan 2006. 21

[86] L. Lopalco. CCR5: From natural resistance to a new anti-HIV strategy. *Viruses*, 2:574–600, 2010. 24

[87] A. Low, W. Dong, D. Chan, T. Sing, R. Swanstrom, M. Jensen, S. Pillai, B. Good, and P. Harrigan. Current V3 genotyping algorithms are inadequate for predicting X4 co-receptor usage in clinical isolates. *AIDS*, 21(14):F17, 2007. 35, 81

[88] A. J. Low, L. C. Swenson, and P. R. Harrigan. HIV coreceptor phenotyping in the clinical setting. *AIDS Rev*, 10(3):143–151, 2008. 35, 81

[89] M. Luckey. *Membrane Structural Biology: with biochemical and biophysical foundations*. Cambridge University Press, 2008. 23

[90] R. D. MacArthur and R. M. Novak. Reviews of anti-infective agents: maraviroc: the first of a new class of antiretroviral agents. *Clin Infect Dis*, 47(2):236–241, Jul 2008. 28, 33

[91] M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M.

Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. I. Alenquer, T. P. Jarvie, K. B. Jirage, J.-B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, Sep 2005. 80

[92] H. Masur, M. A. Michelis, J. B. Greene, I. Onorato, R. A. Stouwe, R. S. Holzman, G. Wormser, L. Brettman, M. Lange, H. W. Murray, and S. Cunningham-Rundles. An outbreak of community-acquired pneumocystis carinii pneumonia: initial manifestation of cellular immune dysfunction. *N Engl J Med*, 305(24):1431–1438, Dec 1981. 12

[93] H. Mayer, E. Van der Ryst, M. Saag, B. Clotet, G. Fätkenheuer, N. Clumeck, K. Turner, and J. Goodrich. Safety and efficacy of maraviroc (MVC), a novel CCR5 antagonist, when used in combination with optimized background therapy (OBT) for the treatment of antiretroviral-experienced subjects infected with dual/mixed-tropic HIV-1: 24-week results of a Phase 2b exploratory trial. In *16th International AIDS Conference*, pages 13–18, 2006. 32

[94] M. L. Metzker. Sequencing technologies - the next generation. *Nat Rev Genet*, 11(1):31–46, Jan 2010. 80, 82

[95] I. Mezzaroma, M. Carlesimo, E. Pinter, D. S. Muratori, F. D. Sora, F. Chiarotti, M. G. Cunsolo, G. Sacco, and F. Aiuti. Clinical and immunologic response without decrease in virus load in patients with AIDS after 24 months of highly active antiretroviral therapy. *Clin Infect Dis*, 29(6):1423–1430, Dec 1999. 18

[96] G. J. Moyle, A. Wildfire, S. Mandalia, H. Mayer, J. Goodrich, J. Whitcomb, and B. G. Gazzard. Epidemiology and predictive factors for chemokine receptor use in HIV-1 infection. *J Infect Dis*, 191(6):866–872, Mar 2005. 29, 32

[97] J. I. Mullins, L. Heath, J. P. Hughes, J. Kicha, S. Styrchak, K. G. Wong, U. Rao, A. Hansen, K. S. Harris, J.-P. Laurent, D. Li, J. H. Simpson, J. M. Essigmann, L. A. Loeb, and J. Parkins. Mutation of HIV-1 genomes in a clinical population treated with the mutagenic nucleoside KP1461. *PLoS One*, 6(1):e15135, 2011. 78

[98] S. Naganawa, M. Yokoyama, T. Shiino, T. Suzuki, Y. Ishigatsubo, A. Ueda, A. Shirai, M. Takeno, S. Hayakawa, S. Sato, O. Tochikubo, S. Kiyoura, K. Sawada, T. Ikegami, T. Kanda, K. Kitamura, and H. Sato. Net positive charge of HIV-1 CRF01-AE V3 sequence regulates viral sensitivity to humoral immunity. *PLoS One*, 3(9):e3206, 2008. 47

[99] K. Nakai, A. Kidera, and M. Kanehisa. Cluster analysis of amino acid indices for prediction of protein structure and function. *Protein Engineering*, 2:93–100, 1988. 37

[100] L. Nanni and A. Lumini. Using ensemble of classifiers for predicting HIV protease cleavage sites in proteins. *Amino Acids*, 36(3):409–416, Mar 2009. 37

[101] K. B. Napier, Z. xuan Wang, S. C. Peiper, and J. O. Trent. CCR5 interactions with the variable 3 loop of gp120. *J Mol Model*, 13(1):29–41, Jan 2007. 28

[102] National Institute for Allergies and Infectious Diseases. HIV/AIDS. http://www.niaid.nuh.gov/topics/hivaids, Jan. 2011. [Online; accessed 26-January-2011]. 16, 17

[103] U. Nations. *United Nations Millennium Development Goals Report 2010*. United Nations, 2010. 13, 14

[104] W. G. Nichols, H. M. Steel, T. Bonny, K. Adkison, L. Curtis, J. Millard, K. Kabeya, and N. Clumeck. Hepatotoxicity observed in clinical trials of aplaviroc (gw873140). *Antimicrob Agents Chemother*, 52(3):858–865, Mar 2008. 28

[105] R. A. Ogert, M. K. Lee, W. Ross, A. Buckler-White, M. A. Martin, and M. W. Cho. N-linked glycosylation sites adjacent to and within the V1/V2and the V3 loops of dualtropic human immunodeficiency virus type 1 isolate DH12 gp120 affect coreceptor usage and cellular tropism. *J Virol*, 75(13):5998–6006, Jul 2001. 73

[106] S. Ong, H. Lin, Y. Chen, Z. Li, and Z. Cao. Efficacy of different protein descriptors in predicting protein functional families. *BMC Bioinformatics*, 8:300, 2007. 37

[107] J. P. Overington, B. Al-Lazikani, and A. L. Hopkins. How many drug targets are there? *Nat Rev Drug Discov*, 5(12):993–996, Dec 2006. 23

[108] K. Palczewski, T. Kumasaka, T. Hori, C. A. Behnke, H. Motoshima, B. A. Fox, I. L. Trong, D. C. Teller, T. Okada, R. E. Stenkamp, M. Yamamoto, and

M. Miyano. Crystal structure of rhodopsin: A G protein-coupled receptor. *Science*, 289(5480):739–745, Aug 2000. 28

[109] A. S. Perelson, A. U. Neumann, M. Markowitz, J. M. Leonard, and D. D. Ho. HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science*, 271(5255):1582–1586, Mar 1996. 21

[110] A. N. Phillips, B. Gazzard, R. Gilson, P. Easterbrook, M. Johnson, J. Walsh, C. Leen, M. Fisher, C. Orkin, J. Anderson, D. Pillay, V. Delpech, C. Sabin, A. Schwenk, D. Dunn, M. Gompels, T. Hill, K. Porter, A. Babiker, and U. K. C. H. C. Study. Rate of AIDS diseases or death in HIV-infected antiretroviral therapy-naive individuals with high CD4 cell count. *AIDS*, 21(13):1717–1721, Aug 2007. 19

[111] S. Pillai, B. Good, D. Richman, and J. Corbeil. A new perspective on V3 phenotype prediction. *AIDS Res Hum Retroviruses*, 19(2):145–149, Feb 2003. 33, 35, 46

[112] B. J. Poiesz, F. W. Ruscetti, A. F. Gazdar, P. A. Bunn, J. D. Minna, and R. C. Gallo. Detection and isolation of type C retrovirus particles from fresh and cultured lymphocytes of a patient with cutaneous T-cell lymphoma. *Proc Natl Acad Sci U S A*, 77(12):7415–7419, Dec 1980. 12

[113] S. Portsmouth, M. Lewis, C. Craig, D. Chapman, L. Swenson, and J. Heera. Long-Term Outcome of Individuals Experiencing an HIV-1 Co-Receptor Tropism Switch in the MERIT Study. In *0th Interscience Conference on Antimicrobial Agents and Chemotherapy (ICAAC 2010), Boston, USA*, 2010. 59

[114] E. Poveda, E. Seclén, M. del Mar González, F. García, N. Chueca, A. Aguilera, J. J. Rodríguez, J. González-Lahoz, and V. Soriano. Design and validation of new genotypic tools for easy and reliable estimation of HIV tropism before using CCR5 antagonists. *J Antimicrob Chemother*, 63(5):1006–1010, May 2009. 33, 83

[115] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006. ISBN 3-900051-07-0, http://www.R-project.org. 49

[116] S. G. F. Rasmussen, H.-J. Choi, D. M. Rosenbaum, T. S. Kobilka, F. S. Thian, P. C. Edwards, M. Burghammer, V. R. P. Ratnala, R. Sanishvili, R. F. Fischetti, G. F. X. Schertler, W. I. Weis, and B. K. Kobilka. Crystal structure of the human beta2 adrenergic G-protein-coupled receptor. *Nature*, 450(7168):383–387, Nov 2007. 28

[117] W. Resch, N. Hoffman, and R. Swanstrom. Improved success of phenotype prediction of the human immunodeficiency virus type 1 from envelope variable loop 3 sequence using neural networks. *Virology*, 288(1):51–62, Sep 2001. 33, 35, 46

[118] J. D. Rose, A. M. Rhea, J. Weber, and M. E. Quiñones-Mateu. Current tests to evaluate HIV-1 coreceptor tropism. *Curr Opin HIV AIDS*, 4(2):136–142, Mar 2009. 33

[119] O. Rosen, J. Chill, M. Sharon, N. Kessler, B. Mester, S. Zolla-Pazner, and J. Anglister. Induced fit in HIV-neutralizing antibody complexes: evidence for alternative conformations of the gp120 V3 loop and the molecular basis for broad neutralization. *Biochemistry*, 44(19):7250–7258, May 2005. 26

[120] V. Roulet, S. Rochas, J. Labernardiere, F. Mammano, J. Faudon, N. Raja, S. Lebel-Binay, and K. Skrabal. HIV-1 PHENOSCRIPT ENV: A sensitive assay for the detection of HIV X4 minority species and determination of non-B subtype viral tropism. In *14th Conference on Retroviruses and Opportunistic Infections (CROI 2007), Los Angeles, USA*, 2007. 32

[121] G. Rozera, I. Abbate, A. Bruselles, C. Vlassi, G. D'Offizi, P. Narciso, G. Chillemi, M. Prosperi, G. Ippolito, and M. R. Capobianchi. Massively parallel pyrosequencing highlights minority variants in the HIV-1 env quasispecies deriving from lymphomonocyte sub-populations. *Retrovirology*, 6:15, 2009. 25, 81

[122] C. M. Ruiz-Jarabo, A. Arias, E. Baranowski, C. Escarmís, and E. Domingo. Memory in viral quasispecies. *J Virol*, 74(8):3543–3547, Apr 2000. 78

[123] D. Ruta and B. Gabrys. Classifier selection for majority voting. *J. Inf. Fusion*, 6:63–81, 2005. 37

[124] M. Saag, J. Goodrich, G. Fätkenheuer, B. Clotet, N. Clumeck, J. Sullivan, M. Westby, E. van der Ryst, H. Mayer, and A4001029 Study Group. A double-blind, placebo-controlled trial of maraviroc in treatment-experienced patients infected with non-R5 HIV-1. *J Infect Dis*, 199(11):1638–1647, Jun 2009. 59

[125] A. Sali and T. L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*, 234(3):779–815, Dec 1993. 48

[126] O. Sander, T. Sing, I. Sommer, A. J. Low, P. K. Cheung, P. R. Harrigan, T. Lengauer, and F. S. Domingues. Structural descriptors of gp120 V3 loop for

the prediction of HIV-1 coreceptor usage. *PLoS Comput Biol*, 3(3):e58, Mar 2007. 33, 35

[127] F. Sanger and A. R. Coulson. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol*, 94(3):441–448, May 1975. 78

[128] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*, 74(12):5463–5467, Dec 1977. 78

[129] E. Seclén, C. Garrido, M. D. M. González, J. González-Lahoz, C. de Mendoza, V. Soriano, and E. Poveda. High sensitivity of specific genotypic tools for detection of X4 variants in antiretroviral-experienced patients suitable to be treated with CCR5 antagonists. *J Antimicrob Chemother*, Apr 2010. 81

[130] A. K. Sethi, D. D. Celentano, S. J. Gange, R. D. Moore, and J. E. Gallant. Association between adherence to antiretroviral therapy and human immunodeficiency virus drug resistance. *Clin Infect Dis*, 37(8):1112–1118, Oct 2003. 18

[131] M. Sharon, N. Kessler, R. Levy, S. Zolla-Pazner, M. Görlach, and J. Anglister. Alternative conformations of HIV-1 V3 loops mimic beta hairpins in chemokines, suggesting a mechanism for coreceptor selectivity. *Structure*, 11(2):225–236, Feb 2003. 26

[132] P. M. Sharp, E. Bailes, R. R. Chaudhuri, C. M. Rodenburg, M. O. Santiago, and B. H. Hahn. The origins of acquired immune deficiency syndrome viruses: where and when? *Philos Trans R Soc Lond B Biol Sci*, 356(1410):867–876, Jun 2001. 12

[133] T. Shioda, J. A. Levy, and C. Cheng-Mayer. Small amino acid changes in the V3 hypervariable region of gp120 can affect the T-cell-line and macrophage tropism of human immunodeficiency virus type 1. *Proc Natl Acad Sci U S A*, 89(20):9434–9438, Oct 1992. 33

[134] W. Siedlecki and J. Sklansky. A note on genetic algorithms for large scale feature selection. *Pattern Recognition Letters*, 10:335–347, 1989. 37

[135] F. P. Siegal, C. Lopez, G. S. Hammer, A. E. Brown, S. J. Kornfeld, J. Gold, J. Hassett, S. Z. Hirschman, C. Cunningham-Rundles, and B. R. Adelsberg. Severe acquired immunodeficiency in male homosexuals, manifested by chronic perianal ulcerative herpes simplex lesions. *N Engl J Med*, 305(24):1439–1444, Dec 1981. 12

[136] T. Sing, A. J. Low, N. Beerenwinkel, O. Sander, P. K. Cheung, F. S. Domingues, J. Büch, M. Däumer, R. Kaiser, T. Lengauer, and P. R. Harrigan. Predicting HIV coreceptor usage on the basis of genetic and clinical covariates. *Antivir Ther*, 12(7):1097–1106, 2007. 33, 35, 83

[137] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer. ROCR: visualizing classifier performance in R. *Bioinformatics*, 21(20):3940–3941, Oct 2005. 49, 50

[138] S. Singh, B. K. Malik, and D. K. Sharma. Targeting HIV-1 through molecular modeling and docking studies of CXCR4: leads for therapeutic development. *Chem Biol Drug Des*, 69(3):191–203, Mar 2007. 28

[139] K. Skrabal, A. J. Low, W. Dong, T. Sing, P. K. Cheung, F. Mammano, and P. R. Harrigan. Determining human immunodeficiency virus coreceptor use in a clinical setting: degree of correlation between two phenotypic assays and a bioinformatic model. *J Clin Microbiol*, 45(2):279–284, Feb 2007. 35, 81

[140] L. M. Smith, R. J. Kaiser, J. Z. Sanders, and L. E. Hood. The synthesis and use of fluorescent oligonucleotides in DNA sequence analysis. *Methods Enzymol*, 155:260–301, 1987. 79

[141] L. M. Smith, J. Z. Sanders, R. J. Kaiser, P. Hughes, C. Dodd, C. R. Connell, C. Heiner, S. B. Kent, and L. E. Hood. Fluorescence detection in automated DNA sequence analysis. *Nature*, 321(6071):674–679, 1986. 79

[142] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *J Mol Biol*, 147(1):195–197, Mar 1981. 84, 86

[143] K. Steegen, M. Bronze, E. V. Craenenbroeck, B. Winters, K. V. der Borght, C. L. Wallis, W. Stevens, T. F. R. de Wit, L. J. Stuyver, and the ART-A consortium. A comparative analysis of HIV drug resistance interpretation based on short reverse transcriptase sequences versus full sequences. *AIDS Res Ther*, 7(1):38, 2010. 101

[144] J. A. C. Sterne, M. May, D. Costagliola, F. de Wolf, A. N. Phillips, R. Harris, M. J. Funk, R. B. Geskus, J. Gill, F. Dabis, J. M. Miró, A. C. Justice, B. Ledergerber, G. Fätkenheuer, R. S. Hogg, A. D. Monforte, M. Saag, C. Smith, S. Staszewski, M. Egger, and S. R. Cole. Timing of initiation of antiretroviral therapy in AIDS-free HIV-1-infected patients: a collaborative analysis of 18 HIV cohort studies. *Lancet*, 373(9672):1352–1363, Apr 2009. 19

[145] L. C. Swenson, A. Moores, A. J. Low, A. Thielen, W. Dong, C. Woods, M. A. Jensen, B. Wynhoven, D. Chan, C. Glascock, and P. R. Harrigan. Improved detection of CXCR4-using HIV by V3 genotyping: application of population-based and "deep" sequencing to plasma RNA and proviral DNA. *J Acquir Immune Defic Syndr*, 54(5):506–510, Aug 2010. 82, 83, 100

[146] A. Tarca, V. Carey, X. Chen, R. Romero, and S. Draghici. Machine learning and its applications to biology. *PLoS Comput Biol*, 3(6):e116, 2007. 36

[147] A. Thielen, N. Sichtig, R. Kaiser, J. Lam, P. R. Harrigan, and T. Lengauer. Improved prediction of HIV-1 coreceptor usage with sequence information from the second hypervariable loop of gp120. *J Infect Dis*, 202(9):1435–1443, Nov 2010. 100

[148] R. Tressler, H. Valdez, E. van der Ryst, I. James, M. Lewis, J. Wheeler, and S. Than. Comparison of Results from the SensiTrop™ vs Trofile™ Assays on 100 Samples from the Maraviroc Expanded Access Program. In *15th Conference on Retroviruses and Opportunistic Infections, Boston, USA*, 2008. 33

[149] L. Trinh, D. Han, W. Huang, T. Wrin, J. Larson, L. Kiss, E. Coakley, C. Petropoulos, N. Parkin, J. Whitcomb, et al. Validation of an enhanced sensitivity Trofile™ HIV-1 co-receptor tropism assay for selecting patients for therapy with entry inhibitors targeting CCR5. *Journal of the International AIDS Society*, 11(Suppl 1):P197, 2008. 32, 83

[150] A. D. Trister and D. A. Hammer. Role of gp120 trimerization on hiv binding elucidated with brownian adhesive dynamics. *Biophys J*, 95(1):40–53, Jul 2008. 15

[151] A. M. N. Tsibris, B. Korber, R. Arnaout, C. Russ, C.-C. Lo, T. Leitner, B. Gaschen, J. Theiler, R. Paredes, Z. Su, M. D. Hughes, R. M. Gulick, W. Greaves, E. Coakley, C. Flexner, C. Nusbaum, and D. R. Kuritzkes. Quantitative deep sequencing reveals dynamic HIV-1 escape and large population shifts during CCR5 antagonist therapy in vivo. *PLoS One*, 4(5):e5683, 2009. 25, 81, 83, 84, 89, 99, 100

[152] V. Tugarinov, A. Zvi, R. Levy, and J. Anglister. A cis proline turn linking two beta-hairpin strands in the solution structure of an antibody-bound HIV-1IIIB V3 peptide. *Nat Struct Biol*, 6(4):331–335, Apr 1999. 26

[153] UNAIDS. AIDS epidemic update 2009. 2009. 13

[154] UNAIDS. *UNAIDS Report on the Global AIDS Epidemic.* UNAIDS, 2010. 13, 14

[155] J. Våbenø, G. V. Nikiforovich, and G. R. Marshall. Insight into the binding mode for cyclopentapeptide antagonists of the CXCR4 receptor. *Chem Biol Drug Des*, 67(5):346–354, May 2006. 28

[156] I. Vandenbroucke, H. V. Marck, W. Mostmans, V. V. Eygen, E. Rondelez, K. Thys, K. V. Baelen, K. Fransen, D. Vaira, K. Kabeya, S. D. Wit, E. Florence, M. Moutschen, L. Vandekerckhove, C. Verhofstede, and L. J. Stuyver. HIV-1 V3 envelope deep sequencing for clinical plasma specimens failing in phenotypic tropism assays. *AIDS Research and Therapy*, 7:4, 2010. 83, 100

[157] C. T. Veldkamp, C. Seibert, F. C. Peterson, N. B. D. la Cruz, J. C. Haugner, H. Basnet, T. P. Sakmar, and B. F. Volkman. Structural basis of CXCR4 sulfotyrosine recognition by the chemokine SDF-1/CXCL12. *Sci Signal*, 1(37):ra4, 2008. 27

[158] R. Vilchez, J. Strizki, J. Quiroz, L. Dunkle, and W. Greaves. Comparison of Trofile and ViroTect Tropism Assays in Treatment-Experienced Subjects. In *16th Conference on Retroviruses and Opportunistic Infections, Montréal, Canada*, 2009. 33

[159] W. F. Vranken, M. Budesinsky, F. Fant, K. Boulez, and F. A. Borremans. The complete consensus V3 loop peptide of the envelope protein gp120 of HIV-1 shows pronounced helical character in solution. *FEBS Lett*, 374(1):117–121, Oct 1995. 26

[160] J. M. Whitcomb, W. Huang, S. Fransen, K. Limoli, J. Toma, T. Wrin, C. Chappey, L. D. B. Kiss, E. E. Paxinos, and C. J. Petropoulos. Development and characterization of a novel single-cycle recombinant-virus assay to determine human immunodeficiency virus type 1 coreceptor tropism. *Antimicrob Agents Chemother*, 51(2):566–575, Feb 2007. 32, 83, 87

[161] Wikipedia. AIDS pandemic — wikipedia, the free encyclopedia. `http://en.wikipedia.org/w/index.php?title=AIDS_pandemic&oldid=411159124`, 2011. [Online; accessed 14-February-2011]. 14

[162] T. J. Wilkin, Z. Su, A. Krambrink, J. Long, W. Greaves, R. Gross, M. D. Hughes, C. Flexner, P. R. Skolnik, E. Coakley, C. Godfrey, M. Hirsch, D. R. Kuritzkes, and R. M. Gulick. Three-year safety and efficacy of vicriviroc, a CCR5 antagonist, in HIV-1-infected treatment-experienced patients. *J Acquir Immune Defic Syndr*, 54(5):470–476, Aug 2010. 30

[163] C. Woese and G. Fox. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America*, 74(11):5088, 1977. 36

[164] S. Wolf, M. Böckmann, U. Höweler, J. Schlitter, and K. Gerwert. Simulations of a G protein-coupled receptor homology model predict dynamic features and a ligand binding site. *FEBS letters*, 582(23-24):3335–3342, 2008. 48

[165] D. Wolpert. Stacked generalization. *Neural Networks*, 5:241–260, 1992. 36, 63

[166] M. Worobey, M. Gemmel, D. E. Teuwen, T. Haselkorn, K. Kunstman, M. Bunce, J.-J. Muyembe, J.-M. M. Kabongo, R. M. Kalengayi, E. V. Marck, M. T. P. Gilbert, and S. M. Wolinsky. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature*, 455(7213):661–664, Oct 2008. 13

[167] B. Wu, E. Y. T. Chien, C. D. Mol, G. Fenalti, W. Liu, V. Katritch, R. Abagyan, A. Brooun, P. Wells, F. C. Bi, D. J. Hamel, P. Kuhn, T. M. Handel, V. Cherezov, and R. C. Stevens. Structures of the CXCR4 Chemokine GPCR with Small-Molecule and Cyclic Peptide Antagonists. *Science*, 330(6007):1066–1071, 2010. 27, 75

[168] H. Wu, D. G. Myszka, S. W. Tendian, C. G. Brouillette, R. W. Sweet, I. M. Chaiken, and W. A. Hendrickson. Kinetic and structural analysis of mutant CD4 receptors that are defective in HIV gp120 binding. *Proc Natl Acad Sci U S A*, 93(26):15030–15035, Dec 1996. 25

[169] G. Zenobi and Cunningham. Using diversity in preparing ensembles of classifiers based on different feature subsets to minimize generalization error. In 2167, editor, *Proceedings of the 12th Conference on Machine Learning*, Lecture Notes in Computer Science, 2001. 37

[170] H. Zhang and W. A. Cramer. Problems in obtaining diffraction-quality crystals of hetero-oligomeric integral membrane proteins. *J Struct Funct Genomics*, 6(2-3):219–223, 2005. 26

List of Publications

Journal Articles

Dybowski JN, Heider D, Hoffmann D. Structure of HIV-1 quasi-species as early indicator for switches of co-receptor tropism. *AIDS Research and Therapy*, 7:41, 2010.

Dybowski JN, Heider D, Hoffmann D. Prediction of co-receptor usage of HIV-1 from genotype. *PLoS Comput Biol*, 6(4):e1000743, 2010.

Horsthemke B, Wawrzik M, Groß S, Lich C, Sauer B, Rost I, Krasemann E, Kosyakova N, Liehr T, Weise A, **Dybowski JN**, Hoffmann D, Wieczorek D. Parental origin and functional relevance of a de novo UBE3A variant. *Eur J Med Genet*, 1769–7212, 2010.

Ehrentraut S, Weber JM, **Dybowski JN**, Hoffmann D, Ehrenhofer-Murray AE. Rpd3-dependent boundary formation at telomeres by removal of Sir2 substrate. *Proc Natl Acad Sci U S A* 2010, 107(12):5522–5527.

Posters

Dybowski JN, Heider D, Hoffmann D. Prediction of Coreceptor Usage of HIV-1 from Genotype and Application to Deep Sequencing. *Proceedings 18th International AIDS Conference (AIDS 2010)*, Vienna, Austria, 2010.

Horsthemke B, Wawrzik M, Groß S, Lich C, Sauer B, Rost I, Krasemann E, Kosyakova N, Liehr T, Weise A, **Dybowski JN**, Hoffmann D, Wieczorek D. Parental origin and functional relevance of a de novo UBE3A variant. *21. Jahrestagung der Deutschen Gesellschaft für Humangenetik*, Hamburg, Germany, 2010.

Dybowski JN, Heider D, Hoffmann D. Prediction of HIV-1 Co-receptor Usage from Genotype and Application to Deep Sequencing. *8. Forschungstag der Medizinischen Fakultät*, Essen, Germany, 2010.

Acknowledgements

> «I just want to thank everyone who made this day necessary.»
>
> – Yogi Berra

In addition to the words of the great Yogi Berra, I want to extend my deep appreciation towards a handful of people who have supported, encouraged, and motivated me throughout this work and my life.

My parents Hubert and Gitty to who I owe everything, but most essentially my life, free spirit, and open mind. Whether in foreign countries, big cities or small villages, you always made me feel at home.

My supervisor and mentor Daniel Hoffmann who taught me everything I know about scientific work. While letting me sample the scientific landscape, Daniel, fair-minded in his decisions and intrigued by every discovery, was always there to give directions.

My buddy in science Dominik Heider with whom I have developed countless ideas, discarded most, accomplished the best, and still cherish the boldest: Ahoi, Captain!

My oldest friend Manuel Prinz who will very soon share with me the development from a computer-playing kid to a number-crunching Doctor of natural science. A friend I could always rely on, work-related or private. Thank you for all the tissues!

My really oldest friend Phil who appreciates the value of a solid pitching performance and small-ball offense over a blow-out, multi-homer win as much as I do. Baseball is fun, but you're great.

My uncle Michael for the invitations to the annual Christmas concert and the enjoyable times we spent in Essen and Düsseldorf.

My colleagues for literally over one thousand cups of coffee, great discussions and tons of laughter.

My love Martina who valiantly endured the years of separation. I am looking forward to sharing a home and my life with you.

Curriculum Vitae

For reasons of confidentiality, the curriculum vitæ is not included in the online version of this work.

Declarations

Erklärung:
Hiermit erkläre ich, gem. § 6 Abs. (2) f) der Promotionsordnung der Fakultäten für Biologie und Geografie, Chemie und Mathematik zur Erlangung der Dr. rer. nat., dass ich das Arbeitgebiet, dem das Thema ”Predicting HIV-1 Co-receptor Usage of the Viral Quasispecies Using Classifier Ensembles” zuzuordnen ist, in Forschung und Lehre vertrete und den Antrag von J. Nikolaj Dybowski befürworte und die Betreuung auch im Falle eines Weggangs, wenn nicht wichtige Gründe dem entgegenstehen, weiterführen werde.

Essen, den ____________ ______________________________________

Unterschrift eines Mitgliedes der Universität Duisburg-Essen

Erklärung:
Hiermit erkläre ich, gem. § 7 Abs. (2) c) + e) der Promotionsordnung Fakultäten für Biologie und Geografie, Chemie und Mathematik zur Erlangung des Dr. rer. nat., dass ich die vorliegende Dissertation selbständig verfasst und mich keiner anderen als der angegebenen Hilfsmittel bedient habe.

Essen, den ____________ ______________________________________

Unterschrift des Doktoranden

Erklärung:
Hiermit erkläre ich, gem. § 7 Abs. (2) d) + f) der Promotionsordnung der Fakultäten für Biologie und Geografie, Chemie und Mathematik zur Erlangung des Dr. rer. nat., dass ich keine anderen Promotionen bzw. Promotionsversuche in der Vergangenheit durchgeführt habe und dass diese Arbeit von keiner anderen Fakultät/Fachbereich abgelehnt worden ist.

Essen, den ____________ ______________________________________

Unterschrift des Doktoranden