University of "Roma TRE"
Faculty of Computer Science Engineering
Department of Computer Science and Automation (DIA)
via della vasca navale, 79 – 00146 – Rome, Italy

Doctor of Philosophy Dissertation

Mathematical Models and Statistical Learning of HIV-1 Genetic Evolution Mechanisms under Drug Pressure for Treatment Optimisation

Mattia C.F. Prosperi

Supervisor: prof. Giovanni Ulivi

Reviewers: prof. Giulio Iannello, prof. Alessandro Cozzi-Lepri

Rome, 2007

a major issue concerning Artificial Intelligence is that Humans deal with it...

Contents

Acknowledgements

This PhD was sponsored by INFORMA Contract Research Organisation srl (via dei Magazzini Generali, 31 – 00154 – Rome, Italy) that contributed with a scholarship, training and many other types of support for the three years of research.

I'd like to thank in rigorous alphabetical order all the people who trusted in me and spent their effort in supporting my studies: Dario Corsini, Andrea De Luca, Francesca Incardona, Giovanni Ulivi and Maurizio Zazzi.

Many thanks to the reviewers prof. G. Iannello and prof. A. Cozzi-Lepri.

Thanks to the Automation group in Roma TRE: Alessandro Longhi, Stefano Panzieri, Federica Pascucci. Thanks to my colleague and friend Andrea, thanks also to Iuri.

Thanks to the Infectious Disease group at University of Sacro Cuore: Simona, Laura, Manuela, Sandrine...

Thanks to all the people from the Bioinformatics dpt of MPI in Saarbruecken where I spent a few months: in particular André, Jorg, Tobias and professor Thomas Lengauer.

I want to thank Mariella for the lovely time we spent together, along with all the friends from Tor Vergata (Ada, Andrea, Francesca, Michela, Valentina...)

Thanks to "Pucca" Bo-Kyung, who makes everyday bright!

Thanks to my apocalyptic friends Gamera ErMano, Ciccio, Dannyno Macca, Franko Sensei, Marco and Godzilla, along with wives and girlfriends and lovers.

Thanks to mom, dad, sister, grandparents and Medea!

Thank God, we have $\sqrt{ing}$ girls and other important things. Thank Buddha. Thank Darwin, and also the Anthropic Principle, even if weak.

Part I

Introduction

Chapter 1

Artificial Intelligence in Medicine and Biology

Recent years have seen medicine and artificial intelligence cross their roads and proceed together: after first genomic regions were sequenced and interpreted, the science scenario changed dramatically, fusing medicine, biology and biochemistry. Statistics have been always a valid support on epidemiology or diagnosis, but since the data became more complex, the needs included new mathematical models and high computational power. Today computer science – through machine learning and intelligent systems – integrates medicine and biology in several fields: from sequence analysis to protein structure and function prediction, to gene regulatory networks modelling, to molecular design, to medical diagnosis. Medical and biological data have been characterised theoretically, with the design of very large data bases with ad-hoc standard structures.

Biological systems are complex systems, medical measures are extremely variable – even under the same conditions – and indirect indicators of real processes. For instance, plasma analyses can describe only partially the body condition through specific markers as viral load, immune cell count, presence of chemicals or micro-organisms.

The HIV scenario sees the drug resistance development by the viral genomic variation under drug pressure: the high mutation rate determines a huge state variable space and the infection-replication (since the virus attacks the immune system which produces viral antibodies) convolve extremely complex mechanisms. There is a large number of different drugs that attack different viral targets genes: they have to be properly combined in order to control the viral suppression and the chance of resistance rise. Moreover, in the human body virus/drug interactions are affected by a host of co-factors, either uncontrollable or unobservable,

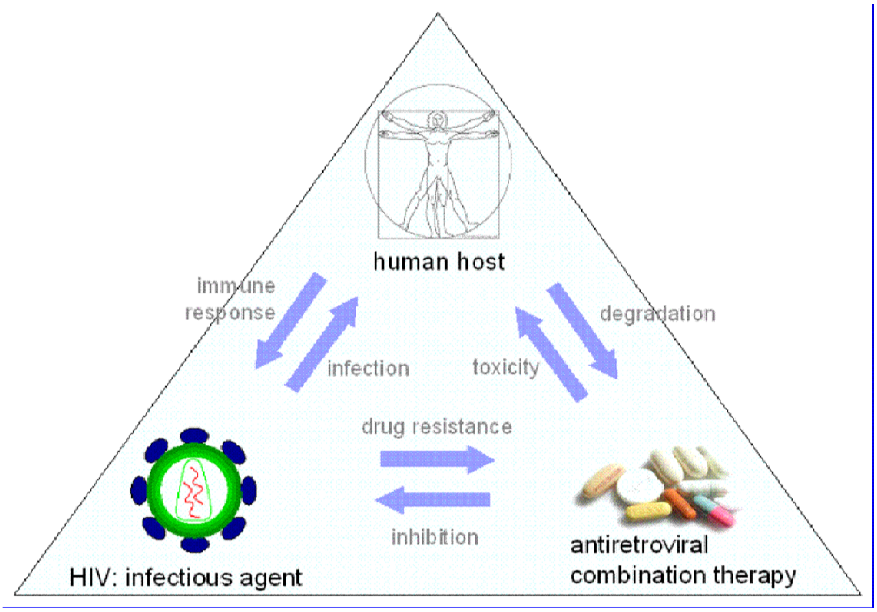

Figure 1.1: The viral infection scenario.

like adherence or toxicity (see figure 1.1).

1.1 Aims of the Thesis

The objective of this thesis is to model mathematically several aspects of the HIV-1 infection, replication and treatment design. The purpose is to investigate a mathematical framework that explains the biological mechanisms of the viral replication and infection in the human body, along with the immune response. This will be the basis on which the treatment settings will be explored in order to model the viral evolution towards the drug resistance: the final aim is to provide a better understanding of the genetic mechanisms involved in the development of drug-escaping mutants, along with predictive models for the treatment optimisation.

1.2 Road Map

This thesis is structured in three parts:

- Part I (Introduction) The aim of the thesis and the mathematical methodologies to be applied in the medicine/biology scenario are presented. An introduction to HIV and AIDS scenario is given, describing the general clinical and biological settings, the data collection policies, the issues concerning treatment design and resistance onset. A detailed biological description

of the infection and replication mechanisms in the human body is finally provided.

- Part II (Mathematical Modelling of HIV-1 Infection, Replication and Evolution) This part covers three main aspects of the viral evolution in the host: (i) viral replication dynamics through differential equations and stochastic modelling (accounting for drug pressure and resistance mutants onset); (ii) modelling of viral mutational pathways under selective drug pressure (through mutagenetic trees and Markov chains); (iii) statistical analyses (univariable, stratified) and unsupervised learning techniques (clustering, dimensionality reduction) for discovering relevant mutational patterns in equilibrium-like situations (naive/treated, success/failure groups).

- Part III (Statistical Learning for HIV-1) A description on the machine learning techniques used is given, along with discussion about validation (in-sample, extra-sample error estimation) and loss functions. Feature selection (filter, embedded and wrapper) and model comparison are extensively investigated. Models for the viral tropism prediction and for the in-vivo therapeutic optimisation are then applied and compared, providing results and discussions on the prediction performances, along with the opinions of biological experts on the model feature importance and prediction behaviours.

Chapter 2

Biological Background On HIV

Many microorganisms can enter the human body and cause harm, including viruses, fungi, bacteria, protozoa. Once inside the body, the primary goal of a microorganism is to survive and reproduce itself. Most antimicrobial agents are designed to kill these pathogens or prevent them from reproducing.

When a microorganism as a virus continues to replicate despite the pressure of a drug, mutants are selected that more efficiently adapt themselves to grow in the presence of a certain drug concentration: this results in the phenomenon of drug resistance. When drug resistance occurs, the efficacy of the drug – or combination of drugs – is reduced. Over time, the treatment can stop working completely. Evolution consists of a selective pressure from the environment that acts on organisms: it selects the best individuals from populations, favouring mutations that appear randomly on the gene pool; advantages acquired from mutations is transmitted to progeny.

The Human Immunodeficiency Virus (a *Lentivirus* belonging to two major families, HIV-1 and HIV-2) has a rapid rate of mutation and has developed through this resistance to antivirals. A brief introduction for non-biologists is given in [88], although we are going to describe more in depth its replication and infection mechanisms. If untreated, HIV-1 causes a progressive deterioration of the immune system leading almost relentlessly to AIDS (Acquired Immune Deficiency Syndrome) and death due to opportunistic infections. Modelling mechanisms of HIV drug-resistance requires the investigation the viral genome (which is in the form of RNA) and genes encoded within. A gene is a sequence of nucleotides (four varieties: Adenine, Cytosine, Thimine, Guanine), while the genome produces proteins that are important in the virus life cycle. A protein is a sequence of amino acids, which are encoded by blocks of three adjacent nucleotides in the genome, called codons. Genomic sequences are the building blocks of biological mechanisms: computer science is today necessary to investigate the

genes and their functions; even simple organisms like viruses are characterized by long character sequences. The basic theory for sequence analysis, which includes (multiple) alignment algorithms and phylogenetics, can be found in the book by Brunak [94], which is also a complete and generic guide for the whole set of derived subtasks.

In this chapter we will describe first in detail the biological mechanisms of HIV: a complete understandment of this part is not mandatory, since the basic needed biological concepts will be re-explained through the following chapters when facing specific modelling scenarios, but it is useful if the reader wants to make a more careful study.

2.1 In Depth on HIV-1 Biology

This section gives a more detailed description of the pathogenesis of HIV-1 infection and of the replication mechanisms. It is composed by material coming from the compendium "HIV Medicine 2006", freely available on the Internet [128]. We used this book as a reference for every biological and clinical aspect of HIV infection.

2.1.1 Pathogenesis of HIV-1 Infection

Since the initial description of the human immunodeficiency virus type I (HIV-1) in 1983 and HIV-2 in 1986, these two viruses have been identified for almost 20 years as the primary cause of the Acquired Immunodeficiency Syndrome (AIDS). As HIV-1 is the major cause of AIDS in the world today, our discussion will be primarily limited to HIV-1 infection. Worldwide, the number of HIV-1 infected persons exceeds 40 million, the majority of whom live in the developing countries of Sub-Saharan Africa, Asia and South America. In addition, new problems relating to the short- and long-term toxicity of drug treatments and the occurrence of resistance mutations in both circulating and transmitted viruses are emerging. In most countries in South East Asia and Africa, the incidence and prevalence of HIV-1 infection continues to increase and surpass that of Europe and North America. However, due to the high costs of drug regimens and the lack of a healthcare infrastructure in these developing countries, the widespread use of Anti Retroviral Therapy (ART) is currently still difficult. Even in countries where there is access to treatment there are only limited options for a first-line regimen and, possibly in the near future, access to a second regimen. The further course of the HIV-1 pandemic, therefore, mainly depends on how and to what degree the developing countries with a high HIV-1 prevalence are able to take advantage of the medical progress achieved in Europe and North America, and whether an effective prophylactic vaccine becomes available in the near future. An understanding of the immunopathogenesis of HIV-1 infection is a major prerequisite

for rationally improving therapeutic strategies, developing immunotherapeutics and prophylactic vaccines. As in other virus infections, the individual course of HIV-1 infection depends on both host and viral factors.

The course of infection with HIV-1 in HIV-infected humans may vary dramatically, even if the primary infections arose from the same source. In some individuals, with a long-term non-progressive HIV-1 infection (i.e. lack of decline in CD4+ T-cell counts, or chronic infection for at least 7 years without the development of AIDS), a defective virion was identified [75]. Thus, infection with a defective virus, or one that has a poor capacity to replicate, may prolong the clinical course of HIV-1 infection. However, in most individuals, HIV-1 infection is characterized by a replication-competent virus with a high daily turnover of virions and, in general, it is more likely that the rate of progression from HIV infection to AIDS is determined by multiple factors related to both the immune system and the virus.

Thus, host factors may also determine whether or not an HIV-1-infected individual rapidly develops clinically overt immunodeficiency, or whether this individual belongs to the group of long-term non-progressors, who represent about 5 % of all infected patients. The identification and characterisation of host factors contributing to the course of HIV infection, including immunological defense mechanisms and genetic factors, will be crucial for our understanding of the immunopathogenesis of HIV infection and for the development of immunotherapeutic and prophylactic strategies [44].

2.1.2 The Structure of HIV-1

HIV-1 is a retrovirus and belongs to the family of lentiviruses. Infections with lentiviruses typically show a chronic course of disease, a long period of clinical latency, persistent viral replication and involvement of the central nervous system. Visna infections in sheep, Simian Immunodeficiency Virus infections (SIV) in monkeys, or Feline Immunodeficiency Virus infections (FIV) in cats are typical examples of lentivirus infections. Using electron microscopy, HIV-1 and HIV-2 resemble each other strikingly. However, they differ with regard to the molecular weight of their proteins, as well as having differences in their accessory genes. HIV-2 is genetically more closely related to the SIV found in sootey mangabeys (SIV) rather than HIV-1 and it seems likely that it was introduced into the human population by monkeys. Both HIV-1 and HIV-2 replicate in CD4+ T-cells and are regarded as pathogenic in infected persons, although the actual immune deficiency may be less severe in HIV-2-infected individuals.

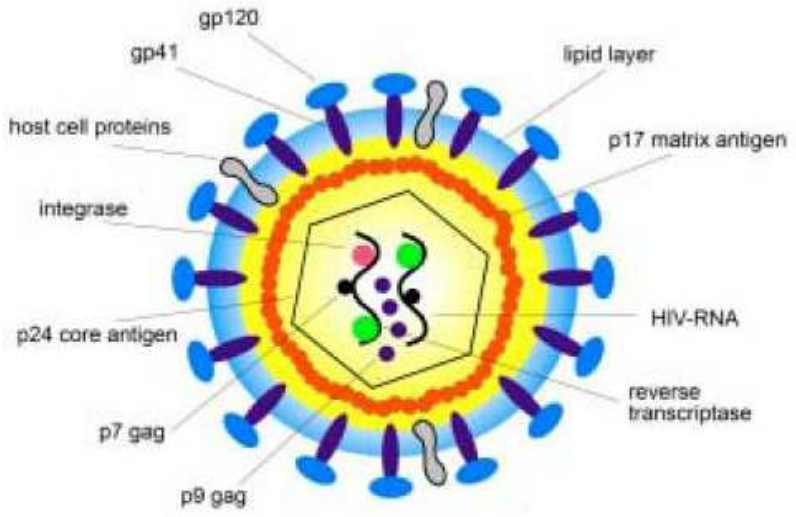

Figure 2.1: Structure of an HIV virion particle.

The Morphologic Structure of HIV-1

HIV-1 viral particles have a diameter of 100 nm and are surrounded by a lipoprotein membrane. Each viral particle contains 72 glycoprotein complexes, which are integrated into this lipid membrane, and are each composed of trimers of an external glycoprotein gp120 and a transmembrane spanning protein gp41. The bonding between gp120 and gp41 is only loose and therefore gp120 may be shed spontaneously within the local environment.

During the process of budding (when the virus after reproducing exits the infected cell), the virus may also incorporate different host proteins from the membrane of the host cell into its lipoprotein layer, such as HLA class I and II proteins, or adhesion proteins such as ICAM-1 that may facilitate adhesion to other target cells. The matrix protein p17 is anchored to the inside of the viral lipoprotein membrane. The p24 core antigen contains two copies of HIV-1 RNA. The HIV-1 RNA is part of a protein-nucleic acid complex, which is composed of the nucleoprotein p7 and the reverse transcriptase p66 (RT). The viral particle contains all the enzymatic equipment that is necessary for replication: a Reverse Transcriptase (RT), an Integrase p32 and a Protease p11 (figure 2.1).

The Organization of the Viral Genome

Most replication competent retroviruses depend on three genes (see figure 2.2): *gag, pol* and *env*: *gag* means group-antigen, *pol* represents polymerase and *env* is for envelope.

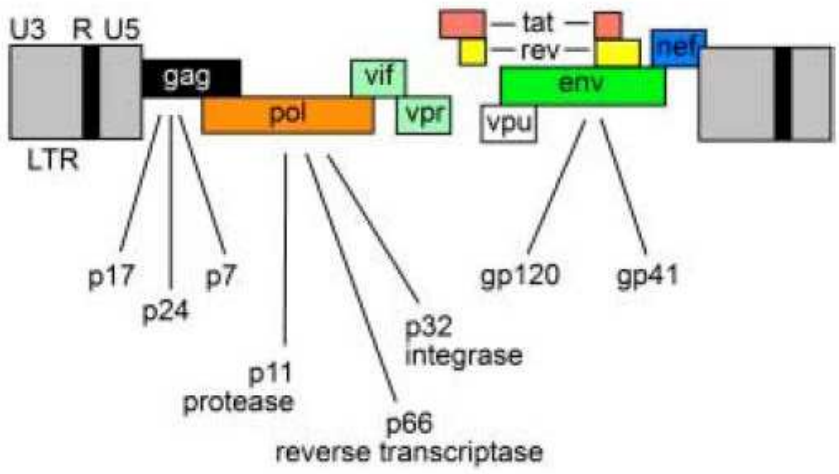

Figure 2.2: HIV genes.

The classical structural scheme of a retroviral genome is: *5'LTR-gag-pol-env-LTR 3'*. The LTR (Long Terminal Repeat) regions represent the two end parts of the viral genome, that are connected to the cellular DNA of the host cell after integration and do not encode for any viral proteins. The *gag* and *env* genes code for the nucleocapsid and the glycoproteins of the viral membrane; the *pol* gene codes for the reverse transcriptase and other enzymes. In addition, HIV-1 contains six genes (*vif, vpu, vpr, tat, rev, nef*) in its 9kB RNA that contribute to its genetic complexity. *Nef, vif, vpr, vpu* were classified as accessory genes in the past, as they are not absolutely required for replication in vitro. However, the regulation and function of these accessory genes and their proteins have been studied and characterized in more detail in the past few years. The accessory genes, *nef, tat, rev* are all produced early in the viral replication cycle. *Tat, rev* are regulatory proteins that accumulate within the nucleus and bind to defined regions of the viral RNA: TAR (transactivation-response elements), found in the LTR; and RRE (*rev* response elements), found in the *env* gene, respectively. The *tat* protein is a potent transcriptional activator of the LTR promoter region and is essential for viral replication in almost all in vitro culture systems. *Rev* is also a nuclear export factor that is important for switching from the early expression of regulatory proteins to the structural proteins that are synthesized later. *Nef* has been shown to have a number of functions: it may induce downregulation of CD4 and HLA class I molecules from the surface of HIV-1-infected cells, which may represent an important escape mechanism for the virus to evade an attack mediated by cytotoxic CD8+ T-cells and to avoid

recognition by CD4+ T-cells. *Nef* may also interfere with T-cell activation by binding to various proteins that are involved in intracellular signal transduction pathways. *Vpr* seems to be essential for viral replication in non-dividing cells such as macrophages. It may stimulate the HIV-LTR in addition to a variety of cellular and viral promoters. *Vpu* is important for the virus budding process, because mutations in *vpu* are associated with persistence of the viral particles at the host cell surface. *Vpu* is also involved when CD4-gp160 complexes are degraded within the endoplasmic reticulum and therefore allows recycling of gp160 for the formation of new virions. Some recent publications have highlighted a new and important role for *vif* in supporting viral replication. *Vif*-deficient HIV-1 isolates do not replicate in CD4+ T-cells, some T cell lines (non-permissive cells) or in macrophages. *Vif*-deficient isolates are able to enter a target cell and initiate reverse transcription, but synthesis of proviral DNA remains incomplete. In vitro fusion of permissive and non-permissive cells leads to a non-permissive phenotype, suggesting that the replication of HIV depends on the presence or absence of a cellular inhibitor.

2.1.3 The HIV replication cycle

HIV entry

CD4 as a primary receptor for HIV CD4 is monomeric glycoprotein that can be detected on the cell surface of about 60 % of T-lymphocytes, on T-cell precursors within the bone marrow and thymus, and on monocytes and macrophages, eosinophils, dendritic cells and microglial cells of the central nervous system. The extracellular domain of the CD4 on T-cells is composed of 370 amino acids; the hydrophobic transmembrane domain and the cytoplasmic part of CD4 on T-cells consist of 25 and 38 amino acids, respectively. Within the extracellular part of CD4, four regions D1-D4 have been characterized that represent immunoglobulin-like domains. Residues within the V2 region of CD4 (amino acids 40-55) are important for the bonding of gp120 to CD4 and this region overlaps the part of the CD4 where its natural ligands, HLA class II molecules, bind. The binding of gp120 to CD4 is not only a crucial step for viral entry, but also interferes with intracellular signal transduction pathways and promotes apoptosis in CD4+ T-cells. In the past couple of years, the idea of blocking CD4 as the primary cellular receptor of HIV has regained interest. CD4, as a primary and necessary receptor for HIV-1, HIV-2 and SIV, was already characterized in 1984. However, experiments, using non-human cell lines transfected with human CD4, showed that expression of human CD4 on the cell surface of a non-human cell line was not sufficient to allow entry of HIV. Therefore the existence of additional human coreceptors necessary for viral entry was postulated (i.e. CXCR4 and CCR5).

Chemokine receptors as coreceptors for HIV entry CD8 T cells from HIV-infected patients are able to suppress viral replication in co-cultures with HIV-infected autologous or allogenic CD4+ T-cells, and this is independent from their cytotoxic activity. In 1995 Cocchi identified the chemokines MIP-1a, MIP-1beta and Rantes in supernatants from CD8+ T-cells derived from HIV-infected patients, and was able to show that these chemokines were able to suppress replication in a dose-dependent manner of some, but not all, viral isolates tested. MIP-1a, MIP-1beta and Rantes are ligands for the chemokine receptor CCR5, and a few months later several groups were able to show that CCR5 is a necessary coreceptor for monocytotropic (M-tropic) HIV-1 isolates. A few weeks earlier, the chemokine receptor CXCR4 (fusin) was described as being the coreceptor used by T-cell-tropic (T-tropic) HIV isolates. Monocytotropic (M-tropic) HIV-1 isolates are classically those viruses that are most easily propagated in macrophage cultures, are unable to infect T-cell lines (i.e. immortalized T-cells), but are able to easily infect primary T-cells from peripheral blood samples. Conversely, T-cell-tropic HIV-1 isolates have classically been identified as being those that are easily propagated in T-cell lines, and grow poorly in macrophages, but are also able to easily infect primary T-cells from peripheral blood samples. Thus, it should be noted that both M-tropic and T-tropic HIV-1 variants can easily infect primary human non-immortalized T-cells in vitro. Chemokines (chemotactic cytokines) and their receptors have been previously characterized with regard to their role in promoting the migration (chemotaxis) of leukocytes and their pro-inflammatory activity. They are proteins of 68-120 amino acids which depend on the structure of their common cysteine motif, and which may be subdivided into C-X-C (alpha-chemokines), C-C (beta-chemokines) and C- chemokines. Chemokines typically show a high degree of structural homology to each other and may share the receptors they bind to. Chemokine receptors belong to the group of receptors with seven transmembranic regions (7-transmembrane receptors), which are intracellularly linked to G-proteins. SDF-1 (Stromal Cell-Derived Factor 1) was identified as the natural ligand of CXCR4 and is able to inhibit the entry of T-tropic HIV-1 isolates into activated CD4+ T-cells. Rantes (regulated upon activation T cell expressed and secreted), MIP-1alpha (Macrophage Inhibitory Protein) and MIP-1beta represent the natural ligands of CCR5 and are able to inhibit the entry of M-tropic HIV-1 isolates into T cells. A schematic model is depicted in Figure 2.3: T-tropic HIV-1 isolates mainly infect activated peripheral blood CD4+ T-cells and cell lines and use CXCR4 for entry into the CD4+ target cell. M-tropic isolates are able to infect CD4+ T-cells, monocytes and macrophages, and depend on the use of CCR5 and CD4 for viral entry. In this scenario, the pharmaceutical company Pfizer developed a CCR5 antagonist (called Maraviroc, which has been recently released in the market) in order to block the virus when entering the cell.

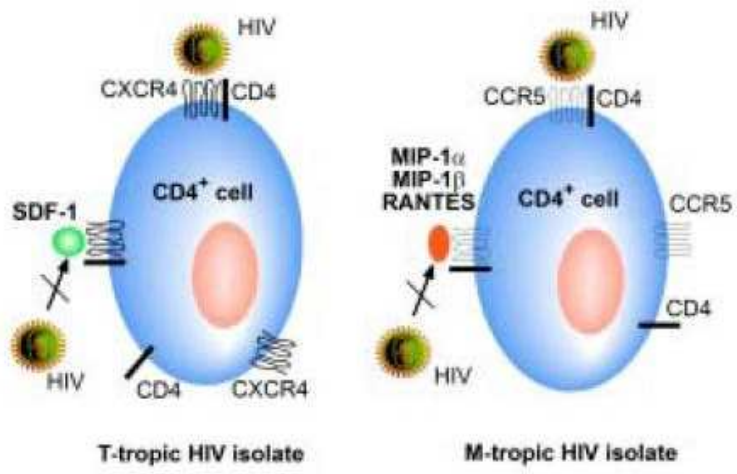

Figure 2.3: Inhibition of virus entry of CCR5-utilizing (monocytotropic) and CXCR4-utilizing (T-cell tropic) HIV isolates by the natural ligands of the chemokine coreceptors CCR5 and CXCR4.

The interaction of gp120 and the cellular receptors is now understood in more detail. Gp120 primarily binds to certain epitopes of CD4. Binding to CD4 induces conformational changes in gp120 that promote a more efficient interaction of the V3 loop of gp120 with its respective coreceptor. Membrane fusion is dependent on gp120 coreceptor binding. Gp41, as the transmembrane part of the envelope glycoprotein gp160, is crucial for the fusion of the viral and the host cell membrane. T20 is the first of several peptides that bind to gp41 and has been tested in clinical trials for suppressing viral replication. Using transfected cell lines, besides CCR5 and CXCR4, other chemokine receptors, such as CCR3, CCR2, CCR8, CCR9, STRL33, Gpr 15, Gpr 1, APJ and ChemR23, were identified and shown to be used for entry by certain HIV isolates. APJ may represent a relevant coreceptor within the central nervous system. Despite this broad spectrum of potentially available coreceptors, CCR5 and CXCR4 seem to represent the most relevant coreceptors for HIV-1 in vivo. The importance of CCR5 as the predominant coreceptor for M-tropic HIV isolates is underscored by another observation. The majority of individuals with a genetic defect of CCR5 are resistant to infection with HIV-1 [31]. In vitro experiments show that lymphocytes derived from these individuals are resistant to HIV-1 infection using M-tropic isolates but not to infection with T-tropic isolates. Lymphocytes from these individuals do not express CCR5 on their cell surface and genetically have a 32 base pair deletion of the CCR5 gene. Worldwide, a few patients have been identified

that have acquired HIV-1 infection despite a homozygous deletion of the CCR5. As expected, all of them were infected with CXCR4-using HIV-1 isolates. In epidemiological studies, the allelic frequency of the CCR5 gene deletion is 10-20% among Caucasians, particularly amongst those of Northern European descent. The frequency of a homozygous individual is about 1% in Caucasians. Studies conducted on African or Asian populations, however, do not find this 32 base pair deletion of the CCR5, suggesting that this mutation arose after the separation of these populations in evolutionary history. Individuals that are heterozygous for the 32 bp deletion of the CCR5 show a decreased expression of CCR5 on the cell surface and are more frequently encountered within cohorts of long-term non-progressors compared to patients who have a rapid progression of disease. In addition to the 32bp deletion of the CCR5, other genetic polymorphisms, withregard to the chemokine receptors (CCR2) or their promoters (CCR5), have been described. Based on the occurrence of these polymorphisms within defined patient cohorts, they were associated with a more rapid or a more favourable course of disease, depending on the particular polymorphism. Recently a whole genome analsis has identified other 3 potential hits associated with a low viral set point and slow rate of progression in untreated individuals [44]. In patients who have a rapid progression of disease (rapid drop in CD4+ T-cell count), virus isolates that use CXCR4 as a predominant coreceptor tend to be frequently isolated from their cells, in comparison to patients with a stable CD4+ T- cell count. The expression of coreceptors on CD4+ T-lymphocytes depends on their activation level. CXCR4 is mainly expressed on naive T-cells, whereas CCR5 is present on activated and effector/memory T-cells. During the early course of HIV-1 infection, predominantly M-tropic HIV-1 isolates are detected. Interestingly, M-tropic HIV-1 isolates are preferentially transmitted regardless of whether or not the donor predominantly harbours T-tropic isolates. At present, it remains unclear whether this in vivo preference of M-tropic HIV-1 isolates is determined by selected transportation of M-tropic isolates by sub-mucosally located dendritic cells or whether the local cytokine/chemokine milieu favors the replication of M-tropic viruses. Recent studies suggest that M-tropic HIV-1 viruses are able to hide more easily from the immune system by replicating in macrophages, in comparison to T-tropic viruses, thus giving them a survival advantage in the infected individual. The blockade of CCR5, therefore, seems to represent a promising target for therapeutic intervention, as we already mentioned citing the Pfizer's Maraviroc CCR5 antagonist.

Postfusion Events

Following membrane fusion the virus core uncoats into the cytoplasm of the target cell. These early events have recently been studied in more detail. HIV can enter into rhesus lymphocytes but replication is stopped before or during

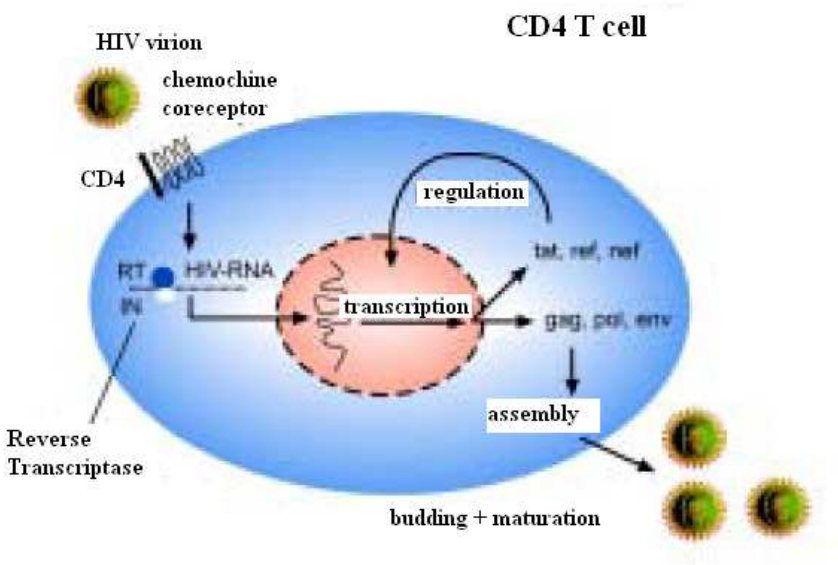

Figure 2.4: Life cycle of HIV.

early reverse transcription. This intracellular blockade is mediated by a cellular factor, TRIM5a, which is a component of cytoplasmic bodies and whose primary function is yet known. HIV-1 entry into quiescent T cells is comparable to HIV-1 entry into activated T cells, but synthesis of HIV-1 DNA remains incomplete in quiescent cells. The conversion of viral RNA into proviral DNA, mediated by the viral enzyme reverse transcriptase (RT), occurs in the cytoplasm of the target cell and is a crucial step within the viral replication cycle (see figure 2.4).

Blockade of the RT by the nucleoside inhibitor zidovudine (AZT) was the first attempt to inhibit viral replication in HIV-1 infected patients. Reverse transcription results in double-stranded HIV DNA with LTR regions (Long Terminal Repeats) at each end. HIV-1 enters into quiescent T cells and reverse transcription may result in the accumulation of proviral, non-integrating HIV-DNA. However, cellular activation is necessary forintegration of the proviral HIV DNA into the host cell genome after transportation of the pre-integration complex into the nucleus. Cellular activation may occur in vitro after stimulation with antigens or mitogens, in vivo activation of the immune system is observed after antigen contact or vaccination or during an opportunistic infection. In addition, evidence is emerging that HIV-1 gp120 itself may activate the infecting cell to enhance integration. Besides monocytes, macrophages and microglial cells, latently infected quiescent CD4+ T-cells that contain non-integrated proviral HIV DNA represent important long-living cellular reservoirs of HIV. Since natural HIV-1 infection is characterized by continuing cycles of viral replication in activated CD4+ T-cells, viral latency in these resting CD4+ T-cells likely represents an

accidental phenomenon and is not likely to be important in the pathogenesis of this disease. This small reservoir of latent provirus in quiescent CD4+ T-cells gains importance, however, in individuals who are treated with cART, since the antivirals are unable to affect non-replicating proviruses and thus the virus will persist in those cells and be replication competent to supply new rounds of infection, if the drugs are stopped. Thus, the existence of this latent reservoir has prevented HAART from entirely eradicating the virus from infected individuals. Persistence of HIV in quiescent CD4+ T-cells and other cellular reservoirs seems one of the main reasons why eradication of HIV is not feasible. If it is ever possible to achieve, a more detailed knowledge of how and when cellular reservoirs of HIV are established and how they may be targeted is of crucial importance for the development of strategies aiming at HIV eradication. Cellular transcription factors such as NFkB may also bind to the LTR regions. After stimulation with mitogens or cytokines, NFkB is translocated into the nucleus where it binds to the HIV-LTR region, thereby initiating transcription of HIV genes. Transcription initially results in the early synthesis of regulatory HIV-1 proteins such as tat or rev. Tat binds to the TAR site (Transactivation Response Element) at the beginning of the HIV-1 RNA in the nucleus and stimulates transcription and the formation of longer RNA transcripts. Rev activates the expression of structural and enzymatic genes and inhibits the production of regulatory proteins, therefore promoting the formation of mature viral particles. The proteins coded for by pol and gag form the nucleus of the maturing HIV particle; the gene products coded for by env form the gp120 spikes of the viral envelope. The gp120 spikes of the envelope are synthesized as large gp160 precursor molecules and are cleaved by the HIV-1 protease into gp120 and gp41. The gag proteins are also derived from a large 53 kD precursor molecule, from which the HIV protease cleaves the p24, p17, p9 and p7 gag proteins. Cleavage of the precursor molecules by the HIV-1 protease is necessary for the generation of infectious viral particles, and therefore the viral protease represents another interesting target for therapeutic blockade. The formation of new viral particles is a stepwise process: a new virus core is formed by HIV-1 RNA, gag proteins and various pol enzymes and moves towards the cell surface. The large precursor molecules are cleaved by the HIV-1 protease, which results in the infectious viral particles budding through the host cell membrane. During the budding process, the virus lipid membranes may incorporate various host cell proteins and become enriched with certain phospholipids and cholesterol. In contrast to T cells, where budding occurs at the cell surface and virions are released into the extracellular space, the budding process in monocytes and macrophages results in the accumulation of virions within cellular vacuoles. The replication of retroviruses is prone to error and is characterized by a high spontaneous mutation rate. On average, reverse transcription results in 1^{-10} errors per genome and per round of replication. Mutations can lead to the

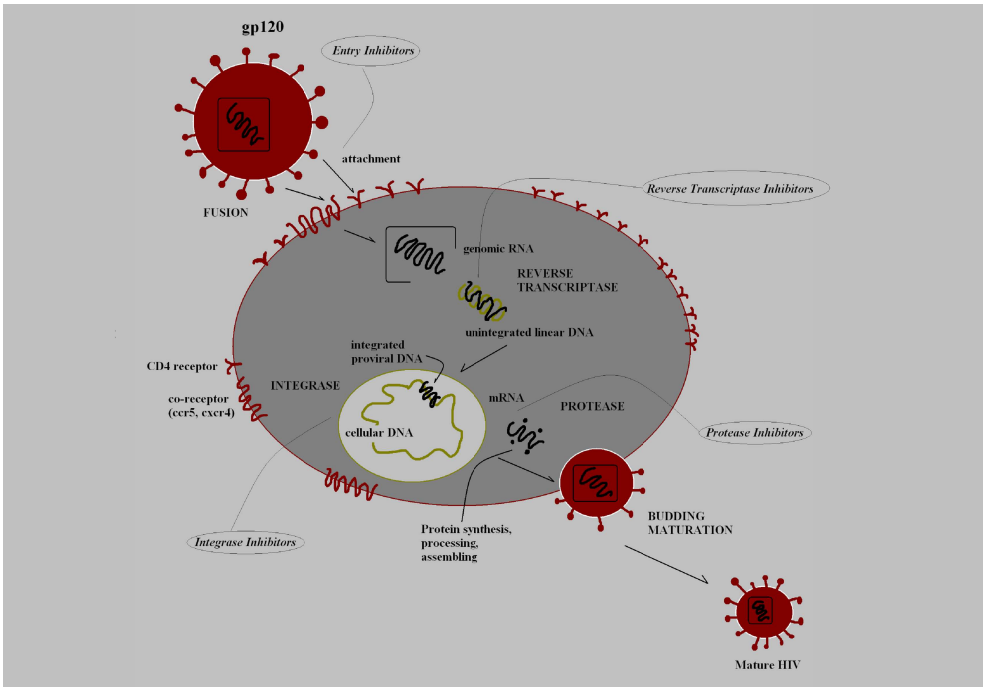

Figure 2.5: HIV-1 Replication Process and Drug Targets

formation of replication-incompetent viral species. But, mutations causing drug resistance may also accumulate, which, provided that there is selection pressure under certain antiretroviral drugs and incomplete suppression of viral replication, may become dominant. In addition, viral replication is dynamic and turns over quickly in infected individuals at an average rate of 10 new virus particles being produced and subsequently cleared per day. Thus, within any individual, because of the extensive virus replication and mutation rates, there exists an accumulation of many closely related virus variants within the population of viruses, referred to as a viral quasispecies. The selection pressure on mostly the pre-existing mutations may not only be exerted by certain drugs, but also by components of the immune system, such as neutralizing antibodies or cytotoxic T cells (CTL). However there is not very strong evidence that CTL plays a major role in controlling HIV infection [12].

2.2 Treatment Design against HIV Replication

Antiretroviral drugs act by blocking the functions of certain viral genes: by now they are reverse transcriptase, protease, integrase and entry – see figure 2.5 – but the virus can develop mutations to escape the drug inhibition.

Summarising the above description, in the virus life cycle (when it reproduces) the genome string has to be copied from a generation to the next one. Soon after

HIV enters the body, the virus begins reproducing at a rapid rate and billions of new viruses are produced every day. In the process, HIV produces both perfect copies of itself and copies containing errors (and copying errors occur frequently). The so-called *wild type* virus is the strain that has naturally evolved as it has the highest replication capacity: before therapy is started, it is the most frequent strain in the body and dominates all other quasispecies.

Mutations can change the viral structure or its functions and then modify its interaction with the environment: the high mutation rate of HIV (combined with the fact that it attacks the immune system) leads to difficulties in the design of a vaccine, and to a rapid selection of mutant strains that are likely to be resistant to antiretroviral drugs. The first Anti Retro-viral drug to be introduced was Zidovudine (AZT) in 1987, initially made from the anchovies' sperm.

At present a few classes of drugs are approved by the FDA (Food and Drug Administration) of USA and EMEA in Europe as antiviral treatment against HIV: these are the Reverse Transcriptase inhibitors (RTI, divided in Nucleoside/tide NRTI and Non-Nucleoside NNRTI), Protease inhibitors (PI) Fusion inhibitors (FI) or Entry Inhibitors (EI) and Integrase Inhibitors (II); each class acts against a step of the viral replication process, and there are around 20 different molecules available in the clinics. A summary is given in table 2.1.

By comparing the nucleotide sequence of a viral isolate against a wild type subtype reference (in fact HIV-1 harbours different subtypes, resulting from different lineages of evolution), mutations in the genome can be extracted. Usually mutations are identified with a number representing a codon (a position in the genomic sequence), headed by a letter that indicates the amino acid present in the wild type (i.e. the standard virus, without mutations) and followed by another letter that describes the amino acid replaced in the mutant. For instance, a mutation that usually confers resistance to the antiretroviral drug Lamivudine (3TC) is the M184V: it indicates that in codon 184 the amino acid Methonine (M) has been replaced by Valine (V).

During infection there is not a single virus in a persons body, but a large population of mixed viruses called quasispecies. Mutant variants are too weak to survive and/or cannot reproduce as efficiently as the wild-type.; as a result, they are under-represented in the body. A drug usually works by blocking a key step of the virus life cycle. Some variants have mutations that allow the virus to be partly, or even fully, resistant to an antiretroviral drug. In a person receiving continuous therapy, mutant resistant strains have an advantage over wild type strains and can become dominant in the patient. This is called selective resistance, because the mutant is selected by the drug and results in treatment losing its efficacy. Treatment of selected mutants is more challenging because therapy options are reduced. If there remains viral replication despite a change in drug regimen, new mutations can be selected and, furthermore, there are mutations (such as

commercial name	acronym
Nucleoside/tide	
retrovir	AZT
combivir	AZT+3TC
epivir	3TC
emtriva	FTC
epzicom	ABC+3TC
trizivir	ABC+3TC+AZT
truvada	TDF+FTC
videx	DDI
viread	TDF
zerit	D4T
ziagen	ABC
Non-Nucleoside	
rescriptor	DLV
intelence	TMC125
sustiva	EFV
viramune	NVP
Protease	
agenerase	APV
lexiva	FPV
norvir	RTV
prezista	DRV
reyataz	ATV
aptivus	TPV
crixivan	IDV
invirase	SQV
fortovase	SQV
kaletra	LPV
viracept	NFV
Fusion-Entry	
fuzeon	T20
selzentry	Maraviroc
Integrase	
isentress	Raltecavir

Table 2.1: List of FDA approved ARV drugs for HIV-1 treatment.

the insertion at codon 69 in the Reverse Transcriptase gene) that cause cross-resistance to a whole class of antiretrovirals (NRTI-inhibitors). It has been shown that during treatment interruptions as short as 2 months, HIV reverts to wild type, but maintains low concentrations of the resistant mutants, so if a previously experienced drug is reused resistance shortly arises.

Because the virological potency of single drugs is limited, combined thera-

pies (i.e. involving multiple drugs used concomitantly) are an approach to avoid resistance. If a virus becomes resistant to a drug, but it is inhibited by many different others, it can still be suppressed to undetectable levels (although complete eradication is not possible with current regimens). Combination therapy is often called cART (combined Anti Retroviral Therapy) or HAART (Highly Active Anti Retroviral Therapy), when more than two drugs in different classes are administrated. Mutations can also occur under cART or HAART, even though at a lower rate. The era of HAART began in 1996 with the introduction of PI and NNRTI regimens and resulted in dramatic improvements in morbidity and mortality of HIV disease, as shown by a decreased incidence of opportunistic infections, tumours, and deaths. Despite all the therapeutic advantages achieved during the last decade, including the development of cART, once an individual has become infected, eradication of the virus still remains impossible. Usually cART produce a marked reduction of plasma viral load within three-four weeks from starting the drugs, and can be sustained for a long period of time (current estimates of the rate of viral rebound in patients starting from drug-naive are in the range of 8% per year [14]. HAART typically containing from three to five different drugs (usually two NRTI and at least one PI – sometimes enhanced by a small dose of RTV – or an NNRTI), are highly potent against HIV but its use is often associated with tolerance and toxicity problems.

2.3 Experimental Settings and Data Collection

Before being approved and commercialised by the FDA, drugs follow a long procedure in which their efficacy is tested through different phases (namely, phase I to phase IV): first they are designed, synthesised and put in viral cultures; if they are proven to be effective in-vitro, they will be tested for absorption levels and toxicity in-vivo, until they are judged to be relatively safe in humans and effective in suppressing HIV replication. In-vitro studies however are always carried out – even after the introduction of the drug into the market – in order to fully characterise the resistance profile of the compund. In-vitro and in-vivo studies provide data that can be used for mathematical and statistical modelling.

In-vitro studies are collections of experiments that measure how a mutated virus responds in culture to inhibition by a single drug, compared with the replication of the wild type under the same drug pressure: the *phenotype* is a numeric indicator of viral replication power, expressed as *Fold Change* of the drug concentration needed to inhibit 50% of the viral replication as compared to a wild type drug-susceptible reference viral strain: the data sets are typically pairs of genotype sequences and fold changes (FC) values. For each antiretroviral drug there are currently thousands of such genotype-FC Pairs freely available and the quality of the data – standing fixed environment conditions and repeatability –

is fairly high. These tests are expensive compared to the cost of sequencing a viral strain, so a first challenge for mathematical modelling was to try to estimate phenotypic FC values associated with the corresponding genotype sequences.

In-vivo studies are usually conducted using data collected from patients in clinical practice , i.e. plasma viral loads measured before and after a therapy switch: usually a therapy is stopped and considered failing when the Viral Load* in the patient's blood is greater than a certain threshold (current HIV-treatment guidelines recommend a switch in therapy if viral load is > 500 copies/mL) and/or the CD4+ T† cell counts are very low; when there are is an appreciable number of copies of the virus in plasma (e.g. > 500 copies/mL), HIV can be sequenced to identify the selected mutations. Unfortunately, both because of technical and political reasons, large data sets containing both patients genotypic and clinical data are rare. Data coming from randomised control trials (based on precise therapeutic protocols usually led by a team of physicians and where patients are controlled weekly) are normally the most reliable ones, but the access to these data is not always free and their sample size is typically small; in contrast, observational cohort studies or clinical cohort databases(collections of clinical reports from one or more hospitals) have a large sample size, but suffer from potential biases due to time delays, missing data and the inclusion of non-adherent patients. In addition, in-vivo viral load measurements are potentially biased by systematic errors due to intrinsic limitations of the viral load assays currently in use: indeed, viral load measures are reliable within 1 Log and assays in use in the clinic cannot detect viral load when the number of copies is under certain limits (e.g. 500 or 50 cp/ml); genotype sequencing methods have an accuracy of 90% in detecting mutations if they are present, but performances decrease using plasma samples with low viral concentrations. Even input errors are not negligible: data bases are often not automated and have inefficient relational structures and implementations, mostly data are recorded manually from paper clinical reports to spreadsheets. In a clinical database, thousands of variables are recorded but the variability of in vivo data is extremely high and the space of investigation is large: for example, in theory there could be $\tilde{}20^{400}$ possible genotypes (in the sequenced region of *polymerase*) and $\binom{20}{1}+\cdots+\binom{20}{5}$ therapeutic combinations. But in practice only certain cART are allowed, either for toxicity reasons, either because the treatment guidelines encourage combinations that have been proven to be effective and discourage a combination different from 2NRTI+1NNRTI or 2NRTI+1PI.

*Viral Load is the virion count in the plasma

†CD4+ T are immune cells targeted and infected by the virus and, therefore, a measure of immuno-suppression

Part II

Mathematical Modelling of HIV-1 Infection, Replication and Evolution

Chapter 3

Overview on HIV-1 Infection, Replication and Evolution

The investigation of HIV-1 infection, replication and evolutionary mechanisms can be devised in two main scenarios:

- spreading and evolution within the population (*macro*-evolution)
- mechanisms of infection, replication and evolution within the single infected organism with or without drug pressure (*micro*-evolution)

The first scenario concerns HIV phylogenetics and epidemiology: the scope of this thesis is not to investigate such a field, but some interesting connections can be found between the spreading of an infectious disease in the population (in terms of subtype characterisation) and the evolution of resistant mutations within one treated single individual. For this reason, is useful to give a set of proper references: for the phylogeny in general, the book by Salemi [124] is probably one of the best in commerce for practical applications, while for more theoretical aspects the books by Felsenstein [45] and Nei [42]; for epidemiology, apart from the standard statistical modelling techniques, we suggest the articles by Goodreau [57] that investigate interesting approaches based on complex-networks.

The second point refers to the mathematical models derived so far in order to describe and simulate the viral infection in the human body, the replication in the cell and the genetic evolution: currently, several methodologies are used: from ordinary deterministic (or stochastic) differential equations (mainly for the infection-replication mechanisms), to cellular automata, to probabilistic networks (for the mutational pathways modelling). They explore (together or separately)

- viral interactions with human cells

- pharmacokinetics/pharmacodynamics
- resistant viral strain evolution

For the case of differential equations, all the models of the infection-replication phase can be seen as extensions/reinterpretations of the basic predator-prey model [127].

There are some crucial points that see a partial failure of most (if not all) of these models: the extreme complexity of human metabolism, the imprecision in measuring and partial un-observability of key variables. For instance, in-vivo clinical markers for the majority of patients consist of viral RNA charge, CD4+ and (sometimes) CD8 cell counts: differentiation between infected and un-infected cells cannot be estimated unless estimations coming from specific experiments are made; no information about latently infected cells is available; viral sequences resemble the predominant population in the plasma.

During most ARV treatments HIV can follow different mutational pathways to develop resistance, depending on drug pressure, viral population size and fitness of the resistant mutants. Knowledge about the probability and time needed to select different resistant strains, joined with pathways description, can increase confidence about long-term therapies. Even if HIV-1 is a simple RNA virus and the resistant mutations develop through copying errors, modelling the *evolution* of resistant quasi-species is far from being a simple problem, being similar to the subtype evolution and spreading of the viral population in the world.

A stochastic model, Markov Chains and Mutagenetic Trees will be the approaches of present thesis towards the objective of dynamics modelling, while univariable and multivariable statistics will be executed on the equilibrium states.

Chapter 4

Differential Equations and Stochastic Modelling

4.1 Overview

Several mathematical models have been investigated for the description of viral dynamics in the human body: HIV-1 infection is a particular and interesting scenario, because the virus attacks cells of the immune system that have a role in the antibody production and its high mutation rate permits to escape both the immune response and, in some cases, the drug pressure. The viral genetic evolution is intrinsically a stochastic process, eventually driven by the drug pressure, dependent on the drug combinations and concentration: in this paper the viral genotypic drug resistance onset will be the main focus addressed. The theoretical basis will be the modelling of HIV-1 population dynamics as a logistic equation with a time dependent therapy efficacy term, while the viral genome mutation rate will be a stochastic process and follow a poisson distribution. Finally, the instant probabilities of drug resistance will be estimated by means of sigmoid functions trained from in-vitro phenotypes. The drug-resistance modelling framework allows also for inclusion of more detailed viral-immune-inhibition population dynamics.

The stochastic-logistic modelling usefully predicted long-term virologic outcomes of evolved HIV-1 strains for selected antiretroviral therapy combinations. Simulations were run for mono- and bi-therapies, for highly active antiretroviral therapies and for unadherent therapies. Sequential treatment change episodes (including drug sparing and recycling) were also exploited with the aim to evaluate optimal synoptic treatment scenarios. For a set of widely used combination therapies, results were consistent with findings reported in literature.

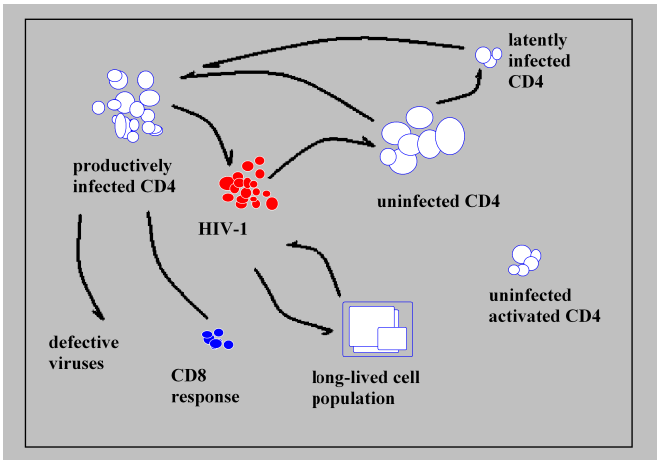

Figure 4.1: HIV-1 Target Cells and Infection Process

4.1.1 Biological Mechanisms of HIV-1 Infection

Infection by HIV-1 is a particularly interesting scenario because this virus directly attacks cells of the immune system and its high mutation rate permits to escape both the immune response and, in some cases, the drug pressure.

HIV-1 primarily infects vital cells of the human immune system such as helper T cells (specifically T cells with *CD4+* receptor), macrophages and dendritic cells. The exact process through which HIV-1 leads to immuno-suppression and eventually to Acquired Immuno-Deficiency Syndrome (AIDS) is still debated, but several mechanisms have been postulated by experimental evidences. Upon entering a target cell, the viral RNA genome of HIV is converted to a double-stranded DNA by the reverse transcriptase gene, within the *polymerase* viral genomic region. The viral DNA is then integrated into the cellular DNA by the integrase gene, so that the genome can be transcribed. Finally, the protease gene allows the production of new infectious virions. HIV-1 can also hide latently in infected cells, waiting quiescently until some environmental condition activates its reproduction.

The host immune response and the viral infection are connected within three principal mechanisms: (i) direct viral killing of infected cells; (ii) increased rates of apoptosis in infected cells; (iii) killing of infected CD4+ T cells by CD8 cytotoxic lymphocytes that recognise infected cells (figure 4.1). When CD4+ T cell numbers decline below a critical level, cell-mediated immunity is lost, and the body becomes progressively more susceptible to opportunistic infections. If untreated, eventually most HIV-infected individuals develop AIDS and/or die for opportunistic infections.

HIV differs from many other viruses as it has very high genetic variability. This diversity is a result of its fast replication cycle, with the generation of 10^9 to

10^{10} virions per day, coupled with a high mutation rate of approximately 3×10^{-5} per nucleotide base per cycle of replication. This complex scenario leads to the generation of many variants of HIV in a single infected patient in a brief time-span. The fact that the immune system is not able to control the infection is also due to this high genetic variability.

4.1.2 State of the Art on HIV-1 In-Vivo Dynamics Modelling

The literature includes a plethora of systems of ordinary deterministic differential equations (ODE), that consider different aspects of the infection, viral replication, immune response and treatment effect (accounting also for pharmacokinetics). So far, from the naive application of *predator-prey* [127] equations, the models have been extended to account for *competition and cooperation* of cohabitant species (i.e. virus, immune response, latent cells, et cetera). In order to model the viral resistance to drug pressure, refinements of these equations have been produced: from simple modifications of the ODE using time-invariant coefficients for drug efficacy, to multi-strain models to describe viral mutants. Stochastic and cellular automata models have been also introduced, investigating mainly population dynamics and immune response behaviours.

A milestone is the system introduced by [11] where the basic model presented is:

$$\frac{dT}{dt} = s + pT(1 - \frac{T}{T_{max}}) - d_T T - KV_i T \tag{4.1}$$

$$\frac{dT^*}{dt} = (1 - \eta_{RT})kV_i T - \delta T^* \tag{4.2}$$

$$\frac{dV_i}{dt} = (1 - \eta_{PR})N\delta T^* - cV_i \tag{4.3}$$

$$\frac{dV_{ni}}{dt} = \eta_{PR} N\delta T^* - cV_{ni} \tag{4.4}$$

where T are uninfected CD4+ T cells, T^* are infected CD4+ T cells, V_i and V_{ni} are infectious and non-infectious virions respectively, η_i are the drug class efficiencies (here protease and reverse transcriptase). Cytotoxic response (CD8), latent and long-lived cells are not included in the model, neither the presence of multiple viral strains. Refinements of this basic model – amongst the many *variation on a theme* – include: further studies by [92], that take into account intra-cellular time delays; a recent work by [92] on latently infected cell activation; several papers by [4] [5] that analyse simplified bi-phasic exponential slopes for viral load changes, along with accurate parameter estimation; similar models by [131] [85] and a recent literature review by [78]. A major concern regarding all the ODE-based models is in the efficacy of (combination) therapies is modelled only by means of constant terms, that shrink the viral reproduction rates. The knowledge of this efficacy must be known a-priori and does not depend on (nor drives) the viral genotype evolution.

Multiple-strain models (i.e. different equations for different viral mutant species) are an attempt to explain the drug resistance onset, where the constant terms for drug efficacy have different values for each mutant specie. A drawback of this approach is that the viral population is simply divided in a few additional equations, that represent resistant and susceptible population for different inhibitors. In [104] population dynamics are analysed using an ODE model and a stochastic one. A different approach is to consider the development of mutations as a process directly driven by the drug pressure, as proposed by [103] and [118].

A paper which has marked similarities with the approach here presented is the one by [80]: the study investigates the evolution of resistant viral strains under time-varying drug combinations. A very simple scenario made by three *ideal* drugs and three corresponding ideal resistance mutations is given, with different replication parameters (arbitrarily fixed) for each mutant under a certain drug pressure. Numerical simulations are carried out to estimate viral evolution, using a sampling from a Poisson distribution to determine whether mutations occur in any one round of replication and to calculate the number of cells created in the next generation. However, it is stated clearly that the model allows only for general considerations on treatment change policies or unadherence problems. There is no attempt to create a specific model using real drugs and actual HIV mutation rates.

The last set of models present in literature uses cellular automata to describe HIV-1 strategies of immune evasion, by [18] [17]: mutation, fitness, viral diversity and predictive markers of disease progression are analysed, but only relative to the infection of host cells, without accounting for drug administration effects, neither drug-resistance emergence.

4.2 Logistic Stochastic Model: Theory

The proposed model is based on a few assumptions about the rate of HIV-1 replication in humans. Most of these are consistent with the current scientific knowledge. First, it is assumed that patients are chronically infected with HIV-1, i.e. they are in a clinical latency stage some time after the date of seroconversion and at low risk of developing constitutional symptoms and AIDS-defining events. This corresponds to a steady-state equilibrium in which the viral load has reached a constant value in the host, with a corresponding constant immune cell count. Secondly, it is assumed that neither CD4+ or cytotoxic T lymphocytes have a direct role in controlling viral dynamics: the virus, when perturbed, tends to go back to the steady-state following a logistic reproduction equation as described by [125]. The quadratic term that shrinks the growth is intended to account intrinsically for the interaction between virions and immune system. Latently

infected cells are instead took into account, allowing for a special reservoir in which the virions are not reached by the drugs. The usage of this simplified population dynamics equation is justified by the fact that has an exact solution when considering constant parameters.

The continuous version of the logistic model is described by the differential equation

$$\frac{dV}{dt} = \frac{rV(V_{max} - V)}{V_{max}} \tag{4.5}$$

where V is the viral load, $r > 0$ is the Malthusian parameter (rate of population growth) and V_{max} is the so-called carrying capacity (i.e. the maximum sustainable population). Dividing both sides by V_{max} then gives the differential equation

$$\frac{dV}{dt} = rV(1 - \frac{V}{V_{max}}) \tag{4.6}$$

The solution to the equation (with V_0 being the initial population) is

$$V(t) = \frac{V_{max}V_0e^{rt}}{V_{max} + V_0(e^{rt} - 1)} \tag{4.7}$$

Usually the growth rate term r is a positive constant, but for negative values it can lead the population to extinction. Suppose then that the application of a therapy reduces the population growth, shrinking the viral reproduction by a $\eta \in [0, 1]$ constant term. It can be modelled by the following modified equation

$$\frac{dV}{dt} = (1 - \eta)rV - \frac{rV^2}{V_{max}} \tag{4.8}$$

which can be rewritten as

$$\frac{dV}{dt} = (1 - \eta)rV(1 - \frac{V}{(1 - \eta)V_{max}}) \tag{4.9}$$

resulting in a new steady state of $(1-\eta)V_{max}$. Alternatively, it would be possible to define the more general equation

$$\frac{dV}{dt} = rV(1 - \frac{V}{V_{max}}) - sV \tag{4.10}$$

in which the term s in not simply a percentage reduction of the growth rate (like it was η in the previous equation). Rewriting the equation in

$$\frac{dV}{dt} = (r - s)V(1 - \frac{V}{\frac{r-s}{r}V_{max}}) \tag{4.11}$$

depending on the values of r and s, different equilibrium could be reached, like the complete eradication if $r < s$ (remember that r is positive by definition).

The assumption of a constant effect of therapy however is fairly unrealistic – this is also the shortfall of the ODE models previously described – because there

is evidence that under drug pressure the virus develops escaping mutations and the viral load ultimately rebounds to V_{max}.

Instead of defining explicitly a set of equations for each possible resistant strain (which is the approach of the multi-strain models), a coefficient $\eta(t) = \eta(V(t), C(t)) \in [0, 1]$ can be introduced, which will depend on the viral population $V(t)$ – identified by a set of mutations in the viral genome present at time t and dependent on the combination therapy C.

The viral population V is a set of single individuals $V = v_1 \dots v_N$ ($N \leq V_{max}$). Each individual v_i has associated a set of mutations: v_i, defined as a binary vector $v_i = [m_{i,1} \dots m_{i,M}] \in \{0, 1\}^M$, where 1 (0) codes the presence (absence) of a mutation. Mutations are intended as amino acidic substitutions in the viral genotype with respect to a HIV-1 wild type reference sequence (namely, subtype *consensus B*). Thus, the viral population can be defined as a binary matrix M of mutations. Through time, this matrix will change randomly according to the distribution of the real mutations observed experimentally.

The aim of this paper is not to derive an analytical solution of the corresponding differential equation, since $\eta(t)$ is a complex time-varying function. In contrast, an algorithmic simulation procedure is designed to simulate it numerically. Of note, this procedure can be applied also to a more complex systems, for example designing approximate numerical solutions for the set of equations in [11] or directly extending the [18] lattice model.

The assumption made for the algorithmic simulation is that η can be considered constant for a very short time (i.e. one day, which is reasonable, since it is known that the virus undergoes significant changes over time over a time-scale of weeks), so that an estimation of the number of virions for the following time step can be calculated using the exact solution of the logistic equation. In order to do this, an average $\hat{\eta}(t) = \hat{\eta}$, $t <= 1$, that is representative of the whole viral population resistance is needed.

The procedure is as follows:

- time t, cART $C_t = \{c_{t,1} \dots c_{t,K}\}$, viral population $V_t = v_{t,1} \dots v_{t,N}$
- for each virion $v_{t,i} \in V_t$ calculate efficacy of each single drug $c_{t,k} \in C_t$ according to the mutations $m_{t,i}$ of $v_{t,i}$: $\eta_{t,i,v,c_k} = f(v_{t,i}, c_{t,k}) = f(m_{t,i}, c_{t,k})$. Combine the single efficiencies into one overall $\eta'_{t,i,C} = \bigcup_k \eta_{t,i,c_k}$, where the union is the probabilistic sum ($\eta_1 \bigcup \eta_2 = \eta_1 + \eta_2 - \eta_1\eta_2$)
- calculate $\hat{\eta}_{t,C} = \frac{\sum_{i=1}^{N} \eta'_{t,i,C}}{N}$ as average constant and fed to the solution of logistic equation
- calculate the number of virions $N_{t+1} = |V|_{t+1}$ at time $t + 1$ and the population change $\Delta N = N_{t+1} - N_t$

- let the population V_t reproduce and mutate (according to the rates above calculated): $V_{t+1} = g(v_{t,i}, \eta'_{t,i,C}, \Delta N)$ where g updates the number of virions, and update the mutation matrix $M_{t+1} = h(M_t)$

Functions f, g, h, represent the efficacy of a single drug against a (mutant) virus isolate, the reproduction function and the stochastic mutation drift respectively.

The function $f(m_{t,i}, c_{t,k}) = \eta_{t,i,v,c_k}$ depends on the mutations of a virus isolate and on the specific drug considered: it can be easily defined using the results coming from in-vitro *phenotypic* tests. In-vitro studies are collections of experiments that measure how a mutant virus responds in a culture to a single drug inhibition: the phenotype is a numeric indicator of viral replication power, expressed as *fold change* of the drug concentration needed to inhibit 50% of the viral replication as compared to the wild type drug-susceptible reference viral strain, under the same drug pressure. Predicting single in-vitro phenotypes from viral genotypic data is a widely explored task. Multiple linear regression, decision trees and support vector machines applied to genotype-phenotype pairs are able to perform predictions that explain correctly up to 80% of phenotypic variance [90]. Analysis of phenotype predictions (among naive and treated patients) reveals in general a bimodal nature of distributions, as can be seen in figure 4.2. In [90], the probability density of predicted phenotypic Log_{10} fold change y is described using a *two-component Gaussian mixture model* for each drug:

$$\alpha\phi(y, \mu_1, \sigma_1) + (1 - \alpha)\phi(y, \mu_2, \sigma_2) \tag{4.12}$$

where $\phi(y, \mu_i, \sigma_i)$ is the density of a normal distribution i with mean μ_i and standard deviation σ_i; α is the mixing parameter. Assuming $\mu_1 < \mu_2$, the resistant population belongs to the Gaussian centred in μ_2. Parameters are estimated by expectation maximization algorithm. A log-likelihood ratio is defined to decide whether a given phenotype is more likely belonging to the resistant or susceptible sub-population:

$$l(y) = \log\frac{\Pr(\text{res}|y)}{\Pr(\text{sus}|y)} \tag{4.13}$$

From Bayes' formula it follows

$$l(y) = \log\frac{\Pr(\text{y}|\text{res})\frac{\Pr(\text{res})}{\Pr(y)}}{\Pr(\text{y}|\text{sus})\frac{\Pr(\text{sus})}{\Pr(y)}} = \log\frac{\phi(y, \mu_1, \sigma_1)}{\phi(y, \mu_2, \sigma_2)} + \log\frac{\Pr(\text{res})}{\Pr(\text{sus})} \tag{4.14}$$

By approximating $l(y)$ with its tangent $l'(y)$ in y_0, the probability of resistance $\Pr(\text{res}|y)$ is the logistic function of $l'(y)$ for a given viral genotype with respect to a drug:

$$\Pr(\text{res}|y) \approx \frac{1}{1 + e^{-l'(y)}} \tag{4.15}$$

The *efficacy* η of a treatment against a given viral strain is obviously

$$\eta(m, c) = \frac{1}{1 + e^{l'(f_c(m))}} \tag{4.16}$$

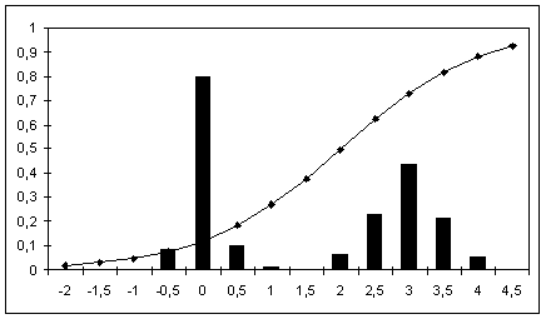

Figure 4.2: Estimation of drug resistance (1-efficacy) probability (y axis) given the distribution of phenotypic Log_{10} fold changes in a viral population sample (x axis), for a given generic drug, using two-component Gaussian mixture model.

where c is a drug compound, m is a viral isolate genotype and $y_c = f_c(x)$ is the Log_{10} fold change phenotype (prediction) for virus m and drug c, having $\eta(c, m) \approx \Pr(\text{sus}|y_c)$.

In this paper, multiple linear regression models were trained on public genotype-phenotype data sets provided by Stanford University (http://hivdb.stanford.edu), considering viral genomic sequences in the polymerase region and reverse transcriptase or protease inhibitors approved by Food And Drug Administration (www.fda.gov): training set cardinality for single drugs ranged from ≈ 200 to ≈ 1000 samples. Mutations were extracted from consensus B reference, considering all positions, even if not previously reported in literature as associated to resistance. Space (≥ 2000 variables) was reduced using sequentially a filter based on univariable rank-sum test on Log_{10} fold changes and stepwise selection based on M5 algorithm [99]. Regression results were consistent with state of the art and yielded correlation values between predicted and observed phenotypes $\rho \in (0.88, 0.97)$ in 10-fold cross validation. The two-component Gaussian mixture models parameters were estimated from a random sample (1500) of polymerase sequences (naive or drug experienced) drawn from the Los Alamos HIV data bases (www.hiv.lanl.gov/), with high concordance with previous estimates by [90].

Function $g(v_{t,i}, \eta'_{t,i}, \Delta N) = V_{t+1}$ updates the number of virions in the population, yielding $|V_{t+1}| = |V_t| + \Delta N$. In order to do this and to assure that only the resistant individuals survive or reproduce, virions $v_{t,i}$ are ordered ascending by the their $\eta'_{t,i}$: if $\Delta N > 0$, ΔN individuals are added to V_t, by drawing them sequentially from the ordered population; if $\Delta N < 0$, individuals are instead deleted sequentially but in the reverse order (i.e. the less resistant viruses are killed). Moreover, since complete eradication is not possible, a fixed number of virions is kept safe from the drug pressure at each time iteration: if the viral pop-

ulation reaches low concentration (< 500 cp/ml), the hidden virions are recalled: this accounts also for latently infected cell activation in the body. For each virion that survives or reproduces, the random mutation procedure is executed at each time step.

Function $h(M_t) = M_{t+1}$ drives the mutation process: for each genome $m_{t,i}$ of $v_{t,i}$ (an individual that reproduces), the number k of possible new mutations for each new generated individual is given by a Poisson distribution

$$\Pr(k) = \frac{\lambda^k e^{-k}}{k!} \tag{4.17}$$

where $\Pr(k)$ is the probability of having k mutations. Since the average error rate in HIV-1 is 3.4×10^{-5} per nucleotide [86] [52] and the *polymerase* (reverse transcriptase plus protease) is ≈ 1500 bases, the resulting λ value is 0.051. The mutations in polymerase do not have all the same probability to appear, thus when a substitution happens, a codon and amino acid are sampled from a distribution that is representative of the prevalence of mutations in the drug-naive HIV+ population (see figures 4.3 and 4.4). The relative frequencies of all mutations in polymerase are calculated considering the whole naive viral population stored in the Los Alamos HIV data bases (www.hiv.lanl.gov/). Clearly, when a mutation which is already in the genome happens to be added, this is removed. In absence of drug pressure, thus, the viral population tends to accumulate mutations following the naive distributions; but when a rare resistant mutation is acquired, maybe resulting from treatment selection, the probability to lose it is extremely low.

4.3 Logistic Stochastic Model: Application and Results

The logistic stochastic model was implemented using *Java* (© Sun Microsystems) programming language. The set of drugs considered was: zidovudine (AZT), lamivudine (3TC), emtricitabine (FTC), tenofovir (TDF), abacavir (ABC), efavirenz (EFV) as reverse transcriptase inhibitors; lopinavir (LPV) as protease inhibitor. The growth rate r was set to 0.44 and latent reservoir size fixed on 500 cp/ml, according to experimental and literature evidences, when the rise in viral load for AZT monotherapy is expected to show before 90 days after the treatment initiation in drug-naive patients. With the same set of parameters, as a proof, LPV monotherapy showed to be more powerful in suppressing viral load than AZT. Initial viral population was 10'000 cp/mL and initial viral mutation matrix generated according to relative frequencies estimated using drug-naive sequences stored in Los Alamos HIV data bases (www.hiv.lanl.gov/). Long-term end point was set to 2 years.

Simulations for specific drug combinations were run for 30 times, collecting the median values of Log_{10} viral load distribution through days. Suboptimal ther-

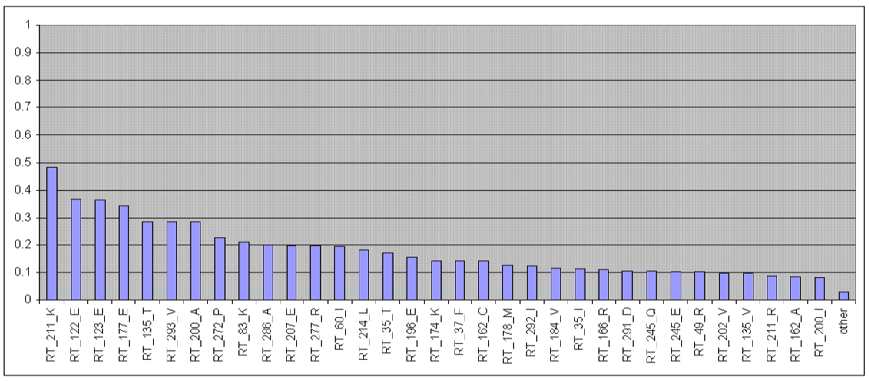

Figure 4.3: Distribution of mutations in Reverse Transcriptase among ART-naive sequences.

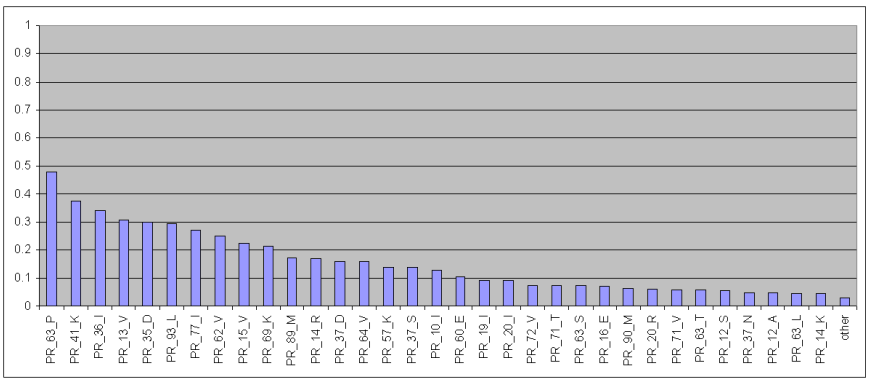

Figure 4.4: Distribution of mutations in Protease among ART-naive sequences.

apies considered were: AZT and LPV monotherapy, AZT+LPV dual therapy. Combination therapies considered were: AZT+3TC+EFV, AZT+3TC+ABC, TDF+FTC+EFV. Additional simulations were run for unadherent AZT+3TC+ABC (1 missed dose every three days), *drug sparing* policy (AZT, then AZT+3TC after 90 days when first viral rebound > 1500 cp/ml, then AZT+3TC+EFV at second same viral rebound), treatment change (whenever viral load > 1500 cp/ml) from AZT monotherapy to 3TC+LPV and then back to AZT, *drug holidays* under AZT+3TC+EFV HAART. An experimental simulation was set up considering the random administration of a single drug every day among the set of {AZT, 3TC, ABC, EFV, LPV}.

Figure 4.5 depicts viral load over time (one simulation run) for suboptimal therapies and HAART given to drug-naive patients. A first test to assess model behaviour was to execute a Wilcoxon rank-sum test on the median viral load distributions over time to compare combination therapies. It yielded that, in first line therapy, HAART significantly ($p < 0.05$) ensured lower viral loads than suboptimal therapies. This result – although already well stated in literature – was consistent with the findings in [30], in which it was shown that patients receiving ≥ 3 active drugs had a significantly lower risk of virological failure than patients receiving ≤ 2 active drugs, using survival analysis. Moreover, it was also shown that resistance scores based on phenotype prediction provided significant prediction of 24 weeks virological failure.

More interestingly, the comparison between AZT+3TC+EFV against TDF+FTC+EFV showed that the latter HAART did not lead to viral load rebound (i.e. resistant mutants did not grow in the population) during the first two years, whilst for the former the median time to the first viral rebound > 1500 cp/ml was 520 days. In fact TDF and FTC are recently approved drugs of the same type of AZT and 3TC respectively, but with higher potency. This has been also shown in [10], where a regimen TDF+FTC+EFV demonstrated superior virologic and immunologic effects compared with a regimen of fixed-dose AZT+3TC+EFV, through 96 weeks in a randomised open-label trial.

Figure 4.6 depicts viral load over time (one simulation run) for the additional set of simulations run for unadherent, random administration, drug sparing, drug recycling and drug holidays. The simulations showed that the poor adherence yielded higher risk of virological failure, with slope comparable to a suboptimal dual therapy, but with the worst result of a three-drug resistant virus. Random administration of a single drug every day did not increase the chance of success over the HAART, yielding large inter quartile ranges in viral load distribution over time. Regarding drug sparing, the scenario deserves some considerations: in [80] it was stated that the sequential addition of a drug to ongoing regimens is not a favourable policy, since selects inevitably a cross-resistant population. This is partly true, because it was not considered the time in which a drug should be

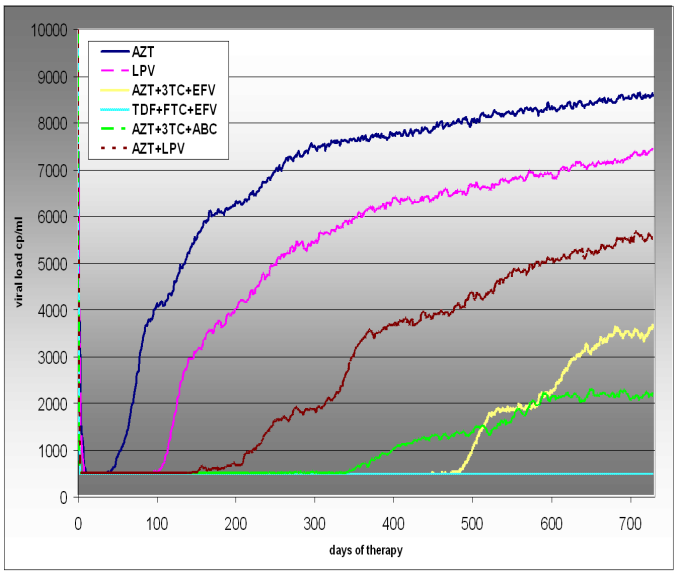

Figure 4.5: Viral load change through time for mono, dual therapies and HAART.

added. Since the probability to select resistant mutations increases when viral load is high, if a drug is added right after the viral load starts to increase, the chance of emergence of cross-resistant variants can be reduced. The drug sparing policy here analysed (first AZT, then AZT+3TC and then AZT+3TC+EFV), considered a drug addition after 90 days when first viral rebound was $>$ 1500 cp/ml and another right after the second same viral rebound. There was no statistically significant difference under the rank-sum test against AZT+3TC+EFV HAART. But in this situation the model undergoes to some limitations, because the role of latent reservoirs and immune response becomes more important. However, the drug sparing policy could have a reason only considering the fact that HAART regimens can be associated with increased toxicity and poor adherence, as pointed out in [47], otherwise it is not worth the risk to miss the optimal timing.

By its design, the model of viral mutation emergence is prone to maintain the (rare) resistant mutations acquired. Thus, in the scenario of switching (whenever viral load $>$ 1500 cp/ml) from AZT monotherapy to 3TC+LPV and then back to AZT at second failure, almost no inhibition effect from the second AZT administration could be measured, even if the combination 3TC+LPV did not share any cross-resistant mutation with AZT.

As it concerns the simulations for drug holidays, in which an HAART was periodically interrupted (in this case AZT+3TC+EFV suspended every 6 months for 3 months), no appreciable differences were found in the selection of resistant mutants with respect to the same HAART. Indeed, if the drug interruption time

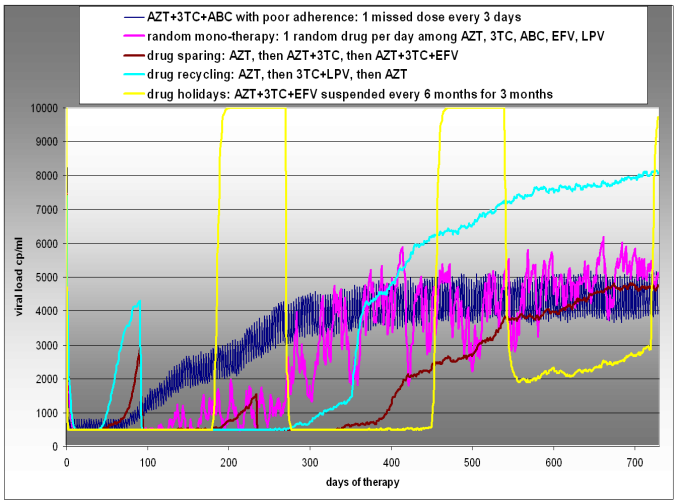

Figure 4.6: Viral load change through time for unadherent therapies, random administration and selected treatment change episodes.

is shorter and more frequent, this is equivalent to an unadherent therapy. But the difference is crucial, since with poor adherence the virus is able to reproduce under selective pressure. In therapy interruption, instead, the virus grows without selecting any particular mutation: thus, if the drug pressure is sufficiently high to avoid the selection of resistant mutants before the interruption, the overall distribution of mutations in the population follows the wild type. Structured drug holidays have the advantage to reduce the toxicity problems, but the disadvantage of a high viral concentration for several months (higher probability to infect other people or for CD4+ depletion, et cetera).

Figure 4.6 shows the evolution of the viral genome for one simulation considering AZT+3TC+ABC triple therapy. Only resistance-associated mutations are listed. The cumulative probability of selecting a particular mutation over time depends either on the prevalence in the population, or on the individual contribution to drug resistance (the coefficient in the linear regression model for phenotype), which can be calculated independently when analysing a single drug pressure. But the logistic stochastic model – by averaging over different runs – can derive distributions for combination therapies.

4.4 Conclusion

The logistic stochastic equation is a relatively simple model intended to describe the principal aspects of viral genotypic drug resistance evolution under the administration of (combination) therapies. The stochastic drift for genotypic variation follows a Poisson distribution and in absence of drug pressure respects the

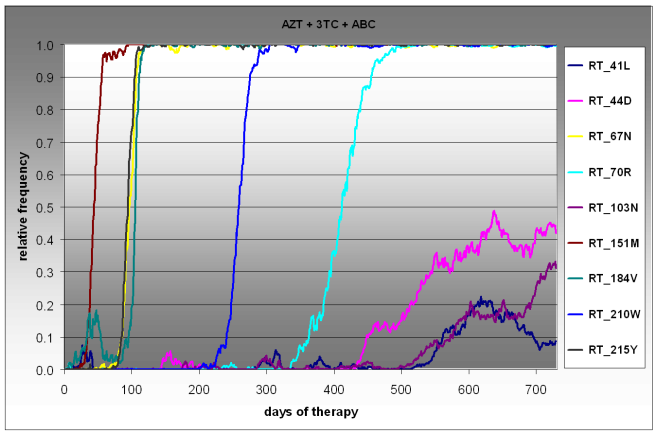

Figure 4.7: Evolution of mutations associated to resistance for reverse transcriptase inhibitors under triple therapy.

wild type mutation prevalence. The definition of a time dependent combination therapy efficacy requires a numerical solution of the model equation. An instant overall efficacy term is designed as a function of mutations in a viral genome, by means of in-vitro phenotypic resistance estimation, combining single drugs. The administration of a therapy changes the mutational drift by selecting resistant mutants that result ultimately in viral load rebound. Under these assumptions, the designed framework allows also for inclusion of more detailed viral-immune-inhibition dynamics rather than the simple logistic behaviour.

The model predicted long-term virologic outcomes and evolved HIV-1 resistant strains. Median values and confidence intervals were assessed via multiple runs. Several simulations were executed considering mono-, bi-therapies, highly active antiretroviral therapies, poor adherence, drug recycling and drug holidays scenarios. Drug sparing policies were also exploited. For a set of commonly used combination therapies, results were shown to be consistent with statistical findings previously reported in literature (corroborated by the usage of rank sum test). The model proved to be an useful tool for devising new treatment strategies. In this sense, multiple simulations can assess the effectiveness (and compare) different HAART, while the exploration and optimisation of sequential treatment change episodes has the aim to ensure the largest and longest-lasting viral load reduction, either when in presence of reduced therapeutic options or toxicity/adherence problems.

Future perspectives foresee the usage of a more complex model – rather than the logistic one – describing virus-host interactions, accounting for CD4+ T cell, cytotoxic T lymphocytes response, latent reservoirs and CTL escaping mutations. The most suitable candidate is the framework provided by [18] [17], that works on discrete time steps and already investigates the interactions with the immune

system. Another possible improvement, finally, is to include in the model the drug concentrations or half lives.

Chapter 5

Statistical and Unsupervised Analyses

The aim of this chapter is to study HIV-1 evolutionary patterns associated with exposure to specific combinations of drugs. The scenario is not dynamic, i.e. only *equilibrium* situation types will be explored, like, for example, the wild type strains or the heavily-treated strains equilibrium. With equilibrium, we refer to the steady state that the virus reaches in the host when the environmental conditions are constant: this means also consant drug pressure and fully acquired resistance, or no drug pressure and wild type condition.

Univariable and multivariable analyses are the ideal methodological tool when the objective is to identify patterns of drug-resistant mutations either from in-vivo data or using information collected from patients' routine clinical practice. Conversely, a more in depth investigation for time-dependant mutational pathways evolution will be presented in section 7 with Mutagenetic Trees and Markov chain models.

5.1 Methods

In this section, univariable (stratified) statistical tests (χ^2) will be performed in order to find out which mutations are significantly associated with resistance/susceptibility to which drug or drug combination, whilst Unsupervised Learning (UL) will be carried out in order to identify patterns of mutations: for this purpose both Hierarchical and Partitional Clustering (HC, PC) will be executed on selected viral population samples. Dimension reduction techniques will be also explored, like Principal Component Analysis (PCA) and Multi-Dimensional Scaling (MDS). These techniques are grouped in the definition of

UL because the data analysis do not require an objective or output function, but explores independently the variable space partition. A general introduction to UL can be found in [66] and [134]. More specifically, Cluster Analysis has a variety of goals. It relates to grouping or segmenting a collection of objects (i.e. observations, individuals, cases, or data rows) into subsets or *clusters*, such that those within each cluster are more closely related to one another than objects assigned to different clusters. Central to all of the goals of cluster analysis is the notion of degree of similarity (or dissimilarity) between the individual objects being clustered.

In *Hierarchical Clustering* the data are not partitioned into a particular cluster in a single step. Instead, a series of partitions takes place, which may run from a single cluster containing all objects to n clusters, each containing a single object. Hierarchical Clustering is subdivided into *agglomerative* methods, which proceed by series of fusions of the n objects into groups, and divisive methods, which separate n objects successively into finer subgroups. Agglomerative techniques are the most commonly used and this is the method that will be adopted in this investigation, even if *divisive* techniques have been also designed. Hierarchical clustering may be represented by a two dimensional diagram known as dendrogram which illustrates the fusions or divisions made at each successive stage of the analysis.

Regarding *Partitional Clustering*, the method *k-means* [69] is one of the simplest unsupervised learning algorithms to solve a clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of fixed *a priori* clusters (i.e. k clusters). The main idea is to initially define k_{t_0} centroids, one for each cluster. These centroids should be placed in a cunning way because the final clustering result can depend on the chosen location of the initial centroids. So, the best choice is to place them as much as possible far away from each other or set up a comparison of different clustering runs with initial random centroid assignments. The next step is to take each observation under study and to associate it to the nearest centroid. When all observations are allocated to a centroid, the first step is completed and A first grouping is done. At this point, k_{t+1} new centroids need to be recalculated so that they are barycentres of the clusters resulting from the previous step. After creating these k_{t+1} new centroids, a new binding needs to be done between the same data set observations and the nearest new centroid. A loop is then generated so that these two steps are repeated a certain number of times until the centroids are stabe (i.e. no longer change their location). Of note, the clustering around a centroid is regulated by the k-mean algorithm that aims at minimizing an objective function (a squared *Euclidean distance* error function) with an iterative procedure. Although it can be proved that the procedure does always terminate, the k-means algorithm does not necessarily find the most optimal configuration,

corresponding to the global objective function minimum. The algorithm is also significantly sensitive to the initial randomly selected cluster centroids. That's why the k-means algorithm can be run multiple times to try to reduce this effect. Moreover, k-means clustering is appropriate only when the Euclidean distance is a suitable metric and the procedure, since uses an average evaluator, is not robust against the presence of outliers (i.e. observations that produce very large distances).

Instead of the naive k-means method, in this thesis, we will use an improved version of this algorithm, named *k-medoids* or *Partition Around Medoids* (PAM), which uses median values instead of mean (and therefore it is more robust to outlier). This approach has also the advantage of being able to handle a distance matrix that can be calculated with any distance rather than just the Euclidean one.

In addition, a model-based clustering approach, using *Gaussian Mixture Models* (GMM) and *Expectation Maximisation* (EM) algorithm, will be executed. It has been shown that k-means is a hard version of the GMM-EM, in the sense that it does not use continuous cluster assignments as the GMM-EM.

Regarding dimensionality reduction techniques, linear combinations of variables to explain the data variance will be explored by using both PCA and MDS, a method to graphically represent multi-dimensional data in lower dimensions.

The list below summarises all the methods used in this chapter with the appropriate references:

- Statistics: non parametric univariable (stratified) χ^2 test on cross tabulations with adjustment for multiple testing (Benjamini-Hochberg and cross-validation) [105]

- Clustering:

  - Hierarchical Clustering HC: agglomerative methods with average grouping, binary distances [68] [97] and multiscale bootstrap resampling [61]

  - Partitional Clustering PC: Partition Around Medoids PAM [66] with silhouette analysis for optimal cluster number assessment; Gaussian Mixture Model Clustering GMM with Expectation-Maximisation EM algorithm for optimal cluster number assessment [50]

- Dimensionality reduction: Principal Component Analysis PCA [15]; Multi-Dimensional Scaling MDS (with Kruskal's Non-metric method on binary distance matrices) [25] [108]

In this chapter we assume that the mathematical domain for viral genotypes is defined as a n-dimensional binary vector which codes the presence or absence of a particular mutation. However, note that the procedure to extract mutations

from a virus isolate against a consensus reference is not a trivial one and becomes more complex when analysing different genes. We posticipate the description of this procedure to section 9.2.1, in which the problem is faced for one of the most problematic genes, the *envelope*. But the procedure has been showed to work even more robustly for *protease* and *reverse transcriptase*.

5.2 State of the Art

Clinical evidences and statistical analyses of resistance data have shown over the past few years that HIV-1 seems to have some preferred evolutionary pathways varying with the class of inhibitors to which is exposed [121]. For Nucleoside and Nucleotide (or Thymidine Analogues) Reverse Transcriptase Inhibitors (NRTI) two patterns frequently emerge: {M41L, L210W, T215Y} (TAM1) and {K70R, D67N, K219QE, T215F} (TAM2)*. Two other rare patterns are also observed: the T69-INSERTION complex and the Q151M complex (that are both often detected along with some TAMs, plus {A62V, V75I, F77L, F116Y}). Finally, additional mutations are significantly associated with the use of nucleosides; some are exclusively selected by a particular drug (i.e. M184VI for Lamivudine, that gives hyper-susceptibility to Zidovudine OR L74V that seems more specific for Abacavir): {E44D, K65R, L74V, Y115F, V118I, M184VI}.

A different set of mutations in the reverse transcriptase region is associated with reduced susceptibility to Non-Nucleoside Inhibitors (NNRTI): this is due to the fact that nucleosides and NNRTI, at molecular level, act in a different way. Also a list of mutations associated with reduced susceptibility to Protease Inhibitors (PI) has been defined for each drug in this class. In general, some mutations shows cross-resistance to other drugs of the same class or other classes, others instead show the opposite behaviour (i.e. if detected they are likely to increase HIV-1 susceptibility to other drugs). The number of mutations that seem to be directly selected by antiretroviarl treatment is constantly updated by panels of experts [121] although a consensus is often not reached by all experts. One possible way to identify mutations that are genounely associated with drug exposure is to perform univariable tests comparing the observed frequency of mutations in populations of treated or ART-naive patients. However, this approach has limitations as it is not able to reveal the complex mechanisms of dependencies between mutations (and drugs) and does not take into account the viral evolution through time.

Multivariable analyses and Unsupervised Learning are a further step, useful to find correlations, clusters, antagonisms, that should be checked against our virological hypotheses.

*please note the different amino acidic substitutions at position 215

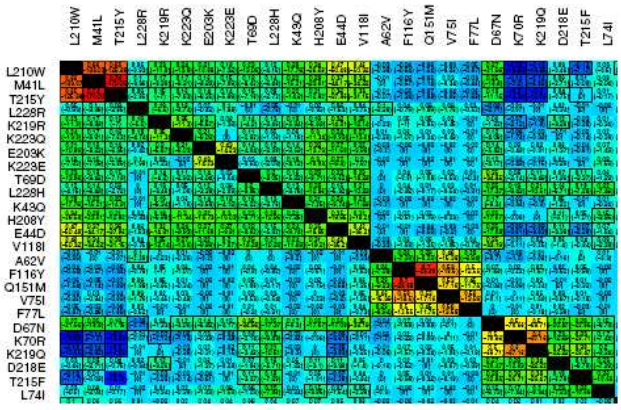

Figure 5.1: Pairwise ϕ correlation coefficients between mutations (part view), with red indicating maximal observed positive covariation and blue maximal observed negative covariation. Boxes indicate pairs whose covariation behaviour deviates significantly from the independence assumption, according to Fishers exact test and correction for multiple testing using the Benjamini-Hochberg method at a false discovery rate of 0.01. The classical mutational complexes form distinct clusters, from left to right: TAM1, Q151M-complex, TAM2.

One of the most complete studies to date is the work by Sing [116], investigating single mutations and patterns associated with drug resistance/susceptibility: it was used univariable analysis, hierarchical clustering and multidimensional scaling, with the aim to confirm/reject the expert panel list (IAS/USA) [121] and to find novel mutations.

Figure 5.1 (from [116]) shows a distance matrix among RT mutations obtained from NRTI treated genotypes using normalised ϕ-coefficient measure: ϕ-coefficient is a measure of correlation for binary data), where $\phi = 1$ means positive covariation, i.e. the probability of detecting two mutations in the same genotype is very high. The TAM1, TAM2 and Q151-complex patterns are clearly evidenced, as can also be seen in the multidimensional scaling picture showed in figure 5.2, where novel (not present in IAS/USA official list [121]) mutations are also included.

Although this type of analysis has improved our understanding of mutational patterns, there is still need to analyse mutational covariation in more depth. Indeed, for NRTI, apart from the well established TAM1 and TAM2 patterns, along with THE Q151M complex, the scenario for the identification of novel relevant mutations and relative pathways has so far led to less clear results. For example, the role of M184V is strongly treatment dependent, suggesting that it may be implicated in other mechanisms beside those already established (the hyper-susceptibility given to AZT and TDF and the resistance to 3TC and FTC).

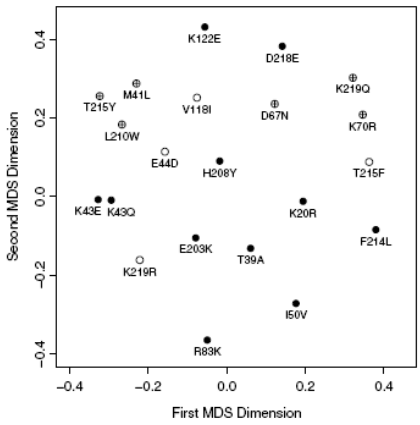

Figure 5.2: Multidimensional scaling plot of novel (shown in black) and classical mutations (in white; main TAMs indicated by a cross), showing a two-dimensional embedding which optimally (according to Sammon's stress function [25] [108]) preserves the distances among the mutations, as derived from the ϕ correlation coefficient. Distances between mutations at a single position were treated as missing values.

The pattern of accumulation of mutation D67N is also unclear as this mutation is usually found in conjunction with the TAM2 pattern, but also often present in TAM1 populations [27]. V118I seems to act in the same way, as no significant difference in prevalence was found when comparing TAM1 and TAM2 profiles, at least when the analysis was restricted to the whole NRTI-treated population, without accounting for treatment covariation.

Therefore, not only specific drugs, but also treatment combinations may be expected to play a crucial role in selecting mutational pathways: for this reason, we decided to carry out a set of stratified analyses. Of note, it would have been also possible to perform a logistic or linear multivariable analysis controlling all factor simultaneously, but with the limitation of analysing only linear combinations.

Finally, the individuation of patterns for Protease mutations gave less encouraging results: the mutations associated to resistance did not seem to cluster in well characterised groups as for the RT.

5.2.1 The Impact of Naive Polymorphisms against the Resistance Development

A more interesting issue is the analysis of the associations between the detection of specific mutations or polymorphism at the time of treatment initiation and the probability of virological success (or risk of failure) on that therapy. Similarly, this type of analyses (for example a survival analysis) can be used to test whether

some polymorphisms detected in a population starting ART from drug-naive is associated with a reduced risk of detecting resistance mutations at virological failure. We carried out this study for the specific case of mutation R83K, a polymorphism in the RT region that was shown to be more prevalent in ART-naive patients as compared to ART-exposed patients in a previous analysis [116]. The analysis was conducted using longitudinal data from patients enrolled in the *Spallanzani, Catholic University of Sacro Cuore* clinical data bases and from the *ICONA* Italian cohort for naive to antiretrovirals. We showed that mutation R83K was negatively associated with TAMs and was associated with a lower risk of detection of TAMs at virological failure of first line therapies. Results of this analysis were presented in [43].

5.3 Results

An univariable analysis was performed to test whether there was an association between the detection of specific mutations and the probability of being treated/drug-naive (grouping together viral sequences treated within the same drug class) and of experiencing success/failure on treatment. In this latter analysis, success was defined as the achievement of an undetectable viral load (< 500 cp/ml according to the less sensitive assay) measured after 8 weeks of treatment. These two analyses had the objective of identifying which mutations were selected by which drug and which mutations where likely to influence the virological outcomes, respectively. This univariable analysis has a limitation: indeed the probability of virological success is higher in patients starting from drug-naive than in those who had previous exposure to treatment and, therefore, the association between mutations that are highly prevalent in the ART-naive population, along with the chance of virological success could simply be due to confounding. For the χ^2 analysis, all results were adjusted either by 10-fold cross-validation or Benjamini-Hochberg method. Only mutations detected with prevalence $\geq 3\%$ were used in this analysis.

The more specific analysis on the potential *protective* role of polymorphisms detected in drug-naive patients against the accumulation of resistance failure, however, was not completely executed because the information in the available data bases was not complete for all the variables in all patients. But we stress again that, when carrying out this type of analysis, it is crucial to adjust for all possible covariates (such as the different RTI backbone, the number of therapy lines previously experienced, etc.) that can mask the associations of interest.

For sake of simplicity, we will not report here the results of the analyses of mutations in the protease gene associated with protease inhibitors: these analyses have been carried out in the same settings, either with respect to (stratified) univariable analysis or cluster analysis.

When analysing data coming from clinical records with the obective of trying to identify novel resistance-associated mutations, the results of the analysis may be biased by the data collection policy. On the other hand, evaluating the existence of significant association between mutations and phenotypic resistance using in-vitro data (i.e. phenotypic tests) is often considered an "easy job": in this setting the statistical results are less likely to be biased by other factors and phenotypic measures are extremely reliable; however it is possible that associations that are very strong in-vivo are not detected by correlation analyses using viral load (when used as a proxi of an in-vivo phenotype). Testing for correlations using in-vivo data and laboratory markers as the independent variables (e.g. CD4 count and viral load), is a difficult task because analyses must be stratified or controlled for a large number of factors (and this typically leads to loss of statistical power). In particular, the potential confounding effect of the specific combination therapy used is not negligible and marked noise effects are expected because of patients metabolic variability.

5.3.1 Data Collection and Descriptive Statistics

A first general analysis was carried out using viral sequences coming from the EuResist data base [98]; patients were divided according to whether – at the time of the test – were still ART-naive or pre-treated with NRTI or PI: a sequence was classified as belonging to the treated group if the patient experienced at least one NRTI or one PI for ≥ 12 months, whilst it was grouped in the ART-naive category if – at the time when the plasma sample was collected – the corresponding patient had never previously received any drug. All sequences were included in the analysis, regardless of patients' subtype. As a result, because the prevalence of mutations varies according to HIV subtype, the estimate of the overall prevalence or resistance may be biased but the selection of mutations under the pressure of a certain drug should be less affected by this potential bias, as long as the analysis is restricted to equilibrium-type scenarios. In other words, it may be possible that *a priori* the development of escaping mutations is dependant on viral subtype, but after a reasonably long period of time it is likely that the natural selection process changes the mutational prevalence towards that of a consensus resistant population. Although we cannot rule out the fact that the selection of resistant patterns may also vary by subtype, by the definition itself, a subtype is an evolved strain from an ancestor.

Figure 5.3 shows the subtype distribution in the EuResist database: as expected, subtype B is the most prevalent (85%), while only complex recombinant form 02_AG and subtypes C, F1 and A1 are present in a prevalence $\geq 2\%$.

The sequences were also grouped with respect to whether a viral load < 500 cp/ml was achieved after 8 weeks of treatment (defined as virological success) and the analysis was stratified by specific treatments. The total number of *pol*

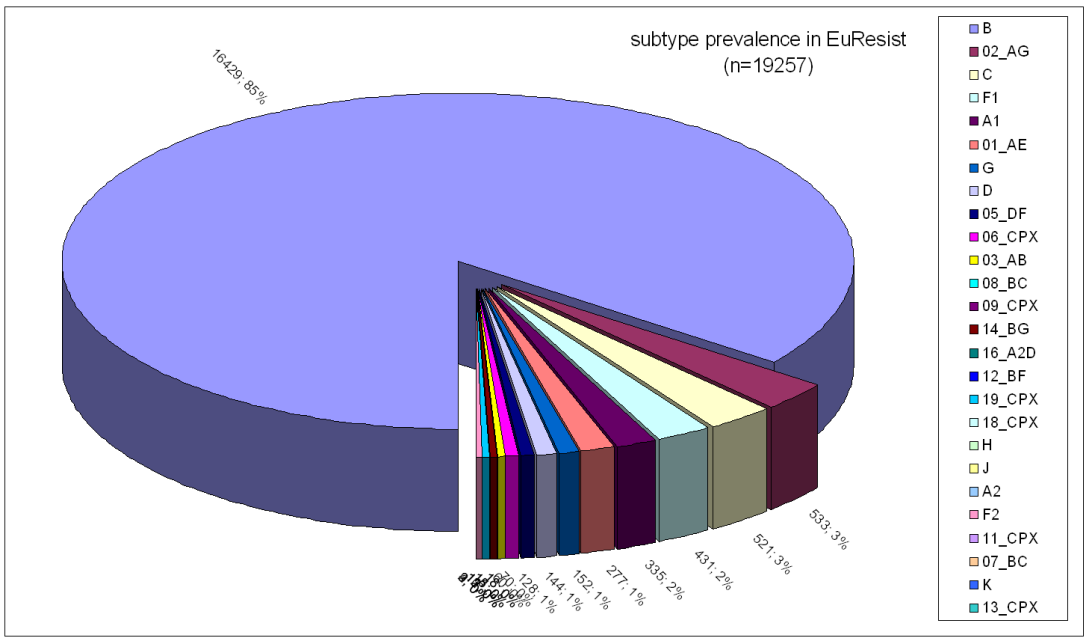

Figure 5.3: Subtype prevalence in EuResist.

sequences included in the EuResist DB was 19257. The study population consisted of 6203 sequences in pretreated patients and 2569 in drug-naive patients for a total of 8772 sequences. Out of 2523 sequences for which a viral load after 8 weeks was available, for 1692 the viral load was < 500 cp/ml. We then calculated the risk of virological failure stratified by specific drugs contained (without controlling for the remaining drugs) in the regimen: these risks were 21% (196 failures out of 799) for AZT, 23% (320 failures out of 1413) for 3TC, 39% (295 failures out of 756) for DDI and 30% (249 failures out of 818) for TDF.

The prevalence of the different therapies used in the whole EuResist DB is depicted in figure 5.4: as expected, there was a large variability in the combination used, where AZT+3TC, D4T+3TC were frequently used nucleoside backbones. The presence of suboptimal therapies (i.e. combinations including less than three drugs) reflects the fact that data were collected in a wide time window, starting before 1996. The distribution of sequences by calendar year or type of treatment is not given here (but that of the specific drugs used is shown in the next paragraph) because not strictly necessary for the purpose of this thesis, although these results should be reported when trying to publish in a peer reviewed journal.

A stratified analysis was then carried out after further dividing the study ppulations in specific subsets: sequences were analysed after stratifying by com-

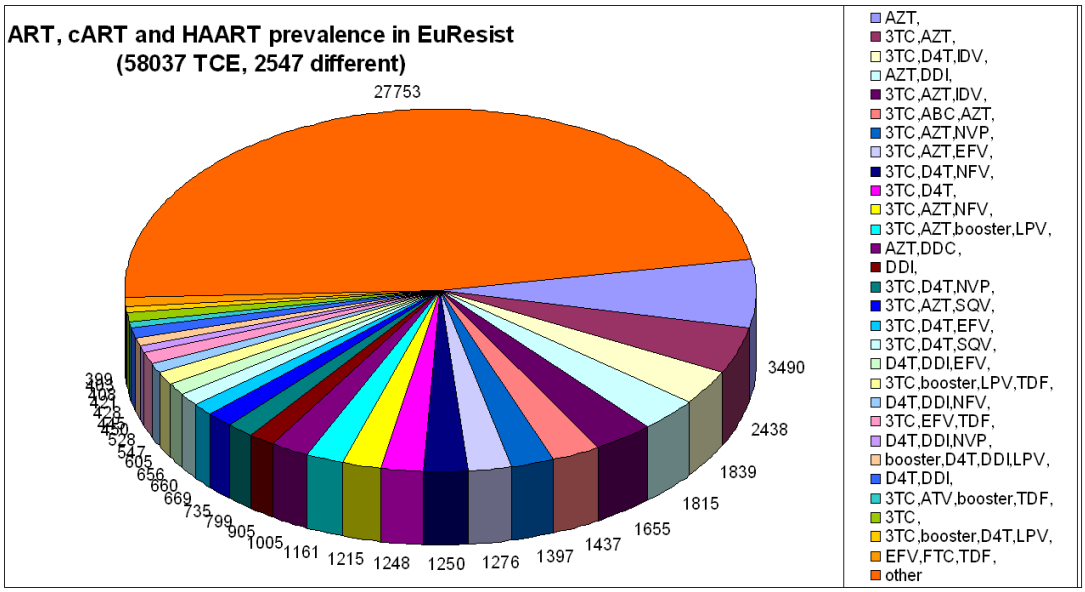

Figure 5.4: Prevalence of (combined) Anti Retroviral Therapies in EuResist.

bination treatments, i.e. for nucleosides backbones: AZT+3TC (n=494, 385 successes), DDI+D4T (n=234, 120 successes) AND 3TC+TDF (n=229, 155 successes). Patient could have concomitantly received a PI or a NNRTI, but they had to be on exactly two nucleosides backbone.

5.3.2 Univariable Analysis

Figure 5.5 shows the frequencies of specific mutations in drug-naive and NRTI_treated patients in Euresist and the results of the χ^2 analysis. According to the prevalence of changes we constructed bunary variables indicating the presence or either a single amino-acidic substitutions or "any" substitution at a codon position. The mutations or codon showed in the figure were all those that were found to be significantly associated with patients' drug exposure with $p \leq 0.01$. After correcting for multiple comparison, $adj.p \leq 0.1$.

Most of these mutations were those included in the ias-usa list [121] list. In contrast, the analysis here is based purely on the difference in prevalence between art-naive and pretreated patients and therefore shows both mutations with particulary high ad those with particularly low prevalence in pretreated as compared to drug-naive. the identification of mutations that are less frequent in

pretreated patients is useful as it has been hypothesised that the accumulation of these mutations carries a cost for the virus and its detection could potentially represents an obstacle for further development of resistance. this was shown for mutation RT_R83K, which seems to be associated with a delayed accumulation of tams [116] in patients receiving zidovudine. Although, these results could not be repeated in the EuResist data set.

The second analysis that investigates the relationship between detection of specific mutations and the risk of virological success/failure (stratified by treatments), is summarised in tables 5.2 and 5.1: since the sample size was not large, we used a cross-validated chi-square test. Again, the majority of mutations identified to be significantly associated with the virological response were also included in the IAS list [121], with a few exceptions: mutations 44D and 69D in the RT region have been recently excluded from the IAS list because they are not thought to have a major impact in reducing susceptibility to nucleosides inhibitors, but they were significantly associated with viral load in almost all subsets of this analysis; similarly both position 43 and mutation 208Y were associated with failure. In a number of subsets, the R83K as well as other polymorphisms were found to be associated with an increased probability of success, although only in the subset of patients receiving DDI. However, in order to establish whether there is something genuinely going with DDI we should formally test for interactions.

The results of this analysis need to be interpreted with caution as they are not adjusted for the exact nucleoside backbone or patients' previous drug history. Unexpectedly, mutation M184V was found to be associated with virological failure to AZT-containing regimens, while there is some evidence suggesting that this substitution produces hyper-susceptibility to this drug. The confounding here is likely to be the fact that AZT is often used in combination with 3TC, and it is well known that M184V confers high resistance to 3TC. In addition, a few mutations selected by NNRTI such as K103N, showed a significant association with risk of virological failure in this analysis. This is also expected as there was a relatively high prevalence of regimens containing NVP or EFV.

It's worth to point out also that for some mutations (like 116Y), although the frequency is higher in virological failure than success, the prevalence in the latter group is nearly zero, so the chances to detect these mutations in pre-treated patients is still very small and probably clinically irrelevant.

Table 5.3 refers to the analysis stratified by NRTI backbones. The number of significant associations decreases dramatically compared to the previous analysis: this was possibly expected, at least because we partially avoid the confounding effect of co-administrated drugs, though still the treatment history could have an effect. Most of polymorphisms typical of drug-naive patients that were found to be significantly associated with virological response in the previous analysis were no longer associated here. This may suggest that the few that remained associ-

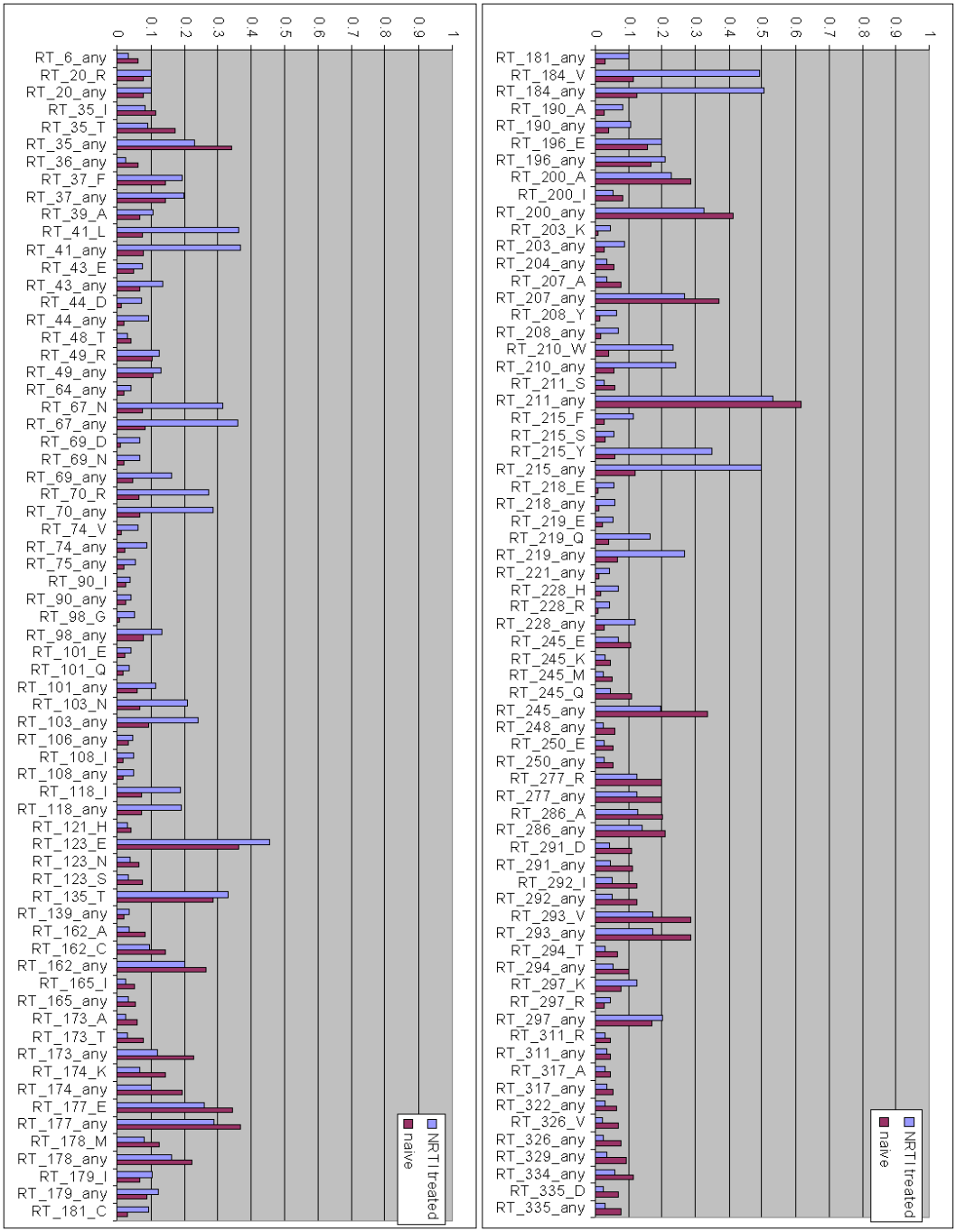

Figure 5.5: Prevalence of mutations in Reverse Transcriptase among NRTI-treated and naive population. Only significant variables showed ($adj.p < 0.1$) among univariable analysis

TDF (249 fail on 818)

χ^2	±	mutation	failure	success
49.569	3.802	210W	0.281	0.091
45.418	4.034	210any	0.317	0.112
43.482	4.224	215any	0.566	0.309
36.664	2.498	118I	0.241	0.083
36.664	2.498	118any	0.261	0.091
31.976	3.539	215Y	0.347	0.166
28.248	3.120	208any	0.120	0.025
27.173	2.132	41L	0.359	0.186
25.853	2.026	41any	0.378	0.200
22.946	2.971	208Y	0.090	0.018
20.533	2.374	69D	0.117	0.037
20.276	4.026	44any	0.124	0.037
17.230	1.729	43any	0.181	0.077
15.928	1.991	43E	0.114	0.040
14.236	2.385	69any	0.237	0.127
14.208	3.575	44D	0.090	0.028
12.924	2.350	67any	0.378	0.248
12.595	3.008	103any	0.361	0.236
12.368	2.777	103N	0.287	0.184
11.158	4.321	123N	0.080	0.030
9.738	5.185	67N	0.307	0.199
8.335	3.081	74any	0.165	0.090

3TC (320 fail on 1413)

χ^2	±	mutation	failure	success
87.729	3.883	215any	0.533	0.261
62.561	3.980	210W	0.293	0.115
59.788	3.441	118any	0.269	0.100
59.724	2.957	118I	0.237	0.088
56.438	4.042	215Y	0.346	0.160
53.009	4.639	210any	0.321	0.144
49.901	4.833	208Y	0.102	0.017
48.276	3.290	41L	0.370	0.190
46.626	3.223	208any	0.119	0.025
44.516	3.288	41any	0.383	0.204
41.490	4.400	67any	0.333	0.170
34.573	3.167	184V	0.336	0.197
34.454	2.955	184any	0.371	0.213
33.536	4.273	67N	0.275	0.141
28.923	3.172	219any	0.262	0.137
28.098	2.535	44any	0.121	0.041
24.739	3.473	116Y	0.037	0.003
24.146	2.657	77L	0.037	0.004
20.155	2.791	69any	0.176	0.089
20.122	3.273	69D	0.078	0.025
18.228	2.296	44D	0.085	0.033
17.442	2.246	203any	0.105	0.043
17.112	2.858	103any	0.274	0.172
17.015	1.611	215F	0.106	0.046
16.314	2.225	151M	0.040	0.009
16.258	2.126	43any	0.169	0.091
15.357	2.060	190any	0.138	0.069
14.894	3.068	98any	0.169	0.094
14.852	2.339	190A	0.092	0.046
14.574	1.772	70R	0.188	0.115
13.613	2.617	181any	0.152	0.084
13.207	2.174	181C	0.130	0.068
11.782	4.445	219E	0.063	0.024
11.709	4.343	39A	0.121	0.063
11.659	4.380	122E	0.436	0.347
11.657	4.292	103N	0.201	0.133
11.642	1.984	228any	0.131	0.072
10.485	1.671	74any	0.114	0.061
10.199	1.251	70any	0.210	0.138
9.521	3.544	188L	0.036	0.010
9.164	4.940	101any	0.112	0.060
8.166	5.785	62V	0.037	0.012

Table 5.1: Stratified analysis on success/failure EuResist data set, $p < 0.01$, $adj.p < 0.1$. The columns for failure/success display the relative frequency of mutations in the subsets.

AZT (196 fail on 799)

χ^2	±	mutation	failure	success
42.762	6.052	184V	0.418	0.233
38.859	4.382	215any	0.337	0.129
30.772	2.742	70R	0.166	0.050
30.545	3.359	184any	0.474	0.254
28.179	4.129	67any	0.204	0.066
26.136	4.593	67N	0.168	0.055
24.498	2.026	70any	0.199	0.070
23.719	3.511	41L	0.207	0.078
22.653	3.222	41any	0.224	0.090
21.180	3.118	215Y	0.158	0.068
20.890	2.633	219any	0.168	0.058
20.686	2.999	210W	0.134	0.045
17.536	2.625	118any	0.168	0.065
17.420	2.919	219Q	0.098	0.027
16.856	2.662	210any	0.163	0.063
16.095	2.395	118I	0.145	0.059
12.398	6.853	215F	0.068	0.015
11.400	2.895	69any	0.143	0.063
9.197	4.792	69D	0.047	0.010
8.696	3.176	245any	0.240	0.363
8.338	1.416	77L	0.015	0.001
8.338	1.416	116Y	0.015	0.000
7.026	4.716	44any	0.061	0.018
6.894	1.317	334any	0.046	0.113

DDI (295 fail on 756)

χ^2	±	mutation	failure	success
30.415	2.745	215any	0.637	0.421
18.968	2.380	118I	0.190	0.086
18.968	2.380	118any	0.224	0.102
18.323	1.766	41L	0.456	0.302
14.933	2.324	210W	0.286	0.171
14.902	2.470	215Y	0.384	0.276
14.407	1.387	41any	0.468	0.323
14.161	2.315	43any	0.214	0.108
13.927	2.309	69D	0.090	0.030
13.480	2.343	210any	0.329	0.204
11.177	4.234	43E	0.099	0.037
10.979	4.185	37F	0.108	0.205
10.699	1.672	67any	0.400	0.280
10.654	4.120	37any	0.112	0.210
9.886	2.417	208any	0.092	0.035
9.881	1.484	122E	0.446	0.342
7.901	4.083	83K	0.141	0.225
7.497	4.076	69any	0.183	0.104

Table 5.2: Stratified analysis on success/failure EuResist data set, $p < 0.01$, $adj.p < 0.1$. The columns for failure/success display the relative frequency of mutations in the subsets.

ated (i.e. mutations R83K and codon 245) really affect the chance of achieving virological success. Of note, mutation M184V is associated with lower chance of virological success in patients receiving the AZT+3TC combination, although phenotypic data show that this mutation confers hyper-susceptibility to AZT. However, it is conceivable that this effect is overcompensated by the reduced susceptibility to 3TC.

No additional comments to the role of specific mutations will be given here, leaving to the physicians/biologists the hard job of validating these findings by comparing them with the current state of knowledge acquired from clinical and laboratory practice. However, because our study population have a large sample size and the analyses have been appropriately adjusted (even if still the hystorical bias was not eliminated), we are confident that this preliminary investigation can be used at least as a starting point for a more detailed future analysis. These results are indeed important for the issue of deriving predictive models for treatment optimisation as they give some indication about which mutation are the most likely determinants of virological response, reducing the large space state of mutations in the *pol* gene.

mutation	χ^2	failure	success
AZT+3TC (n=494, successes=385)			
184V	26.928 ± 3.252	0.275229358	0.093073592
184any	25.027 ± 3.476	0.302752294	0.101298701
70R	23.443 ± 3.54	0.142201835	0.030519481
215any	20.816 ± 3.908	0.302752294	0.114285714
70any	19.308 ± 3.452	0.174311927	0.044155844
67any	10.45 ± 3.799	0.146788991	0.049350649
77L	9.595 ± 1.629	0.027522936	0
116Y	9.595 ± 1.629	0.027522936	0
118any	9.089 ± 4.854	0.183486239	0.075324675
334any	8.083 ± 1.416	0.018348624	0.111688312
215Y	10.362 ± 5.935	0.155963303	0.062337662
245any	6.911 ± 2.643	0.229357798	0.376623377
62V	6.987 ± 4.581	0.032110092	0.002597403
211any	7.863 ± 2.99	0.495412844	0.654545455
D4T+DDI (n=234, successes=120)			
215Y	10.432 ± 4.026	0.456140351	0.279166667
69D	5.656 ± 3.779	0.085526316	0.016666667
83K	7.834 ± 4.171	0.144736842	0.3
118I	11.617 ± 1.661	0.184210526	0.05
118any	11.617 ± 1.661	0.219298246	0.058333333
122E	7.024 ± 4.681	0.451754386	0.2875
215any	10.976 ± 2.313	0.675438596	0.45
3TC+TDF (n=229, successes=155)			
210W	13.353 ± 1.446	0.351351351	0.138709677
215Y	10.804 ± 2.227	0.415540541	0.193548387
103any	8.933 ± 1.969	0.418918919	0.219354839
118I	16.578 ± 1.858	0.304054054	0.096774194
118any	16.578 ± 1.858	0.324324324	0.096774194
122E	7.461 ± 1.198	0.506756757	0.296774194
122any	6.915 ± 1.253	0.567567568	0.374193548
208Y	9.468 ± 1.885	0.114864865	0.019354839
208any	7.909 ± 1.884	0.135135135	0.032258065
210any	12.883 ± 1.595	0.378378378	0.15483871
215any	7.422 ± 1.48	0.554054054	0.35483871

Table 5.3: Univariable analysis on exclusive NRTI backbones from EuResist DB, $p < 0.01$, $adj.p < 0.1$. The columns for failure/success display the relative frequency of mutations in the subsets.

5.3.3 Clustering

A first step to identify mutational patterns in the viral population is to perform an unsupervised analysis using the data of the entire population, without any pre-selection or grouping. This analysis shows how the mutations cluster together, regardless of previous drug experience. It is a starting point that it is usually useful to increase our understanding of the mutational patterns that can arise in treated populations: in fact, several wild-type clusters corresponding to polymorphism aggregations are typically found, along with the clusters formed by the "evolved" resistant sub-population, which can be due to transmitted resistence or pre-existing minority variants.

Figure 5.6 shows the results of Hierarchical Clustering using all the genotypic sequences stored in the EuResist data base (n=19257). The distance measure chosen was the Jaccard coefficient (defined in section 10.3.6 with equation 10.11), with average aggregation method: results were assessed by bootstrap re-sampling (100 replicates). Red boxes enclose branches with $p \leq 0.05$. The usage of Jaccard coefficient is a preferable choice rather than the Hamming distance (Euclidean is not suitable for binary data) since it enhances the common changes.

By focusing initially just on the wild type groups, two main clusters can be identified at the top of dendrogram roughly including the mutation clusters {272P, 293V, 277R, 286A, 281R, 297K} and {291D, 292I, 245Q}. In addition to well characterised clusters that are found in treated populations (also see the tables of the previous section) these mutations, that are detected prevalently in the wild type populations and are considered polymorphisms, showed a significant tendency to cluster. The χ^2 test provided evidence for a higher prevalence of them in drug-naive as opposed to treated populations. Some of these mutations are also less likely to be detected if other resistant mutations are present. the question of whether they actually concur to the risk of long-term development of resistance remains open. To a certain extent, this is also true for mutaion RT_R83K as discussed in previous sections, Although there is more evidence that 83K may have an active role at selecting the pattern of HIV evolution as a results of virological failure. When looking at the clusters observed in the treated populations (bottom of figure), a number of interesting issues deserve to be discussed. First of all, TAM1 {41L, 210W, 215Y} and TAM2 {67N, 70R, 219Q, 215F} patterns were again clearly characterised using this method. Mutations 118I and 135T seem to correlate with TAM1, while {69N, 214L, 218E} are closer to the TAM2 cluster. The significant association between polymorphism 214L and TAM2 mutations has been reported in other studies [20]. Mutations 184V that is usually equally frequent with or without TAM, appeared to cluster with 123E and 211K. Other interesting mutational profiles contained pairs of mutations: {208Y, 44D}, {69D, 228H}, {101E, 190H} and {83K, 35I}.

An interesting feature of Partitional Clustering (PC) is that this method is

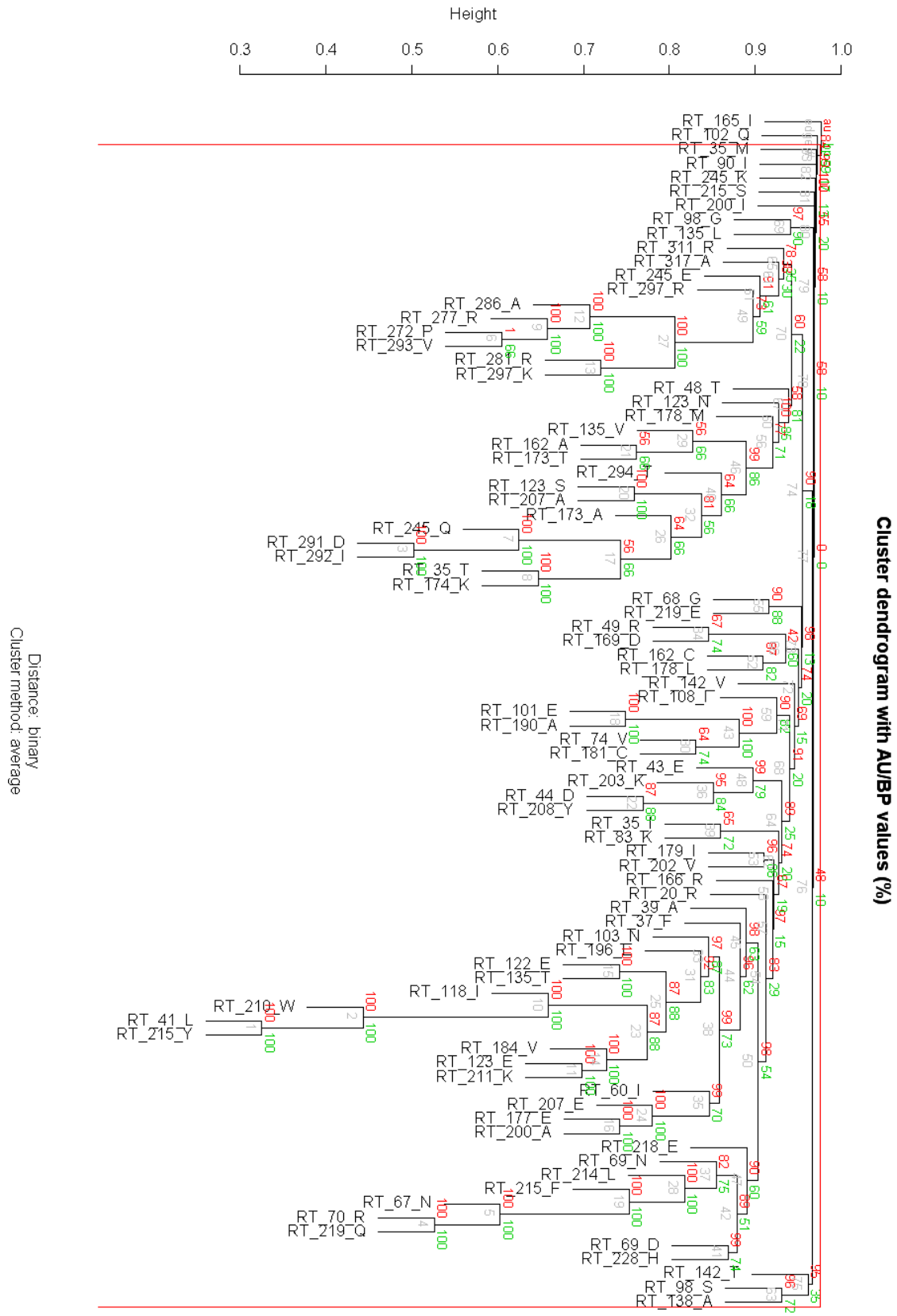

Figure 5.6: Hierachical clustering of Reverse Transcriptase sequences (n=19257)

able to identify overlapping patterns of mutations. Figure 5.7 depicts PC performed on the same population that was used for the HC analysis: a PAM using binary Jaccard similarity was executed, with optimal cluster number selection through silhouette comparison [66]. As shown in the figure, this method not only was able to identify a number of different clusters (similar to those obtained by HC), but also helped to explain the possible roles of mutations such as D67N, that, as already mentioned, are frequently detected both together with TAM1 and TAM2 mutations, thanks to the fact that a mutation can appear in more than one cluster. It is also possible to colour the clusters by decreasing density (in this example going from red to blue) and to draw them according to their relative distances. The two-dimensional representation is achieved by plotting the first two components obtained from a Singular Value Decomposition (SVD, similar to PCA) of the original matrix. However, using this graphical solution, only the 11% of the variance is explained.

Nevertheless, the PAM execution seemed less efficient than other methods at identifying well characterised mutational patterns: a lot of clusters were found (more than 30), most of them with just one mutation.

Using GMM-EM clustering (model based, with optimal number of cluster assignment done by cross validation), we obtained a lower number of clusters, also assigning prior probabilities to each one. A list of the identified clusters with the corresponding prior probability is shown below, including for each cluster the mutations that had a mean prevalence probability ≥ 0.5.

```
{123E} 0.3208
{41L,184V,210W,211K,215Y} 0.1502
{wild-type} 0.1287
{211K,272P,293V} 0.0953
{67N,70R,123E,184V,219Q} 0.0868
{123E,272P,281R,297K} 0.0582
{35T,122E,174K,177E,200A,211K,245Q,286A,291D,292I,293V} 0.0538
{122E,174K,177E,200A,211K} 0.0535
{41L,67N,118I,122E,184V,210W,211K,215Y} 0.0527
```

Again, the identified groups are fairly coherent with those indicated by the HC and PAM analyses: TAM1 and TAM2 were retrieved, although in a mixed cluster including also 184V, as well as the wild-type configurations (including some of the previously clustered together naive polymorphisms). The advantage GMM-EM over the HC analysis is that mutations can belong at the same time to different clusters. Thus, mutations that do not follow any preferential pathway or distribute independently (such as 184V, 211K or D123E) are also evaluated and interestingly can be allocated to clusters despite the fact that they may not have any role in a certain evolutionary pattern.

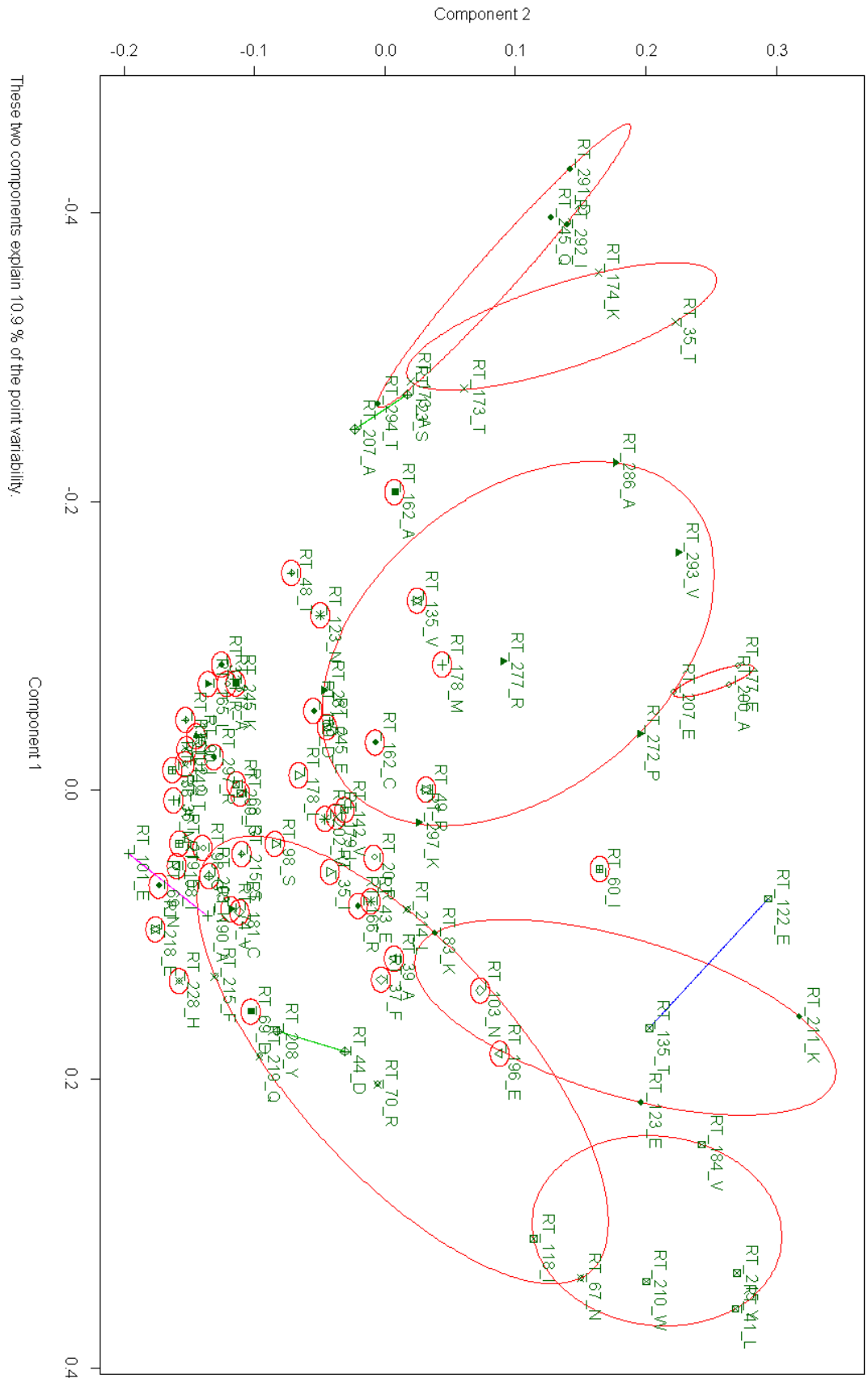

Figure 5.7: Partitional (PAM) clustering of Reverse Transcriptase sequences (n=19257)

As an additional description, I show the results of the MDS method which was executed on the variable distance matrix, reducing the space to two dimensions (figure 5.8), and the PCA solution for the first two components (figure 5.9). As shown in the figure, PCA was only able to discriminate between treated and naive sequences, bacause this takes into account the majority of variance in the mutational distribution.

After restricting the analysis to the set of sequences that were associated with treatment failure, the PCA was able to give a better discrimination (figure 5.10), retrieving TAM1 and TAM2 in the first two components. In contrast, GMM-EM clustering produced a set of clusters in which mutation 41L had a predominant role, Unexpectedly, even in presence of TAM2.

```
{RT_41_L,RT_210_W,RT_215_F,RT_98_any} 0.2289
{wild-type} 0.1529
{RT_41_L,RT_184_V,RT_215_Y,RT_37_F,RT_123_E} 0.1332
{RT_123_E} 0.1078
{RT_41_L,RT_67_N,RT_70_R,RT_215_F,RT_219_Q} 0.1001
{RT_67_N,RT_70_R,RT_219_Q,RT_123_E} 0.0672
{RT_41_L,RT_210_W,RT_215_Y} 0.0484
{RT_41_L,RT_67_N,RT_184_V,RT_210_W,RT_215_Y,RT_203_any} 0.0476
{RT_41_L,RT_67_N,RT_210_W,RT_215_Y,RT_118_I} 0.0454
{RT_41_L,RT_190_A,RT_210_W,RT_215_Y,RT_123_E} 0.0274
{RT_41_L,RT_184_V,RT_215_Y,RT_135_T} 0.0214
{RT_184_V,RT_123_E,RT_135_T} 0.0197
```

5.3.4 Discussion

Given the large number of tests that have been executed, it is difficult to draw specific conclusions on the basis of the results presented in the previous sections, but the investigation was intended to give hints for focusing on more detailed analyses, with the aim of selecting data groups also with more constraints, such as the stratification (or multivariable adjustment) for treatment or treatment history. Moreover, the different clustering techniques revealed that HC – though limited to pairwise agglomerative covariation – depicted very clearly the mutational pattern spaces (confirming the well known TAM patterns and giving new reasonable clusters), whilst PC – that should be more flexible – led to more confusing groups.

Regarding clustering of mutations in the protease gene, we show the results of HC and PCA applied to a very large number of sequences, proving that in this region the patterns are less likely to appear (figures 5.12 and 5.11).

Perhaps PI mutations should be examined specifically one by one in a number of stratified scenario but this should be the purpose of ad-hoc studies.

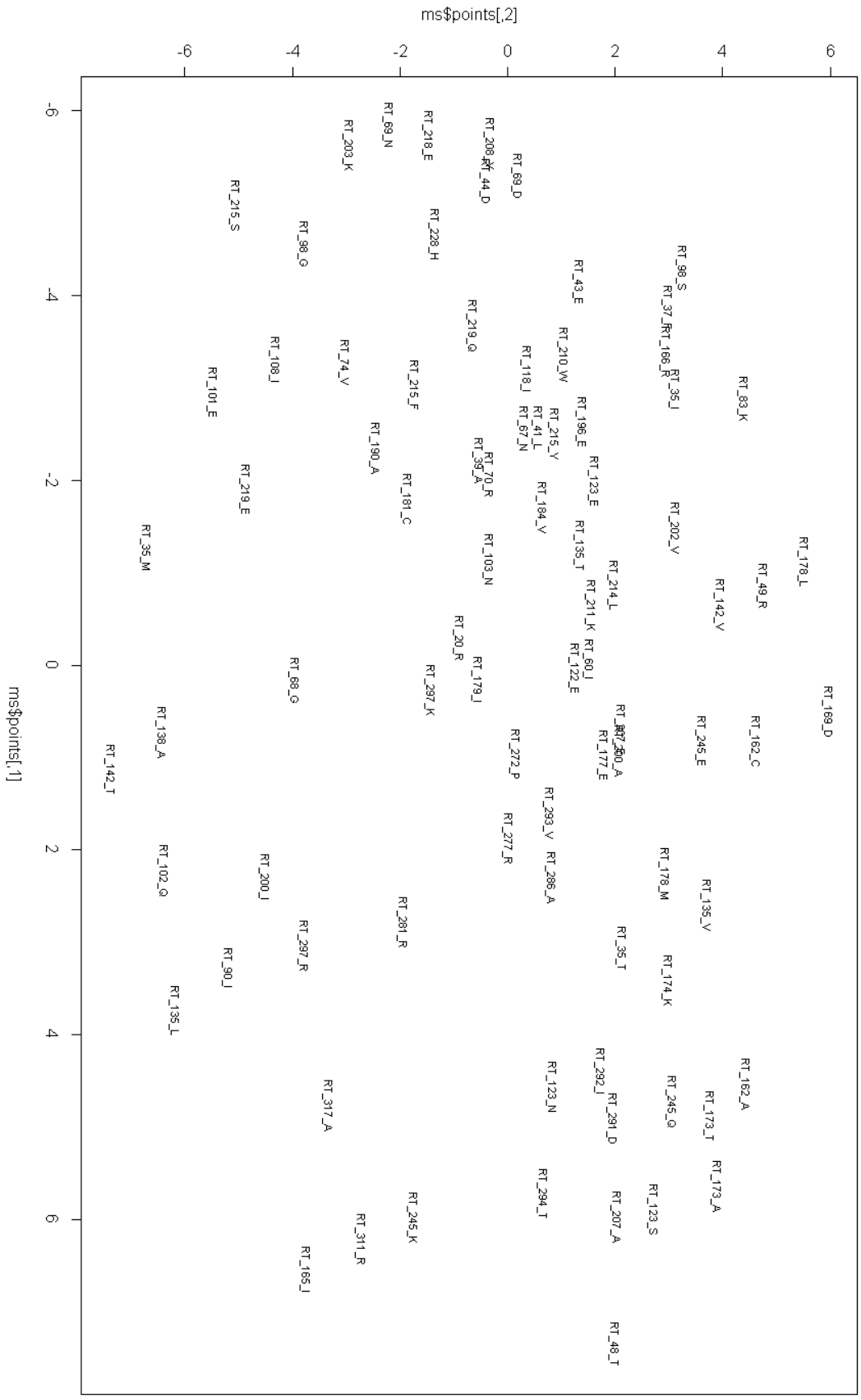

Figure 5.8: Multi Dimensional Scaling of Reverse Transcriptase sequences (n=19257)

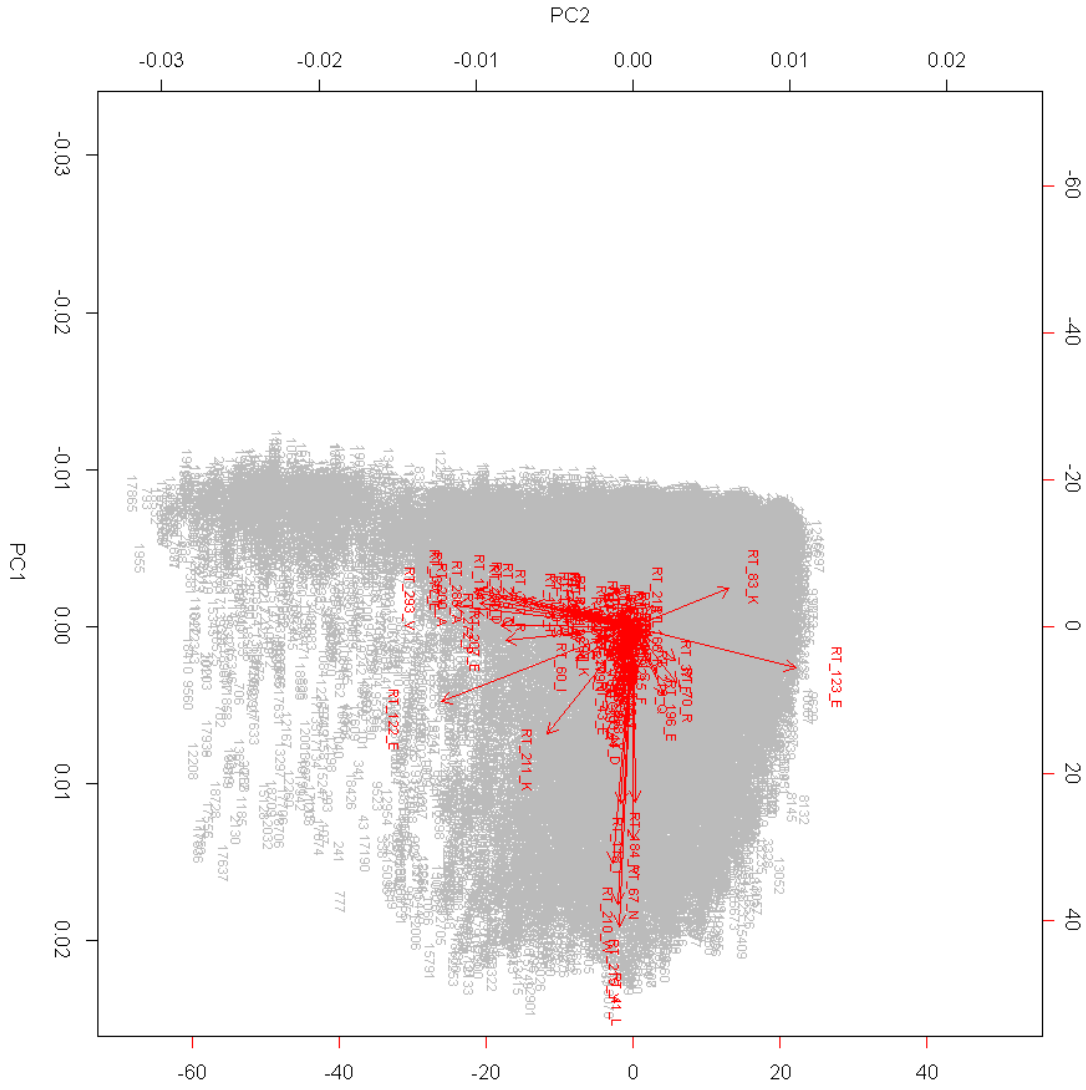

Figure 5.9: Principal Component Analysis of Reverse Transcriptase sequences (n=19257)

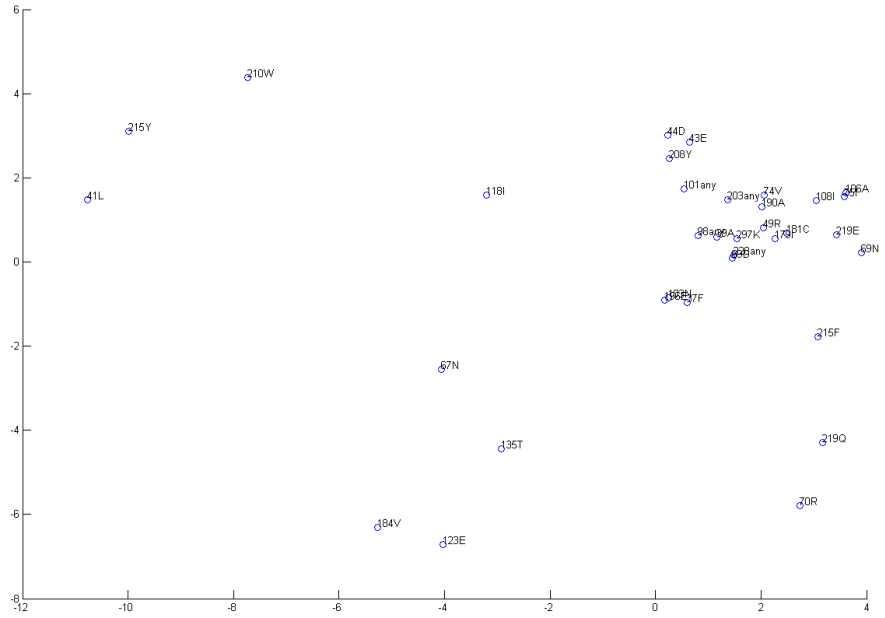

Figure 5.10: Principal Component Analysis of treated sequences leading to virologic failures (n=531).

Because of the incresing variability in type of mutational events, pathway courses and in the effect of drug combinations on HIV evolution, a precise protocol for analysis should be followed. Thus, in conclusion, studies considering genotypes and data on therapies that have been used in the clinics for a long time are generally able to provide more robust results and can also provide useful information for the development of new drugs.

Finally, a number of fundamental steps of the process of data collection and analysis in the context of epidemiological studies of HIV resistance are summarised below:

- collect sequences performed in drug-naive patients
- collect sequences at time of virological failure of treatment (best if the plasma sample was collected at the end of a 6 months period of uninterrupted drug pressure)
- control inthe analysis for treatment backbones or other covariates such as year of test or number of previously used/failed therapy lines
- collect phenotypic data

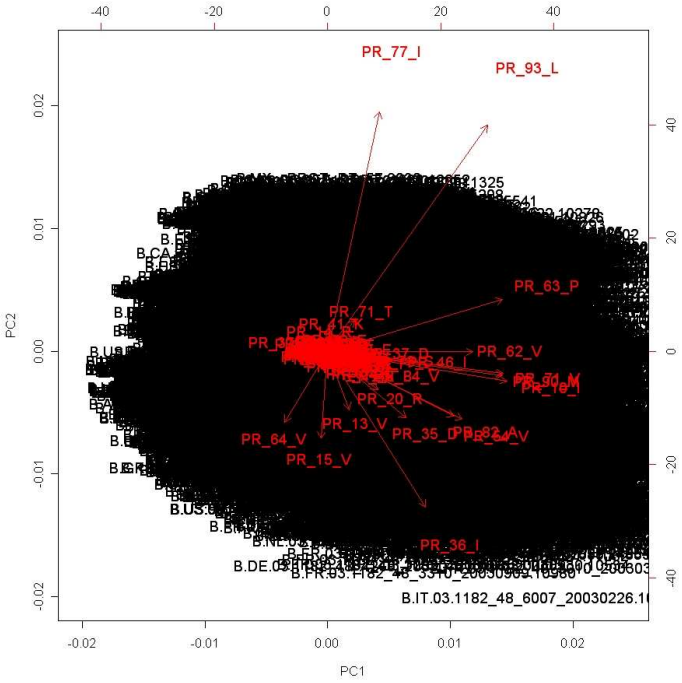

Figure 5.11: Principal Component Analysis of protease mutations (n=27211).

- collect clinical markers before and after therapy initiation and multiple follow up sequences

- perform univariable χ^2 analysis to compare the prevalence of mutations in naive/treated groups after adjusting for multiple testing

- perform univariable χ^2 (stratified) analysis adjusted for multiple testing and multivariable (e.g. linear or logistic regression models including covariates such as RTI/PI backbones, et cetera, looking not only at p-values, but also at goodness of fit) with outcome the percentage of patients with a viral load ¡50 or ¡500 cp/ml in certain time window after starting therapy

- perform adjusted univariable analysis (using the Wilcoxon rank sum test) to compare the median phenotypic log fold change according to the detected mutations, testing also other multivariable models (linear, non-linear like random forests) that give feature evaluation.

- select mutations showing significant association with the outcomes of interest (e.g. drug history, virological response)

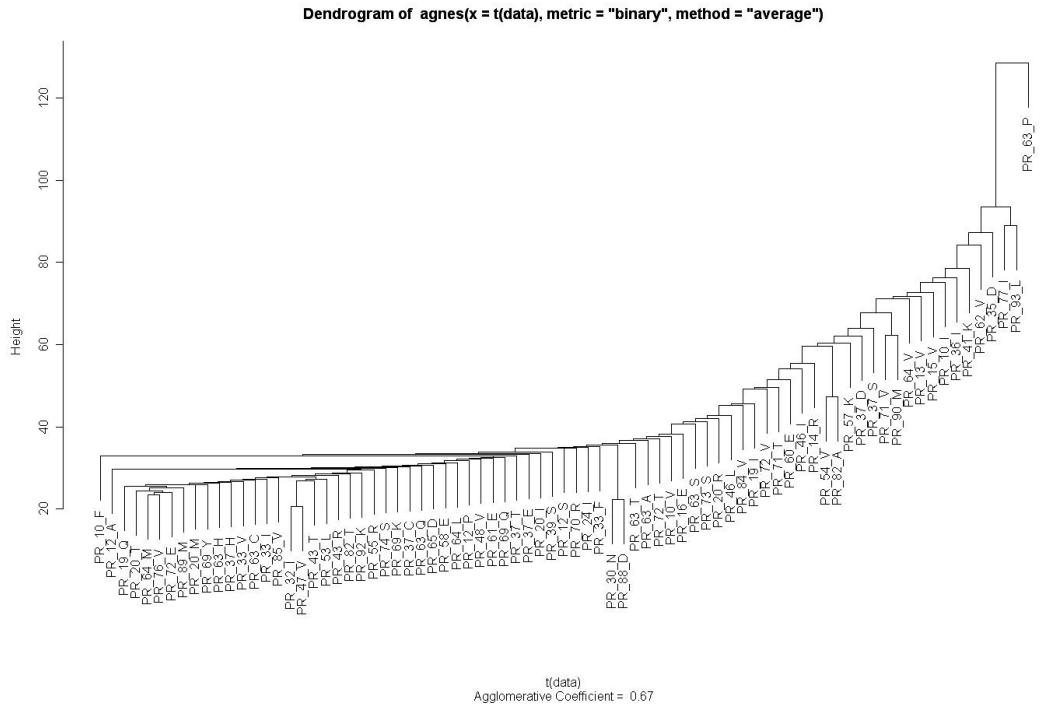

Figure 5.12: Hierarchical Clustering of protease mutations (n=27211).

- perform correlation and covariation analysis among relatively prevalent mutations in order to find positive and negative associations between them (cross tabulations or clustering, eventually tested against random scores)
- perform *time to event* type analyses (aka survival analysis: univariable Kaplan Maier curves and multivariable Cox regression) to test whether a certain mutation is associated to the rate of development of other mutations and/or virological response to treatment

Chapter 6

Markov Chain Models

In this chapter we make the hypothesis that the evolution of HIV genome under a fixed and continuous drug pressure is a stationary ergodic process: in other words, we suppose that the information present in a viral genotype (i.e. the mutations) – during a period in which therapy is unchanged – determines the probabilities for the evolutionary pathways, regardless the evolutionary steps that drove to that particular genotype.

6.1 Methods and Data

6.1.1 Theory

In probability theory, A stationary ergodic process is a stochastic process which exhibits both stationarity and ergodicity. In essence, this implies that the random process will not change its statistical properties over time.

Stationarity is the property of a random process THAT guarantees that its statistical properties, such as the mean value, its moments and variance, will not change over time. In other words, a stationary process is one whose probability distribution is the same at all times.

An ergodic process is one which conforms to the ergodic theorem. The theorem allows the time average of a conforming process to equal the ensemble average. In practicaL terms, this means that statistical sampling can be performed in one instant across a group of identical processes or sampled over time on a single process with no change in the measured result.

In physics and thermodynamics, the ergodic hypothesis says that, over long periods of time, the time spent in some region of the phase space of micro-states with the same energy is proportional to the volume of this region, i.e. that all accessible micro-states are equally probable over a long period of time. The

ergodic hypothesis is often assumed in statistical analysis. The analyst often assumes that the average of a process parameter over time and the average over the statistical ensemble are the same. Irrespective of whether this is true or not (but in some scenarios it can be proven mathematically), the analyst assumes that observing a process for a long time is equivalent to sampling many independent realisations of the same process. The assumption seems inevitable when only one stochastic process can be observed, like, for example, the variations of a price on the market. It can be easily demonstrated that this hypothesis is often erroneous.

In summary, we can say that a process is stationary if its averages are independent of the starting sampling time. A process is ergodic if the temporal averages (calculated on any process realisation) are exactly the same of the ensemble averages calculated at any instant. A stationary process is not necessarily ergodic; in fact, the stationary property implies that the averages are independent of the date of origin, but does not require that these averages are always the same in any process realisation. Vice-versa, an ergodic process is always stationary.

A Markovian process is a stochastic process with limited memory (one backward step for first-order). A Markov Chain is defined as a Markovian process, with discrete time t parameter and x_t random variable, that can assume finite values (each value is a state and the process is said finite-state). To study the evolution of this dynamical system, we calculate the probability that in a certain time t the random variable x_t will be in a state $p(x_t) = i, i = 1..s$, where $S = \{1..s\}$ is the state set.

The two main properties of Markov Chains are, given conditional probability $p(a, b) = p(b|a)p(a)$:

- Markov property: $p(x_k = i_k|x_{k-1} = i_{k-1}, \ldots, x_0 = i_0) = p(x_k = i_k|x_{k-1} = i_{k-1})$, i.e. the probability of being in a certain state at time k depends only on what happened immediately before.

- Stationary property $p(x_k = j|x_{k-1} = i) = p(x_1 = j|x_0 = i)$, i.e. the probability to go from state i to j is the same irrespective of time history.

A stochastic P matrix, whose elements are $p_{ij} = p(x_1 = j|x_0 = i)$, is the transition matrix P. The transition matrix, together with the initial probabilities vector $p_0 = (p_1, \ldots, p_s)$ is sufficient to describe the entire system. Moreover, the probabilities after k steps are: $p_0 P^k$.

The transition matrix P is then defined *regular* if all the elements of any n-th power P^n are positive. If P is regular (neither reducible or periodic) then the following holds:

- $\lim_{n\to\infty} P^n = P^*$

- the system is *ergodic*

- If all the rows of P^* are identical, then the system is completely ergodic and the state of the system ad infinite does not depend on the initial one

Figure 6.1 explains with a simple example a Markov chain graph: suppose that the stock market is defined by two states {positive, negative} and that the trend is assumed to be dependent only on the previous configuration.

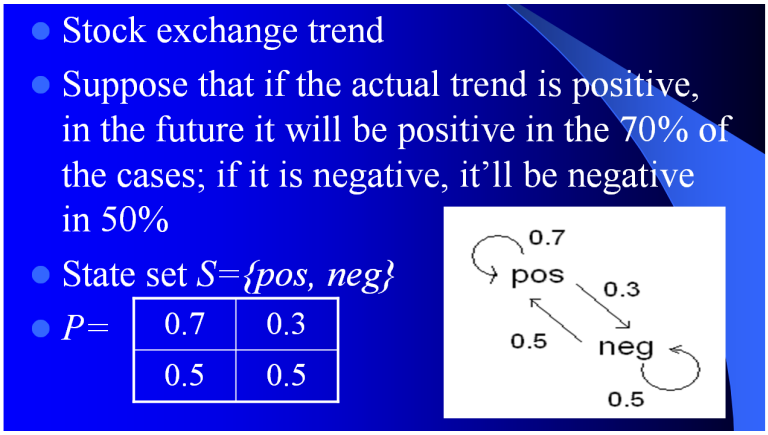

Figure 6.1: Markov Chain: a simple example for stock market trend evolution.

6.2 State of the Art

One attempt to model the HIV-1 evolution has been presented in [122], in which a Markov chain of viral evolution was constructed using data from viral clones coming from Protease Inhibitor-naive patients. The procedure used to estimate the transition probabilities is extremely interesting and it is worth having a look at the assessment of the confidence values for the probability values in the matrices (and because data were *partially* longitudinal). The number of states was created by clustering mutations with the k-means algorithm (see previous chapter for a descritpion of this method). A not so large number of states in the process permits a better estimation of the transition probabilities, but, potentially, only evolutionary end-points are taken into account. Clusters in fact are intrinsically partitioning the data, so the estimated transition matrix is likely to be diagonal (i.e. states are conservative).

In this chapter a more detailed analysis is performed (see [2]): the idea is to focus on a pre-specified therapy combination, in order to remove the confounding effect associated with the use of multiple drugs: moreover, states will be selected on the basis of the most frequent mutations patterns (e.g. using a rougher way

than the partitioning analysis described above) and grouping the remaining states altogether into a "dummy-bridge" state. Of note, only resistant-associated mutations will be considered. Transitions will be estimated by relative frequencies – because longitudinal data is available – and errors assessed through confidence measures on counts.

The proposed solution aims at reducing the huge number of states defined by the full genotype space, which would be theoretically 2^m, where m is the number of codon considered, without losing the intermediate configurations that a clustering could ignore (an improvement from the previous analysis in [122]); in addition, the dummy-bridge state tends to account for low-frequency combinations: another solution will be introduced in chapter 7, where the Markov states will be reduced to a subset of state transitions in which mutations can only be accumulated, while in the model described here mutations can also be lost.

6.2.1 Domain Coding and Descriptive Statistics

Consider a set of states defined by binary codes representing mutational configurations of genotypes. Then, it is possible to estimate the transition probabilities matrix from a data set including consecutive genotypes pairs while a patient was kept on a fixed cART regimen in discrete, roughly constant time steps. Of note, there are more accurate procedures to estimate probabilities from non-constant time steps (see again [122]) but they will not be used here. Figure 6.2 shows the coding for the mutation vectors.

392 consecutive genotype pairs were extracted from the ARCA [102] data base form patients who were kept on a AZT+3TC containing treatment (this regimen could contain NNRTI or PI, but not any other NRTI). 18 codons out of the IAS/USA [121] NRTI resistance list were considered (2004 update): {41, 44, 62, 65, 67, 69, 70, 74, 75, 77, 115, 116, 118, 151, 184, 210, 215, 219}. If there was a change from consensus at specific codon, no effort was made to distinguish between speific amino acidic substitutions.

The mean time between two genotyping was 173 days (st.dev 146), while the mean therapy duration was 780 days (st.dev 499) from samples collected between 1993 and 2004. 119 different mutational patterns were identified from the 784 genotypes. Five states were selected according to uncertainty measure on counts $< 20\%$ (high frequency patterns): {184}, {wild type}, {70, 184}, {41, 184, 215}, {70}, other low frequency patterns were unified in {other} state.

The complete Markov chain transition matrix was estimated by relative frequencies, with time step $t = 6$ months (i.e. rougly 180 days).

AZT+3TC

Pat_IDX	t (months)	t+dt	genotype(t)	genotype(t+dt)	state(t)	state(t+dt)
#1	0	6	wildType	M184V	0,0,0,0	0,1,0,0
#2	0	6	wildType	M41L, T215Y	0,0,0,0	1,0,0,1
#3	0	6	wildType	M184V	0,0,0,0	0,1,0,0
#2	6	12	M41L, T215Y	M41L, T215Y	1,0,0,1	1,0,0,1
#1	6	12	M184V	M184V	0,1,0,0	0,1,0,0
#2	12	24	M41L, T215Y	M41L, L210W, T215Y	1,0,0,1	1,0,1,1

Figure 6.2: State coding for HIV-1 data.

6.3 Results

A graph of the estimated transition probabilities of this Markov chain is shown in figure 6.3. The estimation procedure identified three conservative states: {184}, {wild type} and {41, 184, 215} (range 53-58%). This finding seems to be consistent with evidence showing that strains of HIV harbouring the mutation M184VI have increased fidelity [120], i.e. are less likely to drift.

Interestingly, the model also shows that the viral populations that are kept under the pressure of AZT+3TC containing therapy are likely to remain wild type, i.e. a combination including this nucleoside pair has high genetic barrier: this is because (if present) a third drug was an active PI or NNRTI, in fact under pure dual therapy M184VI would develop more quickly. Of note, the estimated probabilities of the transition matrix are here prone also to the variability of the other drugs present in the regimen.

However, the model also suggests that when the mutation K70R is detected, either with M184V or by itself, the chance of remaining in this state is reduced from 58% to 37% (i.e. the probability of accumulating further mutations increases). States with K70R were predicted to be antagonist with states containing M41L and T215Y (no arrow between the 2 states, in agreement with the

previously reported TAM1 vs. TAM2 patterns, see section 5), while M184VI was linked with either of these two states with similar probabilities, indicating that no preferential association exists with TAM1 or TAM2 profiles.

From the {wild type} state there was a probability of 0.43 to accumulate other mutations, with 14% chance to develop {184} and a smaller chance of suddenly acquiring {41, 184, 215} at the same time or {70, 184} or {70}. Overall, the probability for a mutated strain to revert back to {wild type} was in the order of 0.1. This is consistent with the results of epidemiological studies showing that, under the pressure of continuous treatment, there is very little chance to revert from mutant to wild type. Actually, the only reasonable way in which we foresee this is in the poor patient's adherence or treatment interruption.

From a methodological point of view, the limit of the n-th power of the transition matrix T converged and the rows were identical, indicating that both conditions for the process to be complete ergodic were satisfied, i.e. the system state ad-infinitum does not depend on the initial one.

The n-th power of T is: {0.1754, 0.3294, 0.0356, 0.0470, 0.0312, 0.3813} referring to the state ordering in table 6.1. From this vector, there is an almost equal probability to remain wild type or to develop mutations, but no preferential pathway is found.

Conservative states can be also found through clustering. In order to make comparisons, we executed Partitional Clustering (Gaussian Mixture Modelling GMM using Expectation Maximisation EM algorithm for the selection of the optimal number of clusters, see again 5). The clustering was performed using the same 784 genotype vectors: 5 clusters were selected, as shown in figure 6.4: (i) {41, 184, 215} (prior probability 0.19); (ii) {67, 70, 184, 215, 219} (prior probability 0.09); (iii) {77, 116, 151} (prior probability 0.01); (iv) {wild type} (prior probability 0.58); (v) {70, 184} (prior probability 0.13). The clusters identified with this method nicely match the Markov states. Moreover, GMM-EM WAS able to isolate the rare Q151_complex cluster which, in contrast, was allocated to the dummy-bridge state by Markov modelling because of its low frequency. We foresee here that still a proper configuration for the states is needed.

6.4 Conclusions

In order to model a first-order Markov chain following the procedure described above, a large number of states are required and this may lead to possible wrong estimates for the transition probabilities, especially if the number of genotype pairs is relatively small (119 different states having 392 transition pairs). But we state that the clustering approach could not be appropriate since is more prone to find already evolved patterns. Here, the transition matrix has been estimated

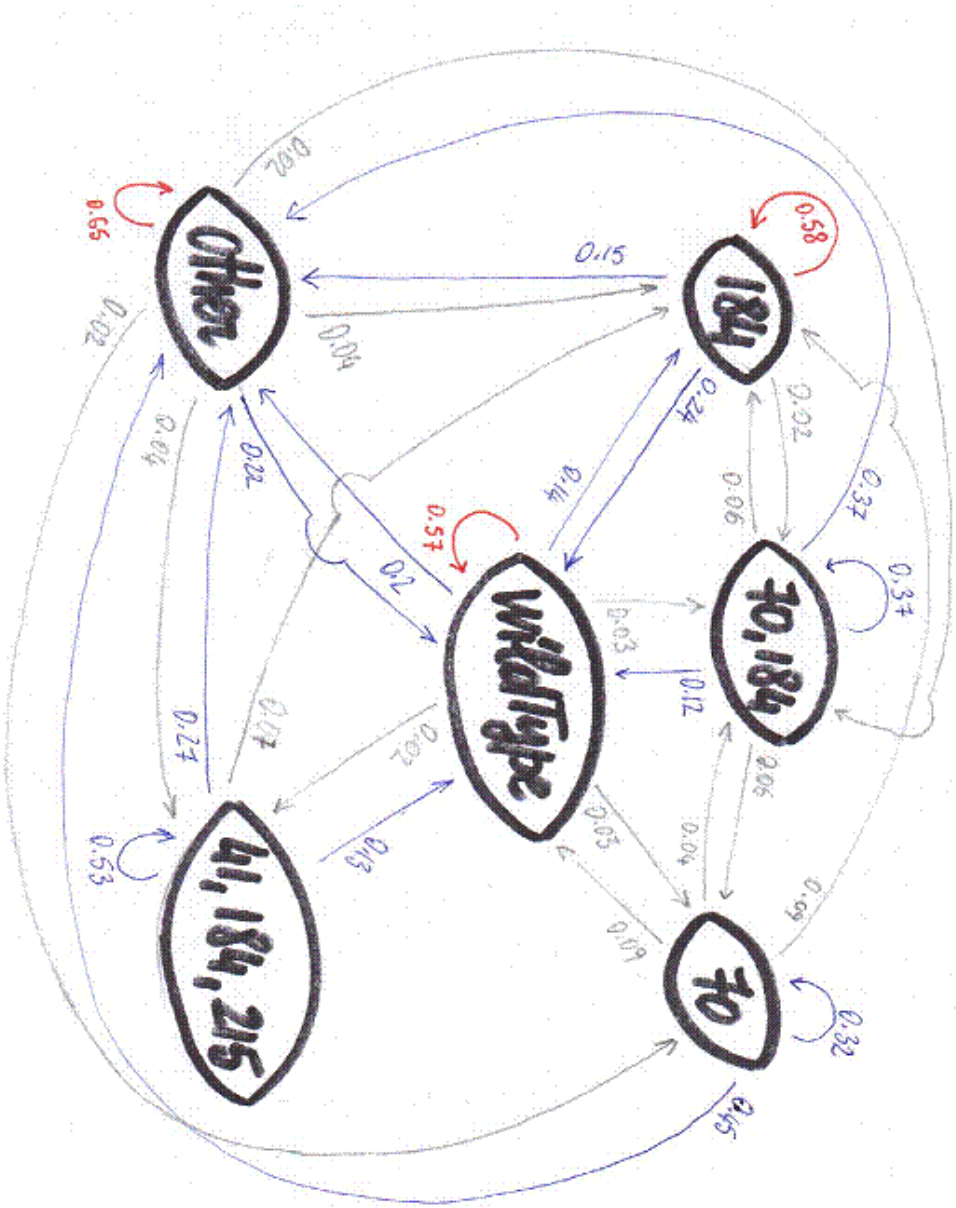

Figure 6.3: Markov Chain Graph for HIV-1 Evolution under AZT+3TC continuous drug pressure.

T	{184}	{wild type}	{70, 184}	{41, 184, 215}	{70}	{other}
{184}	0.5854	0.2439	0.0244	0	0	0.1463
{wild type}	0.1419	0.5743	0.027	0.0203	0.027	0.2095
{70, 184}	0.0625	0.125	0.375	0	0.0625	0.375
{41, 184, 215}	0.0667	0.1333	0	0.5333	0	0.2667
{70}	0.0909	0.0909	0.0455	0	0.3182	0.4545
{other}	0.0467	0.22	0.02	0.04	0.0267	0.6467

Table 6.1: Transition Matrix.

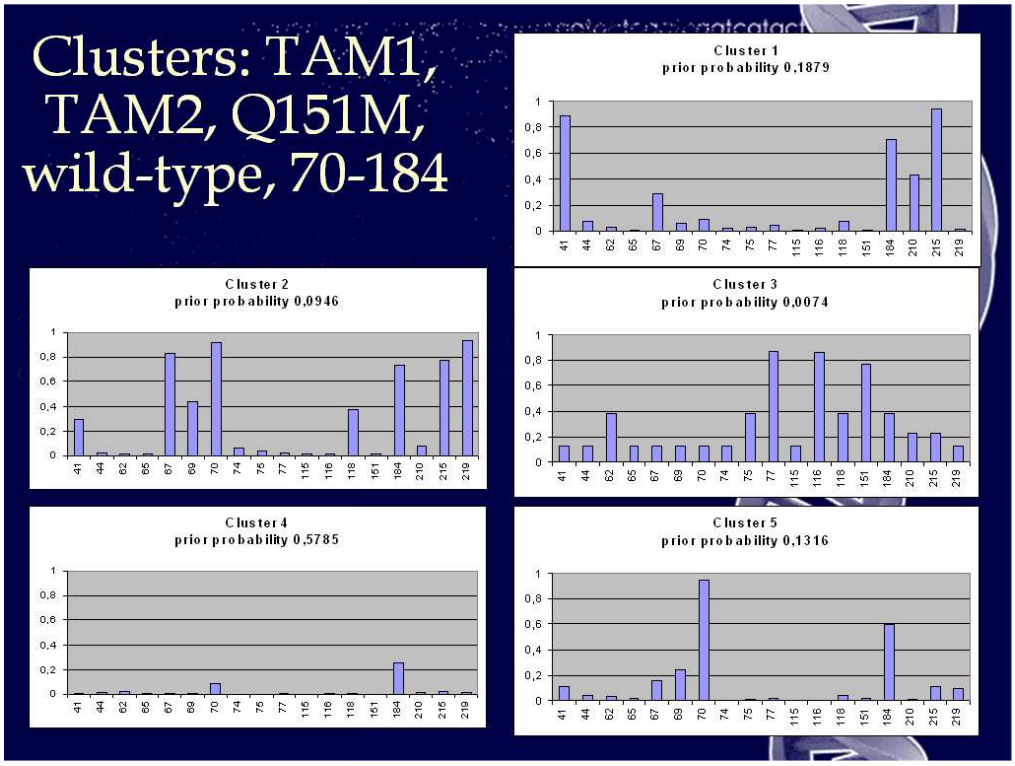

Figure 6.4: GMM-EM clustering for mutational pattern identification under AZT+3TC continuous regimen.

considering constant time steps: a more accurate algorithm needs to be applied to take into account variable time steps.

Dimension reduction must be investigated: a way to do this could be the modelling through Petri Nets (PN).

If opportunely modelled, Petri Nets would allow the estimation of the probability of following different mutational pathways over a long-term period with treatment, thus potentially being able to provide useful information to the physicians, regarding – for instance – the durability of a certain treatment or the risk of developing cross-resistance.

6.4.1 Future Perspectives: Petri Nets

Petri Nets were introduced by A.Petri in 1962 [96]. This type of models are able to give a compact representation of systems with large state space. In fact, PN do not require to represent explicitly all the possible state values of a system, but only the rules that determine its evolution.

A Petri Net is one of several mathematical representations of discrete distributed systems. As a modelling language, it graphically depicts the structure of a distributed system as a directed bipartite graph with annotations. As such, a Petri Net has *place* nodes, *transition* nodes, and directed arcs connecting places with transitions. The difference between a place and a transition, in summary, is that a place represents an event variable (like a mutation in HIV) and a transition represents a causal effect propagating from a place configuration to another. There are then *tokens*, that propagate from places to transitions: roughly, they can be interpreted as activating the places through time being propagated by transitions from a place to another. Arcs run between places and transitions. The places from which an arc runs to a transition are called the input places of the transition; the places to which arcs run from a transition are called the output places of the transition. Places may contain any number of tokens. A distribution of tokens over the places of a net is called a marking. Transitions act on input tokens by a process known as firing. A transition is enabled if it can fire, i.e. there are tokens in every input place. When a transition fires, it consumes the tokens from its input places, performs some processing task, and places a specified number of tokens into each of its output places. It does this atomically, i.e., in one non-interruptible step. Execution of PN is nondeterministic: multiple transitions can be enabled at the same time, any one of which can fire; none are required to fire – they fire at will, between time 0 and infinity, or not at all (i.e. it is totally possible that nothing fires at all). Since firing is nondeterministic, Petri Nets are well suited for modelLing the concurrent behaviour of distributed systems. So far, simple PN are able to describe sequencing, parallelism, synchronicity, concurrency and choice. Moreover, several extensions have been designed in order to take into account other properties (coloured PN,

stochastic PN [71] [62]).

In order to model the mutational pathways, however, my feeling is that PN should be extended with *inhibitory* arcs and stochastic transitions. Stochastic Petri Nets SPN [62] are a formalism developed in the field of computer science for modelling system performance. SPN consist of places and transitions as well as a number of functions. The basic functions are input, output and weight functions. The initial state of the system is represented by the initial marking. SPN can be represented graphically, with places represented as circles and transitions as rectangles, and input and output functions as directed arcs. An example of an SPN might be a computer system consisting of two processors which receive jobs from a common buffer. This system might have 3 places (buffer, processor 1 and processor 2) and 5 transitions (jobs arriving in buffer, jobs being transferred to one of the processors (N=2) and jobs being completed by processors (N=2)).

SPN have discrete state spaces, defined by the number of objects in each place (the marking). Places can be linked to transitions as input places, and transitions can be linked to output places. Transitions are said to be enabled when there are enough objects in each of the input places. Enabled transitions can fire, removing objects from their input places and adding objects to their output places. Enabled transitions fire according to exponential distributions, characteristic of Markov processes.

Petri Nets have been widely applied also in biology [22], for the modelling of gene-regulatory networks or chemical reactions. In a schematic way, we can describe the configurations for modelling chemical reactions:

- Place $\leftrightarrow$ Molecular Species
- Transition $\leftrightarrow$ Reaction
- Input Arc $\leftrightarrow$ Defines reagent of reaction
- Output Arc $\leftrightarrow$ Defines product of reaction
- Weight Function $\leftrightarrow$ Rate of reaction
- Marking of SPN $\leftrightarrow$ State of reaction system

Having the number of molecules of each species, the initial marking is the initial state of the system. The interpretation for a transition being enabled is that enough of all the reagents must be present for the reaction to complete, while the interpretation of a transition firing is a single molecular reaction.

In a similar way, we could define a PN that models the mutational pathways of HIV, where a place correspond to a mutation (plus an initial place that correspond to the wild-type) and a transition allows for the emergence of new mutations, with the possibility to lose or maintain in the dominant population the one accumulated. The graphical representation is shown in figure 6.5: this

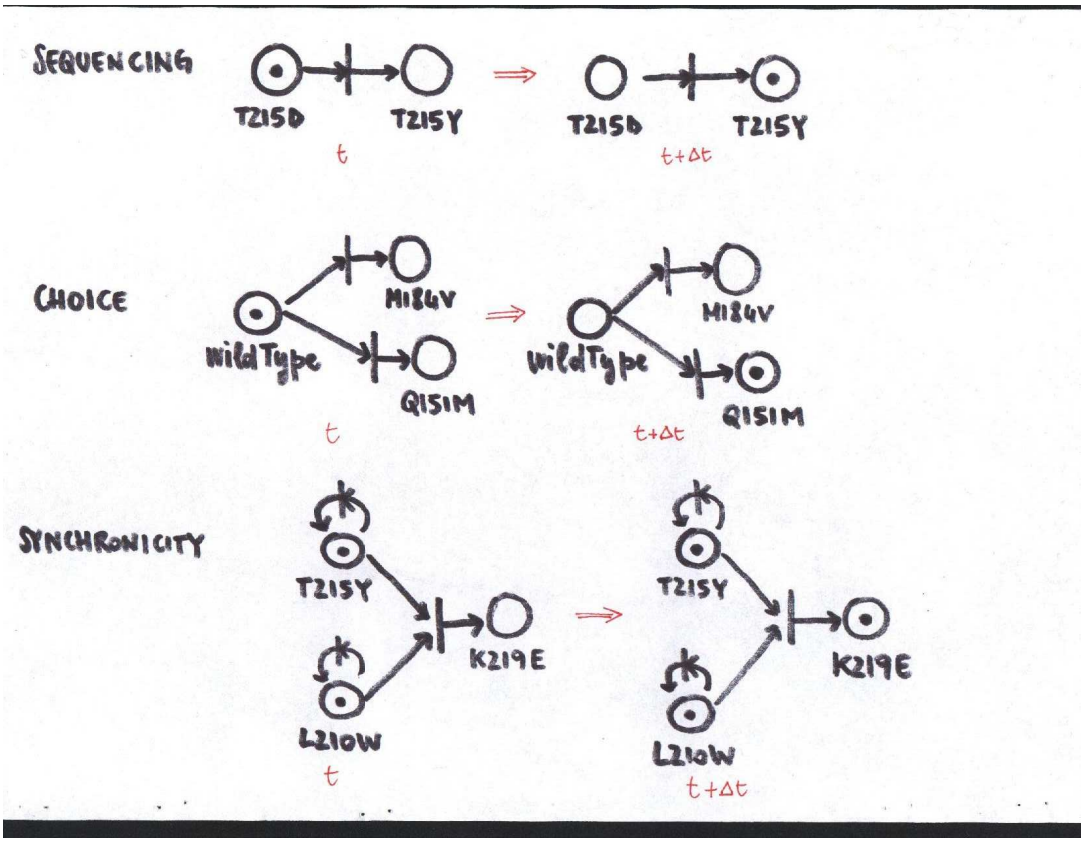

Figure 6.5: Petri Net Modelling of Mutational Pathways.

is a simplified graph, because the transition probabilities are not shown. The difference between an empty circle and a circle including the token is in the sense that the presence of a token codes the presence of a mutation in the population.

Within the context of using PN models to model HIV evolution, it would be easy to define the suitable configurations for mutational events, but there is still a main problem: the topology of the net that best suits this phenomenon needs to be discovered. In fact, the number of possible transitional events is the same as the number of states of a complete Markov chain and this issue needs to be addressed. Rather than a search on the power set, a possible way of dealing with this problem would be to use heuristic algorithms, maybe starting with low-order interactions (one-place transitions, two-place, et cetera...), or genetic algorithms.

The advantage of PN against Markov chain models would be that no selection of states is needed, whilst the advantage against Mutatgenetic Trees models would be the possibility to explore more flexible pathways, allowing also for mutation loss.

Chapter 7

Mutagenetic Trees

In this chapter we'll introduce *Mutagenetic Trees* and *mixtures* of Mutagenetic Trees. These models represent a restricted Markov chain model (and also a special case of a Bayesian network) in which only subsets of the space state can be reached: the assumption is that there can be *only* an accumulation of single events, i.e. – if the acquisition of a mutation is an event – mutations in the viral genome cannot be lost. This constraint permits the shrinkage of a considerable number of states (a problem that we partially solved in the previous chapter with the usage of a "dummy" bridge state), but has the drawback to ignore all the configurations in which mutations revert to the wild type amino acid.

In the next section a set of theoretical bases will be given, along with the validation of different trees configurations on large clinical data sets. Results will show that – however – the fashionable Mutagenetic Trees Mixture models are not able to describe data (in terms of likelihood) than simpler models that take into account either only independent behaviour of mutations or mutation accumulation which is not position-specific (i.e. only the number of acquired mutations is important).

7.1 Theoretical Bases

7.1.1 Data Representation

Let $\{1, \ldots, l\}$ be a set of genetic events. A *mutagenetic tree* on l events is a tuple $\Upsilon = (V, E, r, p)$:

- $V = \{0, 1, \ldots, l\}$ is the vertices set with *null event* X_0 always observed and binary random variables $X_1, \ldots, X_l$ indicating random occurrence of mutational events

- E is the set of edges, with $p : E \to [0, 1]$ such that
  - (V, E, r) is a connected branching rooted at 0
  - for all edges $e = (u, v) \in E$: $p(e) = Pr(X_v = 1|X_u = 1)$
  - we set $Pr(X_v = 1|X_u = 0) = 0 \quad \forall (u, v) \in E$ (i.e. an event can occur only if its predecessor in the tree has occurred)
- each vertex has at most one entering edge in a branching, so $p_v = p(u, v)$ for $e = (u, v) \in E$

The mutagenetic tree model Υ is a Bayesian tree model with transition matrices

$$(P(X_v = b|X_u = a))_{a,b=0,1} = \begin{pmatrix} 1 & 0 \\ 1 - p_v & p_v \end{pmatrix}$$

on each edge (u, v).

7.1.2 Likelihood Computation

A mutagenetic tree Υ induces a probability distribution on the set of all mutational patterns. the probability that Υ generates a sample x is:

- Let $S \subseteq V$ the set of events specified in x
- If there exists a subset $E' \subseteq E$ such that S is the set of all vertices reachable from r in the induced subtree $\Upsilon = (V[E'], E')$ then x can be generated by Υ and
$$P(x|\Upsilon) = \prod_{e \in E'} p(e) \cdot \prod_{e \in (S \times V \setminus S)} (1 - p(e))$$
- If there is no such edge subset, the topology of Υ does not allow for generating x and hence $P(x|\Upsilon = 0)$

The likelihood computation can be done efficiently by traversing the mutagenetic tree in a breadth-first search starting from the root.
We call Υ a *star* if all edges $e \in E$ leave the root vertex r. the star topology models events as being independent of each other. In terms of the likelihood a star is characterized as:

Lemma 1. *A mutagenetic tree is a star if and only if all 2^l possible patterns of events have positive likelihood.*
Proof: If Υ is a star,

$$L(x|\Upsilon) = \prod_{j|x_j=1} Pr(j) \cdot \prod_{j|x_j=0} (1 - Pr(j)) > 0$$

since Pr(j)=Pr(j|r) $\in (0, 1)$ by definition. If Υ is not a star, there's at least one edge (j_1, j_2) with $j_1 \neq r$ and any pattern with $x_{j1} = 0$ and $x_{j2} = 1$ has likelihood zero.

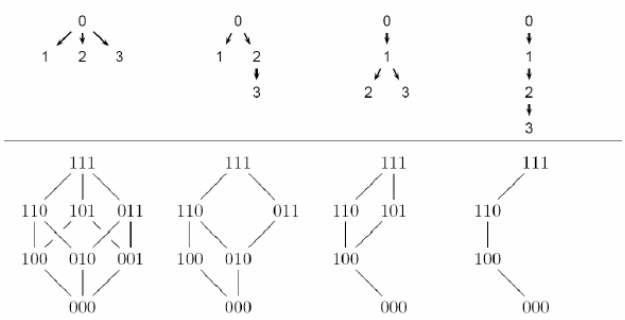

Figure 7.1: Different trees and corresponding reachable Markov states

7.1.3 Tree Reconstruction

The structure and the parameters of a mutagenetic tree can be estimated efficiently from all pairwise probabilities of genetic events applying Desper's algorithm [3]. The structure is reconstructed as the solution to the maximum weight branch problem [65] in the complete graph on $l+1$ vertices with weights w that depend only on the pair probabilities as

$$w(u,v) = logPr(u,v) - log(Pr(u) + Pr(v)) - log(Pr(v))$$

easily estimated from counts in the data. Since trees can only represent a limited set of acyclic dependency structures, this model class is too restricted for many applications.

7.1.4 Explanation

So, a mutagenetic tree is a tree and from its root (that represents the event that always happens) depart (many) arrows. Each arrow is an event that can happen with a certain probability p and going through vertices means that you acquire that event and don't lose it. Multiple arrows branching from the same node explain different pathways, but are not mutually exclusive as one could think at first sight. That's why the sum of probabilities in each level of the tree can be more than one. Each p of an edge (u,v) in fact describes the probability to acquire event v after you gained event (u). If you want to calculate the probability of a certain pattern x you simply multiply the probabilities of the path that drives you to your pattern (if you can get it, otherwise is zero), multiplying again by 1 minus the first step of the other paths and the path that is following. Look

at figure 7.1 to see how the compact tree draws represent the reachable Markov states.

7.1.5 Mixture Models

Definition

Suppose that $Y_1, \ldots, Y_k$ are multivariate discrete random variables with range Ω that are distributed according to mutagenetic trees

$$\Upsilon = (V, E_k, r, p_k), \quad k = 1 \ldots K$$

respectively. Let $\Delta_1, \ldots, \Delta_k \in \{0, 1\}$ be binary random variables with $\Pr(\Delta_k = 1) = \alpha_k$. We call the model

$$\Psi = \sum_{k=1}^{K} \alpha_k \Upsilon_k$$

with $\alpha_k \in [0, 1]$ and

$$\sum_{k=1}^{K} \alpha_k = 1$$

that generates the random variable $Y = \sum_{k=1}^{K} \Delta_k Y_k$, a *K-mutagenetic tree mixture model*.
Thus, the likelihood of a pattern of events x in the mixture model is

$$L(x|\Psi) = \sum_{k=1}^{K} \alpha_k L(x|\Upsilon_k)$$

Throughout we'll consider mixture models that have a special structure in the first mutagenetic tree Υ_1. We assume that, in addition to different pathways of accumulation of events, there's a certain probability β of any event occurring spontaneously independent of all other events. Thus, Υ_1 is a star with $p(e) = \beta$ for all $e \in E_1$ (but it can be possible also to define different βs for each event). Υ_1 is called the *noise* component of the model. Including a star in the mixture model ensures that all patterns of events have positive likelihood.
In figure 7.2 we show a 3-mutagenetic tree mixture model estimated from HIV-1 genotypic data coming from patients treated with zidovudine (AZT). From the wild type (null event) each node represents the accumulation of a resistant mutation.

7.1.6 EM-like Learning Algorithm

Given the number of trees K, we want to reconstruct a K-mutagenetic trees mixture model from observed patterns X. This task would be easy if we knew for

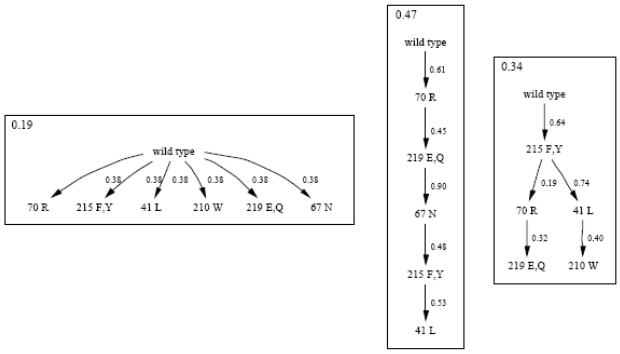

Figure 7.2: 3-mutagenetic trees mixture model $\Psi = 0.19\Upsilon_1 + 0.47\Upsilon_2 + 0.34\Upsilon_3$ for the development of AZT resistance

each pattern of events from which component(s) of the model it has been generated: we would apply K times the reconstruction technique for a single tree. However this information is missing and we have to estimate it from the data too. This procedure results in an algorithm similar to the Expectation Maximisation (EM) algorithm [28]. Our goal is to find mutagenetic trees $\Upsilon_1, \ldots, \Upsilon_K$ and mixture parameters $\alpha_1, \ldots, \alpha_K$ that maximise the log_likelihood of the data, which can be written as

$$\sum_{i=1}^{N} log \sum_{k=1}^{K} \alpha_k L(x_i|\Upsilon_k)$$

if the x_is are independent. The *responsibility* of model component k for sample x_i is defined as

$$\gamma_{ik} = \Pr(\Delta_k = 1|\Psi, x_i)$$

Let $N_k = \sum_{i=1} N\gamma_{ik}$ be the weighted number of samples generated by Υ_k. In an iterative fashion, we estimate γ (E step) and Ψ (M step) from the data. Given an estimate of $\Psi = \sum_{k=1}^{K} \alpha_k \Upsilon_k$, we can estimate γ by

$$\gamma_{ik} = \frac{\alpha_k L(x_i|\Upsilon_k)}{\sum_{m=1}^{K} \alpha_m L(x_i|\Upsilon_m)}$$

Given an estimate of γ we update Ψ as follows. For the noise component ($k = 1$) we choose the star topology and estimate β as the rate of occurrence of any event in this component,

$$\beta = \frac{1}{lN_1} \sum_{j=1}^{l} \sum_{i=1}^{N} \gamma_{i1} x_{ij}$$

For $K \geq 2$ we first estimate all joint probabilities between pairs of evemts within the kth component:

$$p_k(j_1, j_2) = \frac{1}{N_k} \sum_{i=1}^{N} \gamma_{ik} x_{ij_1} x_{ij_2}$$

Next we reconstruct Υ_k from p_k by solving the maximum weight branch problem described in the previous section. Edges with $p_k(j_1, j_2) \leq 0.01$ are previously deleted from the complete graph in order to avoid weakly connected components within one mutagenetic tree. Finally the mixture parameters are updated by the equation

$$\alpha_k = \frac{N_k}{N} = \frac{1}{N} \sum_{i=1}^{N} \gamma_{ik}$$

We iterate the E step and the M step until the log_likelihood function does not increase any more. To run the algorithm we need initial values for the responsibilities γ_{ik}. The starting solution can be picked up at random, but in general this strategy yield poor results. The two common approaches to overcome this problem are either to sample many random starting solutions, or to identify a single promising initial solution. To limit computational costs we decided for the latter approach and perform an ordinary k-means clustering (with best starting points chosen from 100 independent random runs) with $k = K - 1$ on the set of patterns using squared Euclidean distance as dissimilarity measure [66]. From the k-means clusters assignments we derive the initial responsibilities

$$\gamma_{ik} = \begin{cases} 1/2 & \text{if sample } x_i \text{ belongs to cluster } K-1 \\ \frac{1}{2(K-1)} & \text{else} \end{cases}$$

The procedure is summarized in figure 7.3. It is differing from a true EM algorithm in the fact that the tree reconstruction step does not provide a maximum likelihood estimate. Thus, unlike with a true EM algorithm, this modified version is not guaranteed to converge to a local maximum of the log_likelihood function.

7.1.7 Model Selection Criteria

The determination of the number of tree components in the mixture model is a model selection problem. Model selection aims at identifying models that provide an accurate fit to the data, generalize well, and are no more complex than needed to explain the data. The problem can be viewed as an optimization problem involving two basic components [134]:

- a strategy for searching through the family of possible model structures efficiently
- a measure (criterion) for scoring different model structures

INPUT:

- Patterns of events $X = (x_{ij})_{\substack{1 \le \ell \le N \\ 1 \le j \le \ell}}$
- Number of mutagenetic trees $K \geq 2$

OUTPUT:

- K-mutagenetic trees mixture model $\sum_{k=1}^{K} \alpha_k \mathcal{T}_k$

PROCEDURE:

1. Guess initial responsibilities:
   (a) Run $(K-1)$-means clustering algorithm
   (b) Set responsibilities
   $$\gamma_{ik} = \begin{cases} \frac{1}{2}, & \text{if } x_i \text{ is in cluster } k-1 \\ \frac{1}{2(K-1)}, & \text{else.} \end{cases}$$
2. *M-like step.* Update model parameters:
   Set $N_k = \sum_{i=1}^{N} \gamma_{ik}$ for all $k = 1, \ldots, K$.
   Let $\mathcal{T}_1$ be a star with edge weights
   $$\beta = \frac{1}{\ell N_1} \sum_{j=1}^{\ell} \sum_{i=1}^{N} \gamma_{i1} x_{ij}.$$
   For $k = 2, \ldots, K$:
   (a) For all pairs of events (j_1, j_2), $1 \leq j_1, j_2 \leq \ell$, estimate their joint probabilities
   $$p_k(j_1, j_2) = \frac{1}{N_k} \sum_{i=1}^{N} \gamma_{ik} x_{ij_1} x_{ij_2}.$$
   (b) Compute the maximum weight branching T_k from the complete digraph with weights w derived from p_k.
   (c) Compute the mixture parameter $\alpha_k = \frac{N_k}{N}$.
3. *E step.* Compute responsibilities:
   $$\gamma_{ik} = \frac{\alpha_k L(x_i|T_k)}{\sum_{m=1}^{K} \alpha_m L(x_i|T_m)}.$$
4. Iterate steps 2 and 3 until convergence.

Figure 7.3: EM-lke algorithm for learning a K-mutagenetic trees mixture model from data

In the context of mixtures of mutagenetic trees, the model family is indexed by the number of tree components K such that searching is trivial. Here we address the second issue, namely the choice of the model selection criterion. A widely used and simple approach to model selection is cross-validation [66]. The score implicitly used by cross-validation is the estimated extra-sample performance. The motivation for this score is that we would like to fit a model which not only accounts for the training data, but also generalizes well to unseen data. In order to obtain an approximately unbiased estimate, the number of partitions used in cross-validation needs to be set large enough and this makes the procedure time-consuming. Another popular criterion for approximating the extra-sample performance is the Bayesian Information Criterion (BIC) [53]. Let M be a statistical model with parameters θ and effective number of parameters d. The effective number of parameters is also called the effective number of degrees of freedom, or the dimension, or the complexity of M. The BIC score is defined as

$$\text{BIC} = logP(D|\theta_{ML}, M) - \frac{d}{2} logN \tag{7.1}$$

where θ_{ML} denotes the maximum likelihood estimate of the parameters of the model using the training data set D. N denotes the size of the training data set. BIC implicitly implements Occam's Razor [73]: A good model is supposed to be highly likely, given the training data (the first term in 7.1) while having low complexity (the second term in 7.1). Thus, this trade-off penalizes both too complex and too simple models, and is aimed at selecting a model with intermediate complexity that is just right to account for the data set. BIC is only asymptotically consistent, i.e. it chooses the true model if it is contained in the model space as $N \to \infty$, which is often suboptimal for finite data in practice. Another difficulty with applying BIC is that identifying the effective number of parameters can be deceiving (see [117] for details).

Modified BIC We define the similarity between two tree components Υ_i and Υ_j of the mixture model by

$$S_{ij} = S_{ji} = 1 - \frac{\|A_i - A_j\|_\infty}{l} \in [0, 1] \tag{7.2}$$

where A_i and A_j denote the adjacency matrices of Υ_i and Υ_j, respectively. The matrix infinity norm is defined by $\|A\|_\infty = max_i \sum_j |a_{ij}|$, the maximum absolute row norm. Therefore, the term $\|A_i - A_j\|_\infty$ measures the maximum difference of outgoing edges between the two trees. We define the redundancy R of a mixture model as the maximum similarity among its tree components

$$R = max_{i \neq j}(S_{ij}) \tag{7.3}$$

For large K and for models containing redundant structural parts, we want to incorporate the redundancy R into our model selection criterion. For a fixed data set D and all $K \geq 1$, let Ψ_K be the K-mutagenetic trees mixture model estimated from D and let d_K be its dimension. Consider

$$\text{BIC}_R = logP(D|\theta, \Psi_K) - (1 + R) \times \frac{d_K}{2} logN \tag{7.4}$$

which doubles the penalty term for two identical tree components. We use the weighted average between standard BIC and BIC_R defined by

$$\text{BIC}_w = w \times \text{BIC} + (1 - w) \times \text{BIC}_R \tag{7.5}$$

with weight $w = min(\frac{1}{l+1} max(d_K - d_{K-1}, 0), 1)$. The idea of this weighting is that for a large increase in d (due to new model structure) we prefer standard BIC, while for a small increase in d (due to repetitive model structure) we consider BIC_R and hence penalise redundancy more heavily. The min and max terms are used to bound w between 0 and 1, because in rare cases d or R may actually decrease with increasing K.

7.1.8 First Discussion

A mutagenetic tree is a way to model a Markov chain in a compact way, and it's indeed a Bayesian net with some restrictions. Which is the advantage? First of all, when you're dealing with accumulation and dependencies of events – and maybe a lot of events – you don't have to model all the Markov states, thus obtaining a more compact (and robust) model.

The mixtures complicate the understandability, but offer some advantages. Depending on its topology, a tree can have different expressive power. A single-edge branching tree describes a unique pathways and does not allow any other kind of conditional dependencies, while a star tree is capable to describe all patterns, but in the naive-Bayes assumption. So a (linear) combination of trees can be an acceptable tradeoff between expressive power and compact representation: but one can augment the described pathways also modifying the internal structure of the trees; two trees can represent the same pathways of a single tree, but with different conditional probabilities.

7.2 Evaluation of Tree Models on Data

7.2.1 Previous Work

Mutagenetic tree mixture models have been proved to be a good model for learning HIV-1 evolutionary pathways under drug pressure. The following results are coming from [113]. Model selection was assessed through cross-validation and BIC value, while the goodness of fit was explored not only in terms of likelihood, but also generating random samples from the tree mixture models and comparing their frequency histogram. Data set was consisting of 364 instances of AZT genotypes with resistant phenotype.

Goodness of Fit A random sample from a K-mutagenetic trees mixture model $\Psi = \sum_{k=1}^{K} \alpha_k \Upsilon_k$ can be drawn by generating a uniform random number and decide according to the mixture parameters αs which mutagenetic tree to use. In the selected tree we draw each edge $e \in E$ independently with probability $p(e)$. the sample consists of all events that are reachable from r in the induced subgraph. We want to quantify how closely a trained mixture model reproduces the empirical probability distribution on $\Omega = 2^{1,\ldots,l}$. To compare two histograms $H_1, H_2 \in \mathbf{N}^{2^l}$ we use the cosine distance, defined in terms of angle spanned by the histogram vectors as

$$\mathrm{dist}(H_1, H_2) = 1 - \cos\angle(H_1, H_2) = 1 - \frac{< H_1, H_2 >}{\parallel H_1 \parallel \parallel H_2 \parallel} \in [0, 1]$$

Using cross-validation we calculate for each partition of the data into training and test set the histogram distance between test data and

- training data
- simulated data drawn from the optimal mixture model
- simulated data drawn from a single mutagenetic tree model
- simulated data drawn from a single star model with non-uniform edge weights

The three models are estimated from the training data and the size of the simulated sample equals that of the training data. The first histogram distance measures only the effect of finite sampling, whereas the other distances include losses that are due to imperfect model assumptions and/or parameter estimates. We compare the optimal mixture model with a single mutagenetic tree model and with a star model representing the null hypothesis of independence of events. Figure 7.4 shows the distribution of all distances for 100 runs of 10-fold-cross-validation each. Histograms generated from the estimated mixture model closely resembled the observed data. In contrast, both the single tree model and the independence assumption provide inferior model fits.
In the same cross-validation runs we have determined the percentage of samples that remain unexplained by the non-trivial components of the mixture model. The mean percentage of samples with likelihood zero in all but the noise component was 13%. Thus, the mixture model maps 87% of the observed patterns onto the identified mutagenetic trees. For the optimal model on the full data it happens to be the case that the only pattern that can be generated by both non-trivial trees is the null pattern. The following table reports the distribution of samples among trees in detail.

$L(x\|\Upsilon_1)$	$L(x\|\Upsilon_2)$	$L(x\|\Upsilon_3)$	fraction	description
> 0	> 0	> 0	31.6%	null patterns
> 0	> 0	$= 0$	30.2%	70-219 pathway
> 0	$= 0$	> 0	25.0%	215-41 pathway
> 0	$= 0$	$= 0$	13.2%	noise

7.2.2 Mutagenetic Tree Mixture Model Validation on Large Clinical Data Sets

The investigation will be focused on comparing different models of evolution using a large training set obtained from the German AREVIR data base [29] and a large test set coming from the Italian ARCA retrospective cohort [102].
We selected 9 different drugs, according to large test sets available ($\geq$ 395 instances, which was the smallest size, belonging to LPV set): genotypes were selected simply collecting all the sequence made under each treatment.
Various models were tested:

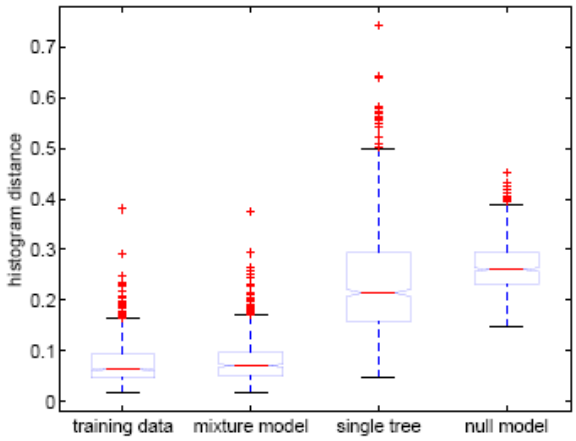

Figure 7.4: Box-plot of histogram distances from 100 10-fold cross-validation runs. From left to right: distances between histograms of test data and training data (first box), simulated data from the optimal mixture model (second box), simulated data from the single tree model (third box), simulated data from the null model (fourth box). Simulated data are drawn from the models that have been fitted on the training data. The null model refers to the independence assumption of events and is a single star with non-uniform edge weights

- the mutagenetic tree mixture
- a single mutagenetic tree
- a star model (which assumes independence of events)
- an accumulation model (which does not takes into account the singular events, but only their accumulation in unspecified order)

Tree Selection and Topology The first thing that we can notice, before assessing performances, is that the modified BIC criterion for the optimal number of trees selection does not work well: for instance, for 3TC drug (see figure 7.8) a set of three identical trees was selected, in despite of the penalty for tree similarity. Could it be due to the large training set size?
Another surprising thing is that the pathways for AZT are different from the ones discovered formerly using instances obtained from genotype/phenotype resistant pairs (in-vitro tests, compare figures 7.5 and 7.2, M41L and K70R have exchanged places). A possible explanation for this can be the fact that the examples extracted from in-vivo data are under combined treatments and are not discriminated by a resistance cut-off value (in fact we found some wild type genotypes using this selection criterion).
The trees obtained for the protease inhibitors (we show here the mixture model

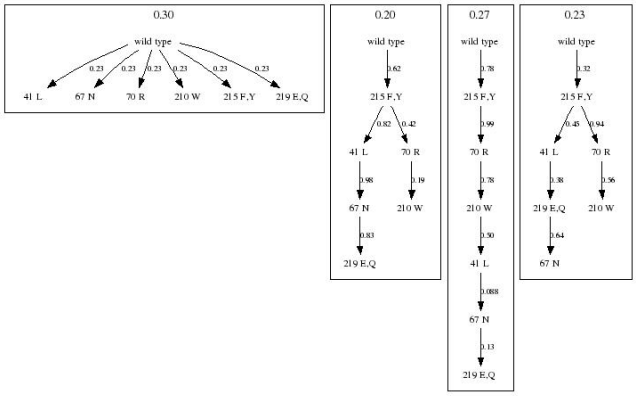

Figure 7.5: mutagenetic tree mixture model for AZT

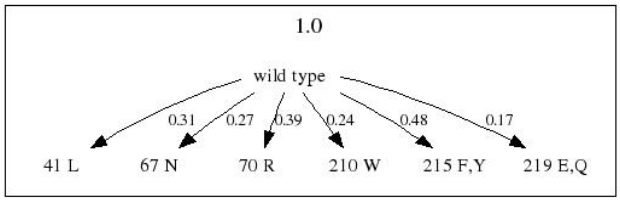

Figure 7.6: independence model for AZT

obtained for IDV in figure 7.10) are not showing defined pathways, they resemble much more a star model.

Likelihoods and Responsibilities We collected results for training set data in table 7.1 and for test data in table 7.2.

The log_likelihood (and responsibilities in the mixture model or percentage of explained pathways in the single tree model) results obtained from the test data can be used to assess significant differences among models and among drugs. However, the choice of a suitable statistical test is not so trivial. We have a vector of log_likelihoods (calculated on the samples) for each model and for each drug: fixing the attention to one drug, clearly the distribution of log_likelihoods is not gaussian, so a non-parametric test must be used. A Wilcoxon-signed-rank

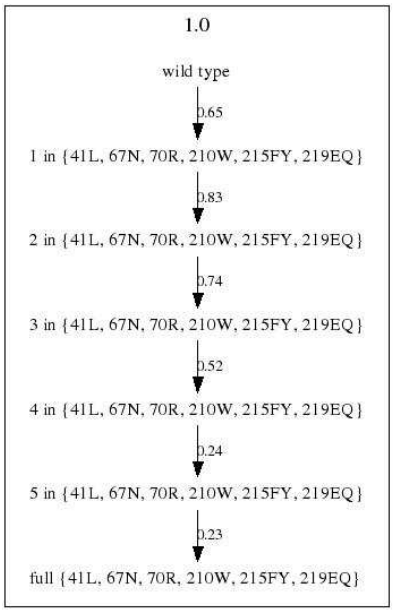

Figure 7.7: accumulation model for AZT

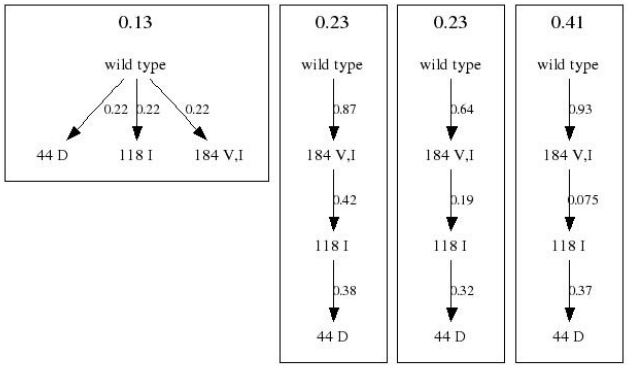

Figure 7.8: mutagenetic tree mixture model for 3TC

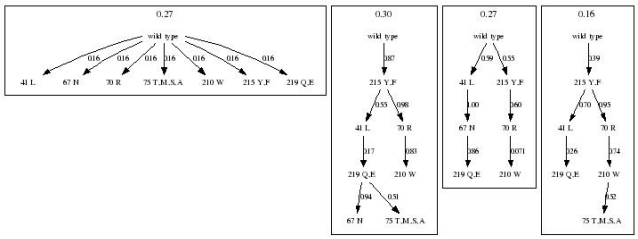

Figure 7.9: mutagenetic tree mixture model for D4T

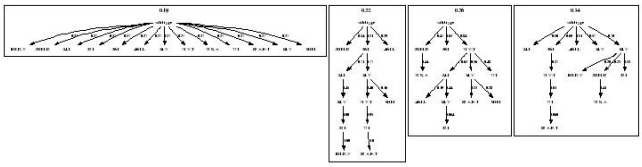

Figure 7.10: mutagenetic tree mixture model for IDV

Table 7.1: AREVIR data base (training set). Maximum K was set to 4, modified BIC was used to select optimal K, a tolerance value of 0.05 was set for the model fit.

			mixture model				oneTree	indep.	accum.
drug	n=	E	K	loglike	mod.BIC	resp.	resp.	loglike	loglike
D4T	2109	8	4	-6301.58	-6400.81	0.15	0.26	-8357.4	-3775.5
AZT	1629	7	4	-4488.60	-4566.43	0.21	0.21	-5831.4	-2727
3TC	2845	4	4	-3517.07	-3556.84	0.05	0.05	-3632.2	-2953.2
DDI	1177	9	4	-4381.76	-4470.14	0.13	0.27	-5168.6	-2101.5
IDV	770	14	4	-4629.60	-4770.60	0.14	0.35	-5144.3	-1628.5
NVP	663	8	2	-1743.74	-1773.73	0.07	0.03	-1716.8	-759.01
SQV	664	10	4	-3143.99	-3247.09	0.13	0.22	-3517.5	-1320.8
LPV	317	17	4	-2164.86	-2327.46	0.26	0.34	-2482.2	-603.81
NFV	1089	10	4	-5035.96	-5153.44	0.13	0.20	-5540.7	-1919.1

Table 7.2: ARCA data base (test set)

		mixture model		oneTree	independ.	accumul.
drug	n=	loglike	resp.	resp.	loglike	loglike
D4T	2122	-10687.00	0.47	0.51	-8424.80	-3674.00
AZT	2434	-7736.80	0.36	0.36	-7760.90	-3746.80
3TC	2690	-3549.60	0.04	0.04	-3723.90	-3065.80
DDI	1255	-5422.80	0.25	0.39	-5455.30	-2245.30
IDV	817	-5311.60	0.58	0.61	-5822.10	-1737.30
NVP	596	-1515.16	0.07	0.03	-1517.80	-731.84
SQV	873	-3589.10	0.33	0.33	-4201.00	-1706.30
LPV	395	-4038.30	0.83	0.83	-4473.30	-1318.10
NFV	662	-3848.90	0.49	0.50	-3497.70	-731.84

test for two groups (or a simpler sign test, that does not requires simmetry or a Kruskal-Wallis for multiple comparisons) seems appropriate, but relies on median comparisons. Judging a log_likelihood vector "better" than another one involves some arbitrary steps: a model can fit better a fraction of the data, while the other can be better in average.

The sign test was made comparing -for each drug- the log_likelihood vectors obtained from the test data: the mixture model was tested against the independent and the accumulation model. All results were significant ($p< 10^{-3}$)*. Figures 7.11, 7.12 and 7.13 show the log_likelihood distributions in the test set for the three different models regarding AZT drug, while figures 7.14 and 7.15 plot the differences between the mixture model log_likelihoods and the star and accumulation model respectively. Finally, figures 7.16 and 7.17 plot mixture model likelihoods versus star and accumulation models (jittered with some uniform random noise). Again, for the remaining drugs we plot mixture vs independent and mixture vs accumulation likelihoods (figures 7.18 7.19 7.20 7.21 7.22 7.23 7.24 7.25 7.26 7.27)

The accumulation model outperforms both models: this can be useful if we need to calculate a time-indicator of resistance developing, for which just the number of accumulated mutations is important, but the evolutionary pathways are lost.

Now we can think about the comparison of models in the general framework. A model has to perform good in each drug scenario: so, we can use the log_likelihood sum (maybe normalized, having different sample set sizes) as indicator value and a drug as a random sample. The normal distribution again does not hold, and again a Wilcoxon test can be appropriated, because anyway the models should at least behave coherently among different drugs. In the previous work ([113]) mixture

*except for AZT (mixture vs accumulation p=0.6827), LPV (mixture vs star p=0.02722) and NVP (mixture vs star p=0.5131)

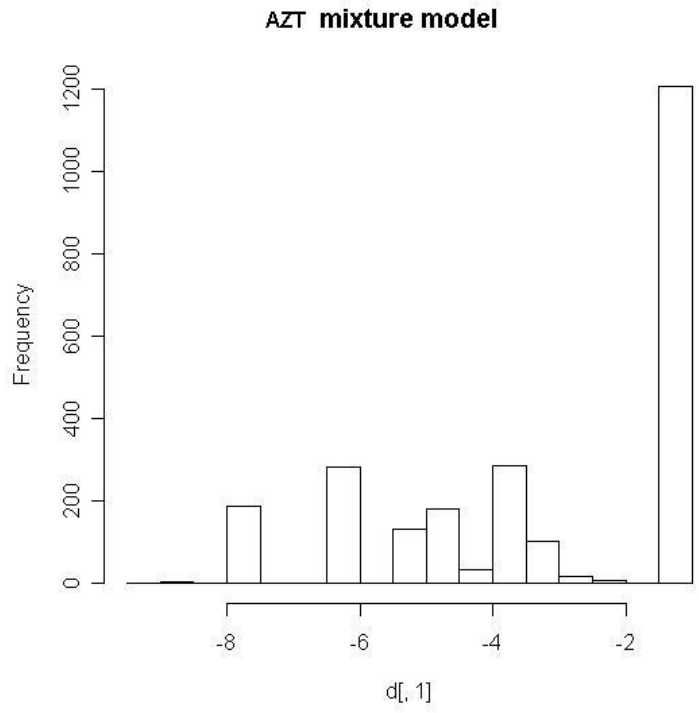

Figure 7.11: AZT: Likelihood distribution for mixture model

model was proven to be better than single tree and independent model through cross-validation: however, the test was made only for AZT. Surprisingly, with AREVIR and ARCA data sets we obtained different results. Even though the sign test gave -almost for all drugs- significant differences between the models, the Wilcoxon test made on the overall drug scenario was not significant comparing the mixture model with the star, while it was (though not extreme) with the accumulation (see table 7.3).

In terms of responsibilities, no significant differences were found between the percentage of samples explained by the noise component in the mixture model and the percentage of unexplained patterns in the single tree model, either in the training set or in the test set (p=0.12 and p=0.7997). The responsibilities due to the noise component however increase from the training set to the test set (except for 3TC, NVP, and less in SQV), while the mean log_likelihood remains more stable for all models.

7.2.3 Conclusions

Still two issues have to be investigated: first, the effect of combination of treatments in the tree estimation and in the genetic barrier computation; secondly, the set of mutations (events) to be included in the trees. For the first problem,

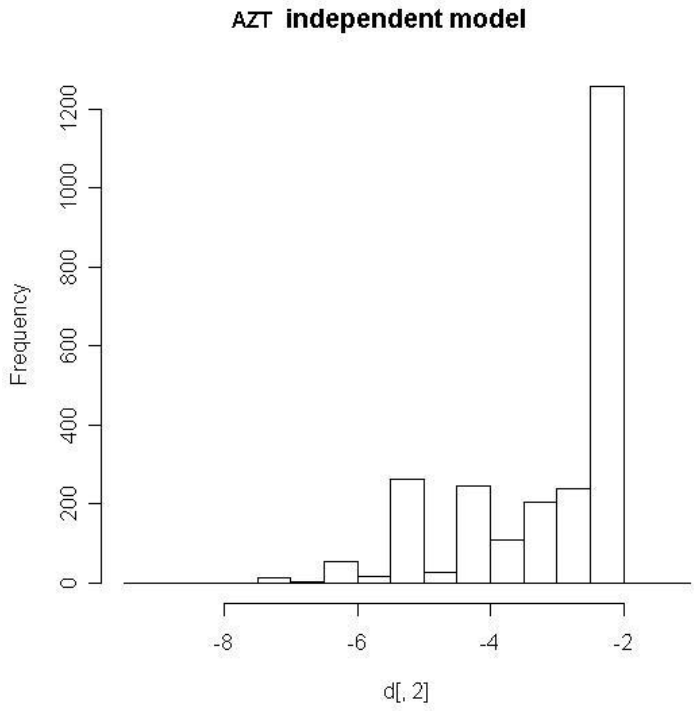

Figure 7.12: AZT: Likelihood distribution for independence (star) model

one important correlated issue is the selection criterion of genotypic sequences for each drug. Actually it's not completely true that every genotype made under a treatment is a resistant genotype, and this could bias a bit our analyses: among the several reasons, one can be the un-adherence or the bad adsorption; for instances -the worst case- test set for SQV contained 103 wild type patterns out of 873, while other drugs assessed to very low percentages. A better choice would be to select only genotypes with resistant phenotype, but again we could end up with less data and assumption that in-vitro resistance is in-vivo resistance. Moreover, the nuisance effects due to combined treatments can be significant as well: as we pointed out in the multivariate analysis, different drug combinations can lead to different evolutionary pathways compared to the monotherapy. So, why do we estimate single-drug trees? A natural solution should be to train models using different sets and choose the best validation results (so, single-drugs, drug-combinations, phenotypic-resistant.. but test set should be comparable with a common indicator).
The second problem is also crucial: the event set for each drug, in fact, is made by resistance associated mutations approved by the IAS/USA [121]. We know that there are these novel mutations that participate to the resistance development, then they should be included in the event set (as also discrimination between dif-

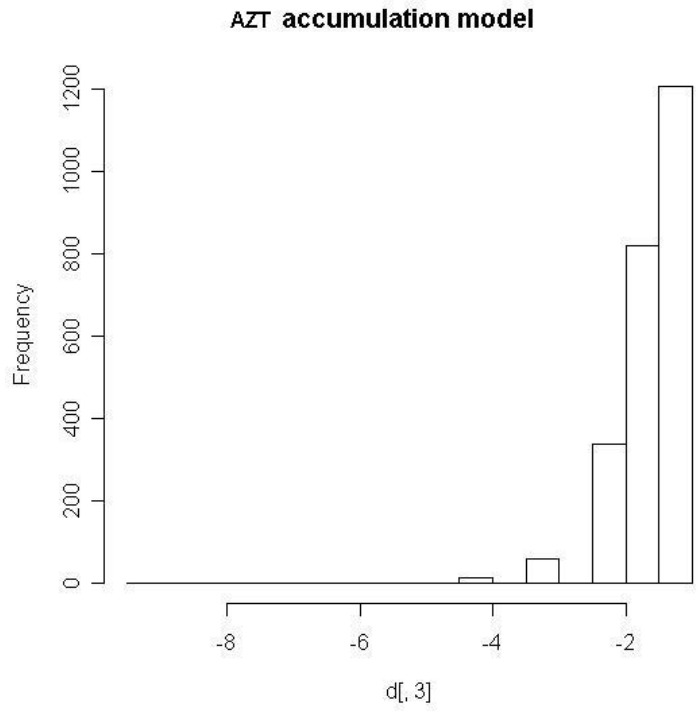

Figure 7.13: AZT: Likelihood distribution for accumulation model

ferent amino acidic substitutions, as RT_215_Y and RT_215_F that are in TAM1 and TAM2 respectively). Can a large number of events create problems in the tree building process? How about a heuristic event selection pre-processing? Finally, under this large test sets the mixture model seemed not to be significantly better than the independent model -though almost for all drugs a small fraction of the data was always better explained- and more surprisingly the responsibilities of the mixture model due to the noise component were comparable to the fraction of data unexplained by a single tree model. The accumulation model still should be an option, considering also the fact that the increasing number of mutations has always been proven significantly correlated with the resistance in clinical studies.

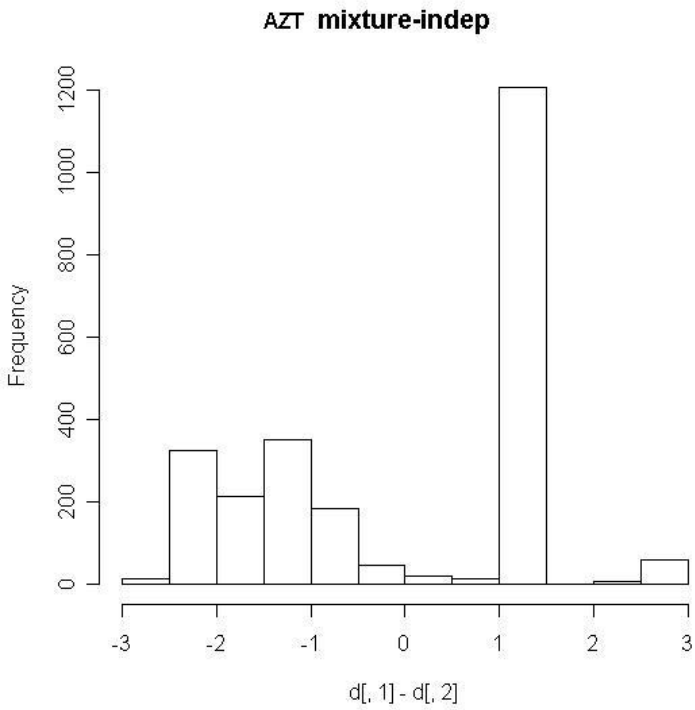

Figure 7.14: AZT: differences between mixture model and independence model

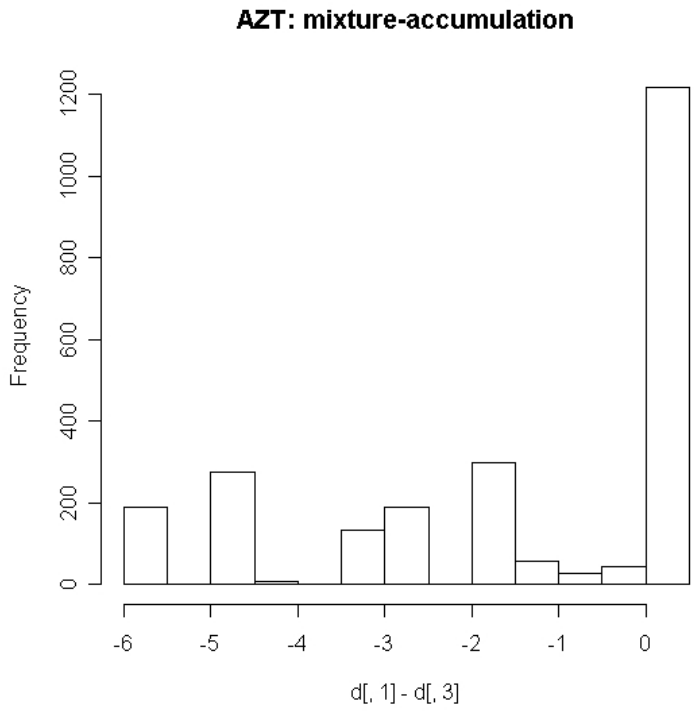

Figure 7.15: AZT: differences between mixture model and accumulation model

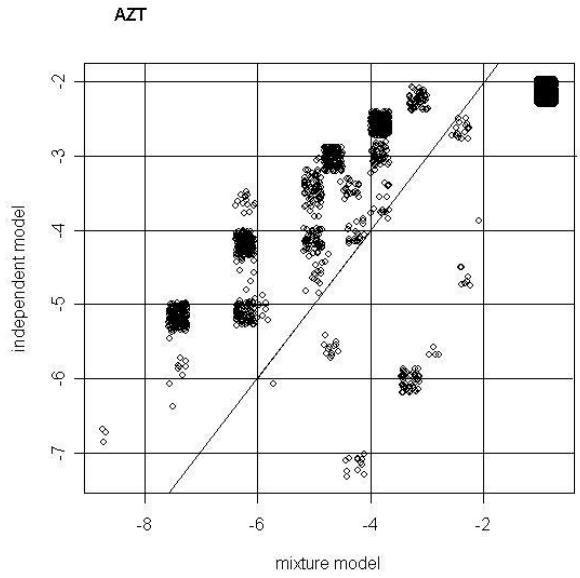

Figure 7.16: AZT: scatterplot of mixture model vs independence model likelihood

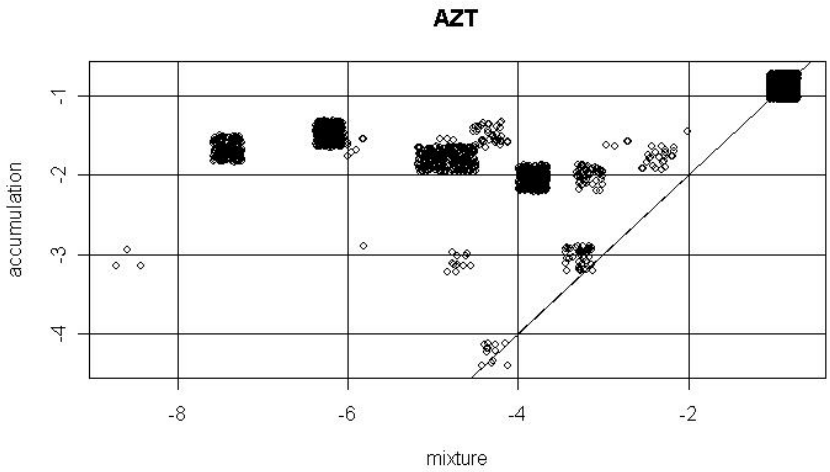

Figure 7.17: AZT: scatterplot of mixture model vs accumulation model likelihood

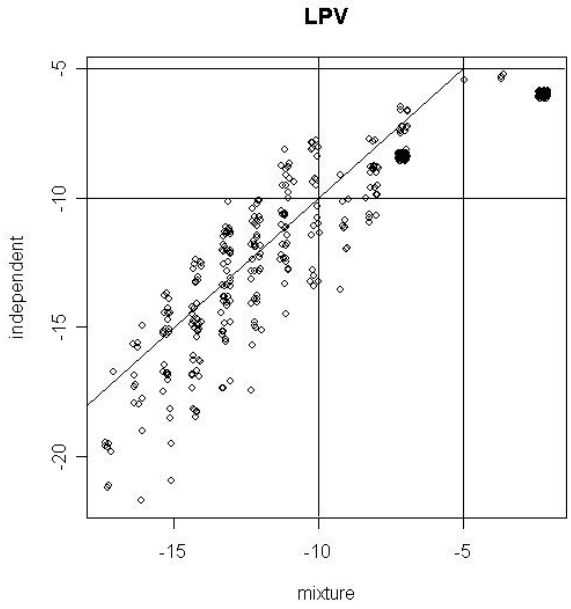

Figure 7.18: LPV: scatterplot of mixture model vs star model likelihood

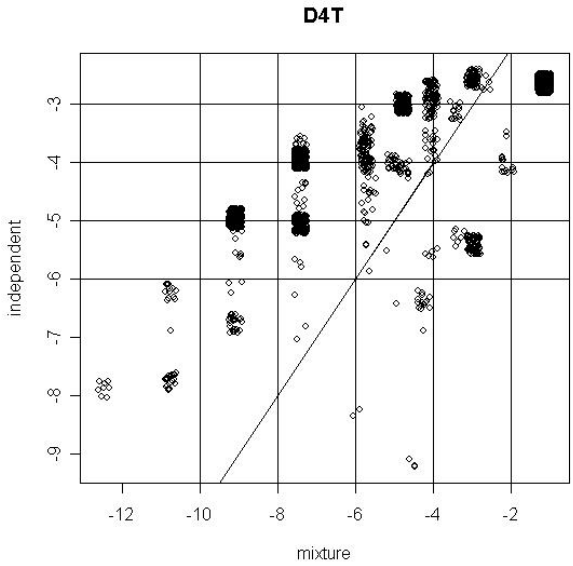

Figure 7.19: D4T: scatterplot of mixture model vs star model likelihood

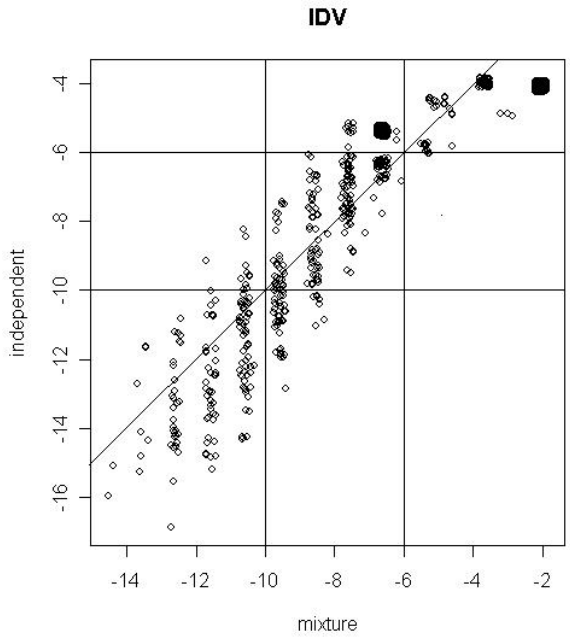

Figure 7.20: IDV: scatterplot of mixture model vs star model likelihood

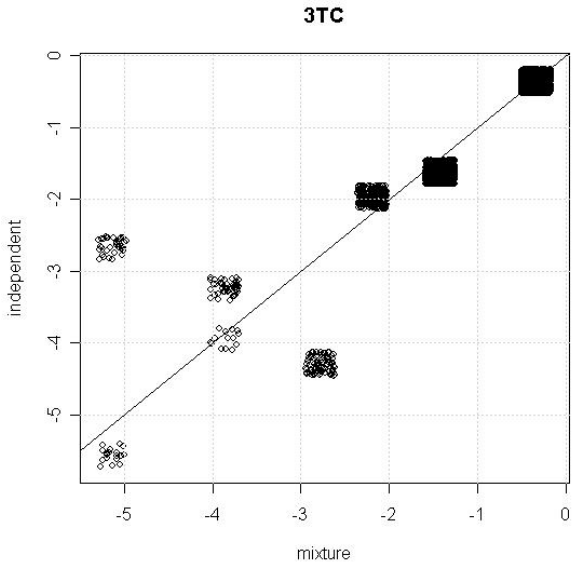

Figure 7.21: 3TC: scatterplot of mixture model vs star model likelihood

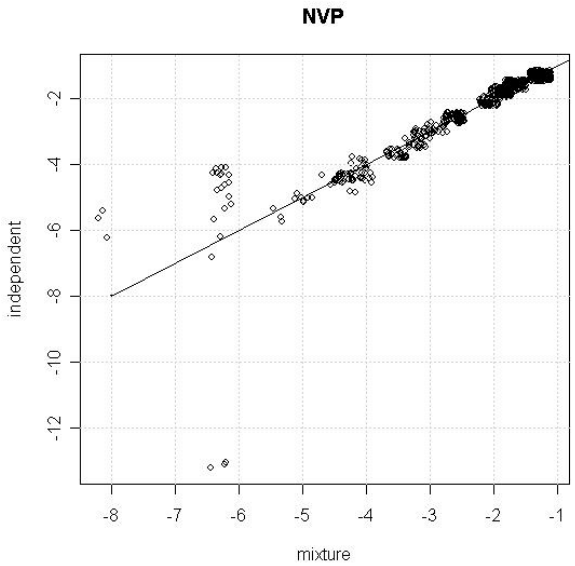

Figure 7.22: NVP: scatterplot of mixture model vs star model likelihood

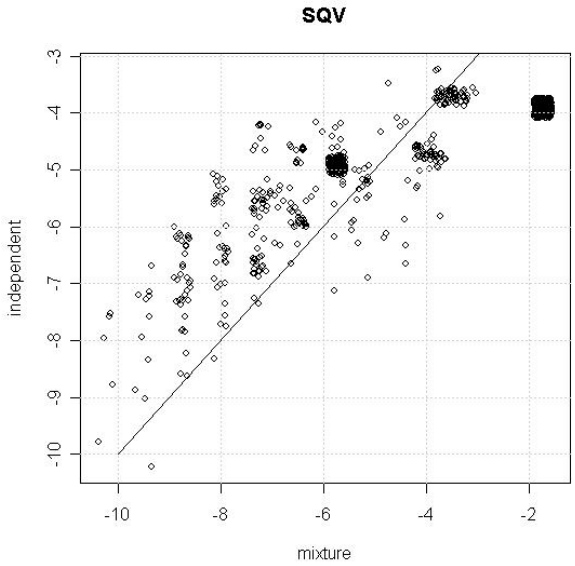

Figure 7.23: SQV: scatterplot of mixture model vs star model likelihood

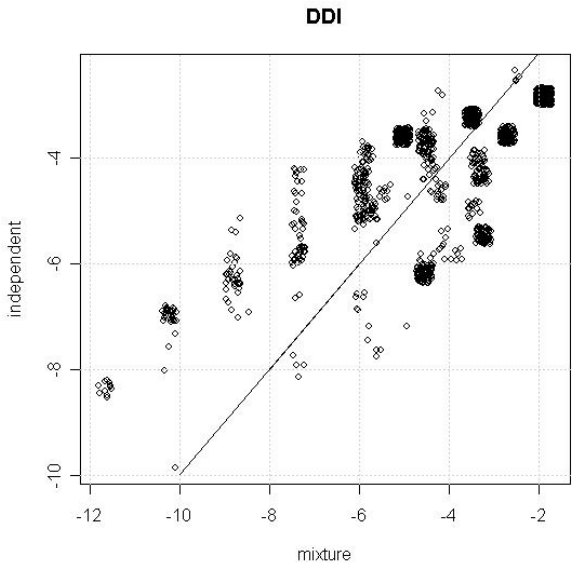

Figure 7.24: DDI: scatterplot of mixture model vs star model likelihood

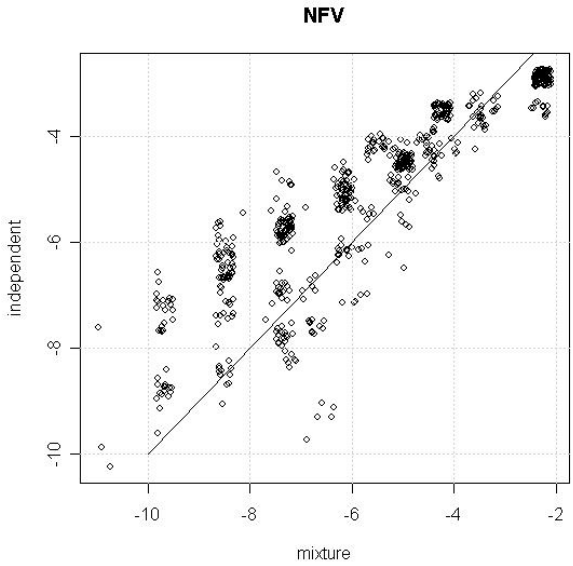

Figure 7.25: NFV: scatterplot of mixture model vs star model likelihood

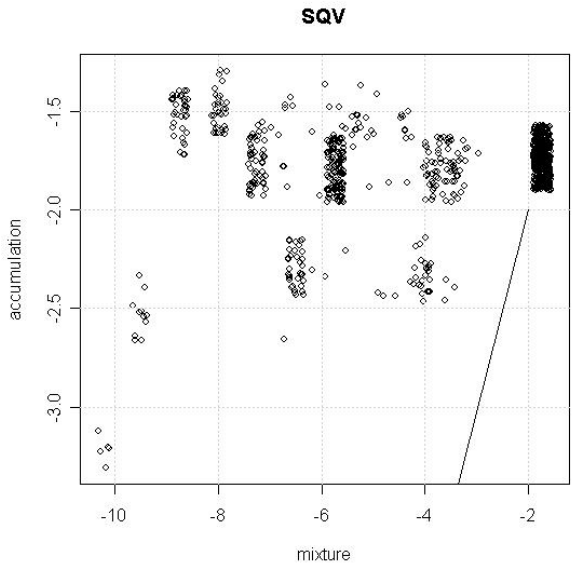

Figure 7.26: SQV: scatterplot of mixture model vs accumulation model likelihood

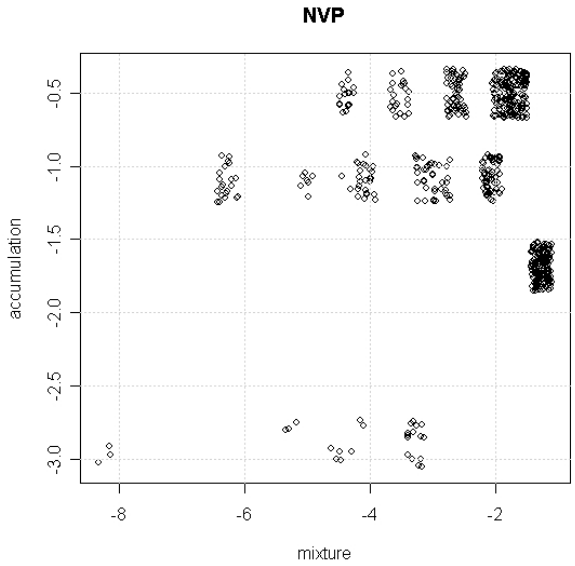

Figure 7.27: NVP: scatterplot of mixture model vs accumulation model likelihood

	mixture vs star	mixture vs accumulation
training set	p=0.6048	p=0.001851
test set	p=0.9314	p=0.002756

Table 7.3: Wilcoxon test for model comparison

Part III

Statistical Learning for HIV-1

Chapter 8

Overview and Methods

In this part we'll explore the application and comparison of several machine learning techniques applied to different classification and regression problems regarding HIV-1. In this first chapter we'll give a quick review of the methods used in general (that will be specified again case by case in the different sections). The aim of this introduction is not to give deep theoretical explanation of the methods (for which proper references will be given), but to show the general methodologies and frameworks that can be applied in data mining scenarios relative to biology and medicine fields: a major problem when facing data mining in fact is not only the model choice and performance assessment, but also – and mainly – the input domain coding, the feature selection and extraction, the model selection criteria.

The following chapters will describe the applications of such methods for the development of expert systems on HIV-1 classification/regression problems:

- Prediction of in-vitro phenotypic resistance to drugs from viral genotypic sequence
- Prediction of viral tropism (coreceptor usage) from viral genotypic sequence and clinical attributes
- Prediction (classification and regression) of clinical outcomes for in-vivo antiretroviral therapeutic combinations from clinical, historical and viral attribute domains

The problems are closely related, either for the domain specifications or for the input-output scenario. In fact, the main issue is the building of in-silico prediction models that use clinical and genomic can mimic in-vitro experiments (often much more expensive to be carried on than a simple viral sequencing) or can help in taking clinical decisions for in-vivo therapy design.

8.1 Machine Learners

The following methods were extensively investigated. We won't describe the theoretical aspects of each technique, since the purpose of this thesis is not to review, but to apply and compare statistically them, in conjunction with feature analysis and robust validation, on the biology-medicine data scenario. We assume that the reader is confident with the methods, giving appropriate references.

- Multiple Linear Regression (MLR)
- Logistic Regression (LR) [106] [35]
- (Bagged) Decision Trees, Random Forests (DT, RF) [23] [77] [74] [76]
- Support Vector Machines (SVM) [67]
- Rule Based Methods (RB) [132]
- Instance Based Reasoning (IBR) [7]

MLR and LR have the advantage to be yet easy to interpret models and the possibility to assess statistical significance of variables, along with the suitability to be enhanced with several feature selection techniques. DT are understandable models and can explore non-linear dynamics, but they suffer of poor predictive ability, thus have to be enhanced with bagging techniques or within the RF framework. SVM are among the best models regarding optimisation, moreover are extremely flexible when using different kernels: they've been criticised for poor interpretability, though some considerations can be done regarding variable importance analysing support vector weights. RB probably are the models that are more similar to human reasoning: there are several rule finding policies, but so far their power can be considered similar to DT.

Maybe surprisingly, we avoided the usage of Neural Networks: a motivation raises from the fact that the parameter optimisation (in the sense of activation functions, hidden layer topology, etc) is often cumbersome, the training algorithms have to be tuned properly in terms of epochs, the over training needs to be faced with care (maybe with weight decay).

Regarding the implementations (for example we'll use C4.5 for DT and Platt's algorithm for SVM [67]), we will always use the procedures given in the references.

8.2 Loss Functions and Validation

8.2.1 Loss Functions

Along with the validation techniques, different loss functions are available for classification problems in literature [66]. In the different scenarios, we'll use losses that can catch different nuances:

1. accuracy ($\frac{TP+TN}{TP+TN+FP+FN}$)
2. specificity & sensitivity ($\frac{TN}{TN+FP}$, $\frac{TP}{TP+FN}$)
3. AUC of ROC curve
4. f-measure ($\frac{2 \cdot PPV \cdot sensitivity}{PPV+sensitivity}$)

where TP, TN, FP, FN stand for True Positives, True Negatives, False Positives, False Negatives respectively. PPV is the Positive Predicted Value and is equal to $\frac{TP}{TP+FP}$. Depending on the problem settings, one function will be preferred to the others: for example, in the tropism prediction the sensitivity towards the X4 detection will take a major role.

For regression, correlation and RMSE will be used.

8.2.2 Validation

The main issue in validating a model is to derive an error estimation for the entire input space, i.e. a general performance measure. For this reason the error calculated for the training samples is not a reliable indicator. There are several statistical procedures designed for the achievement of a robust prediction error indicator, which can be divided into two main families:

- In-Sample Prediction Error Estimate
  - Akaike Information Criterion AIC [58]
  - Bayesian Information Criterion BIC [53]
  - Minimum Description Length MDL [66]
- Extra-Sample Prediction Error Estimate [66]
  - Test Set Error
  - k-fold Cross Validation (CV)
  - Bootstrap

In-Sample Prediction Error Estimates

The usage of adjusted measures on the training error (as in-sample prediction error estimate) is not only useful for optimising a more robust model, but can be used also as a *quick* indicator of the model performances for unseen data: although these indicators are demonstrated to be inferior to the error estimation coming from k-fold CV, they are computationally inexpensive. The most famous error measure that can be directly calculated from the training error is the Akaike Information Criterion.

Akaike Information Criterion The Akaike Information Criterion (AIC) is a statistical model fit measure. It quantifies the relative goodness of fit of various previously derived statistical models, given a sample of data. It uses a rigorous framework of information analysis based on the concept of entropy. The driving idea behind the AIC is to examine the complexity of the model together with goodness of its fit to the sample data, and to produce a measure which balances between the two. The formula is

$$AIC = 2k - 2ln(L) \tag{8.1}$$

where k is the number of parameters, and L is the likelihood function. When errors are assumed to be normally distributed, AIC is computed as $AIC = 2k + n \cdot ln(RSS/n)$, where n is the number of observations and RSS is the residual sum of squares. A model with many parameters will provide a very good fit to the data, but will have few degrees of freedom and be of limited utility. This balanced approach discourages overfitting. The preferred model is that with the lowest AIC value.

The Bayesian Information Criterion is another indicator, introduced by Schwarz in 1978 [53], which definition was already given in section 7.1.7.

Performance Assessment, Model Comparison and Selection

The state of the art for error evaluation, independently from the loss functions, is to execute the x-fold cross validation or the bootstrap [66], when a large test set is unavailable. These procedures however cannot tell about the distribution of the error estimation and about comparisons of different models. One solution is to execute multiple independent runs of x-fold cross validation, obtaining a normal distribution under mild conditions. In this way, two different models can be evaluated and compared (each against the other or each against a null hypothesis) using a t-statistic. The usage of a naive Student's t-test can be however biased due to the sample overlap: an adjusted Student's t-test was proposed by Bengio [133] and holds for multiple cross validation runs:

$$adj.t = \frac{\frac{1}{kr}\sum_k \sum_r x_{ij}}{\sqrt{\left(\frac{1}{kr} + \frac{n_2}{n_1}\right)\sigma}} \tag{8.2}$$

where k, r are the number of folds and runs respectively, n_1 is the training set size, n_2 is the test set size, $\sigma = \frac{1}{kr}\sum_k \sum_r (x_{ij} - m)$ is the standard error, $x_{ij} = a_{ij} - b_{ij}$ is the error difference between the two distributions, $m = \frac{1}{kr}\sum_k \sum_r x_{ij}$ and the degrees of freedom are $df = kr - 1$.

8.3 Feature Selection

The loss functions above described tend to minimise the error or maximise the correlation between observed and predicted vectors. They do not take account for the number of parameters used. In usual engineering scenarios the parameters to be optimized are few and related to significant variables as position, speed, acceleration: in a biological framework instead there is a huge number of variables (for example all the mutations in the viral genotype) and a corresponding large parameter space. Many of the input variables can be non-significant for the model and many parameters mean that the system can be easily overfitted.

Feature Selection is closely related to the Occam's principle, for which models that use a minor number of parameters are preferred under the same prediction performances. This is useful when dealing with high-dimensional data sets, where many input attributes could be irrelevant and redundant to the dependant variables and act just as a noise. By allowing learning algorithms to focus only on highly predictive variables, their accuracy can be even improved. Feature selection – following the definition by Weston et al. [123] – is the problem to find the feature subset of a certain size that leads to the largest possible generalisation or equivalently to minimal risk. In literature the feature selection techniques have been grouped in three methodologies [56]:

- *Filter*
- *Wrapper*
- *Embedded*

Filter methods independently rank and select relevant features using different criteria, regardless the prediction method, thus they do not incorporate learning; under certain independence or orthogonality assumptions, they can be optimal with respect to a given predictor. Wrapper methods assess subsets of variables according to their usefulness to a given predictor: these methods use a learning algorithm to measure the quality of subsets of features without incorporating knowledge about the specific structure of the prediction function. In contrast to filter and wrapper approaches, in embedded methods the learning part and the feature selection cannot be separated.

8.3.1 Filters

- Variable Importance Ranking
- Correlation-Based Feature Selection

The variable importance ranking can be done via any statistical test made with respect to the class or output attribute. Usually univariable tests are carried on

and ranked by their statistical significance. Depending on the variable types, non-parametric or parametric (Wilcoxon, Mann-Withney, Student's, Chi-Square, et cetera) tests can be used, also taking into account for multiple comparisons. The execution of an univariable test however cannot tell much about variable inter-dependencies, confounding terms or higher-order interactions. Stratification can be a way to handle this, or the usage of multivariable techniques.

A way to handle the inter-correlation between variables and select subset of relevant features, yet under the independence assumption, is the Correlation-based Feature Selection method CFS [83]. At the heart of the CFS algorithm is a heuristic for evaluating the worth or merit of a subset of features. This heuristic takes into account the usefulness of individual features for predicting the class label along with the level of inter-correlation among them. The hypothesis on which the heuristic is based can be stated as

```
Good feature subsets contain features highly correlated with
the class, yet uncorrelated with each other
```

In CFS a merit function is defined:

$$Merit_S = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \tag{8.3}$$

where $Merit_S$ is the heuristic merit of a feature subset S containing k features, r_{cf} is the mean feature class correlation $f \in S$ and r_{ff} is the average feature-feature inter-correlation. The numerator can be thought of as giving an indication of how predictive of the class a group of features are; the denominator of how much redundancy there is among them. The heuristic handles irrelevant features as they will be poor predictors of the class. Redundant attributes are discriminated against as they will be highly correlated with one or more of the other features. In order to have a common criterion, all the numeric features are discretised and *entropy* plus *information gain* measures are used.

The purpose of feature selection is to decide which of the initial features to include in the final subset and which to ignore. If there are n possible features initially then there are 2^n possible subsets. The only way to find the best subset would be to try them all; this is clearly prohibitive for all but a small number of initial features. Various heuristic search strategies are often applied to search the feature subset space in reasonable time (we describe them in the Wrapper section). Standard implementation of CFS starts from the empty set of features and uses a forward best first search with a stopping criterion of five consecutive fully expanded non-improving subsets.

CFS assumes that features are conditionally independent given the class. Experiments show that CFS can identify relevant features when moderate feature dependencies exist. However, when features depend strongly on others given the class, CFS can fail to select all the relevant features.

8.3.2 Embedded Methods

Overfitting during the training phase can be major problem, especially when in presence of a large input attribute space. Different Machine Learners can be more or less prone to overfitting, depending on their internal optimisation mechanisms. Random Forests and Decision Trees, for instance, can perform feature selection in the training phase (see the next chapter). Neural Networks, on the other hand, do not have any criterion: for this reason mechanisms such as *weight decay* have been proposed (see [66]).

- Decision Trees: Decision Trees can be viewed as embedded methods for feature selection, in the sense that they are iteratively built by splitting the data depending on the value of a specific feature: in most of cases only a subset of input features will be included in the tree. The split points are chosen according to a specific criterion used to evaluate the feature importance regarding the classification: among the criteria, *mutual information*, *cross-entropy*, *gini index* are the most used. Well known tree building methods in literature are CART [23] and C4.5 [74]. We will also use the LogitBoost implementation for Logistic Model Trees [35].

- Ridge Regression: see [106]

- Fuzzy Criteria: The idea for a Fuzzy criterion (developed by the author) rises from the AIC definition and the penalty term of Ridge Regression: it's re-interpreted in fuzzy terms. While the AIC formula is fixed and selects variables only based on their statistical significance and the penalty term in Ridge Regression penalises the weights, families of parameterized fuzzy functions that take into account the number of parameters and the loss function can be designed, in order to decide with more flexibility how much the model has to be simple (i.e. how many parameters are included) joined with its goodness of fit. This can be viewed as an embedded method because the parameter selection is performed in the same phase of parameter optimisation during the training process. For details, see [87].

8.3.3 Wrappers

As introduced, wrapper methods consist in using the prediction performances of a given learning machine to assess the relative usefulness of variable subsets. In practice three steps have to be followed:

- Search space of the variable subsets

- Loss and Risk functions to assess the prediction performances of a learning method in order to guide and evaluate the subset search

- Machine Learning Models to be used

Exhaustive search on the power set of the input attributes is feasible only if the number of variables is not too large (the problem is known to be NP-hard). Greedy-heuristic searches are used in practice: best-first, branch and bound, genetic algorithms, simulated annealing, hill climbing, random searches. Among these cited, *forward* and *backward selection* (known as *nested subsets* methods) are fast and easy to implement greedy strategies. Forward selection iteratively adds variables if the are proven to improve performances under a certain risk indicator, while backward removes variables if they result not in affecting the predictor's error. It has been showed that simple (single variable addition) forward selection however – for non-linear cases – can be problematic when it's true the case that two variables are not significant alone, but together they are.

8.3.4 Heuristic Functions and Optimisation Algorithms

Simple optimization algorithms are often limited to regular convex functions. Actually, most real problems lead to face multi-modal, discontinuous, non-differentiable functions. To optimise such functions traditional research techniques use gradient-based algorithms [34], while new approaches rely on stochastic mechanisms: these latter base the search of the next point basing on stochastic decision rules, rather than deterministic processes, requiring then weak conditions for the objective functions. Genetic Algorithms, Simulated Annealing and Random Searches (see again [34]) are among these, and often are used either when the problems are difficult to be defined, either when "comfortable" properties – such as differentiability or continuity – are missing. For the most known heuristics in literature, we can give the following references

- Forward or Backward subset selection [66]
- (Fuzzy) Genetic Algorithms, developed by the author [87]
- Simulated Annealing [19]
- Random Searches [34]
- Bubble Selection, developed by the author [81]

Bubble Selection

Since the Bubble Selection will be used in a Wrapper FS in conjunction with IBR for in-vivo therapy optimisation, we describe the algorithm in detail in table 1: Bubble Selection it is a derivative-free quasi-random search that considers size-varying feature sets and accomplishes the Occam's principle. In Bubble Selection elements in the power set of the attributes are heuristically evaluated, preferring

low-cardinality subsets. Since not only one variable is added at each iteration, the independence assumption is relaxed.

Algorithm 1: Bubble Selection Algorithm

available feature set C_{tot} of size K, learning model M
current feature set $C = \emptyset$
best performance indicator $p_B = p(C, M)$

while *stop criterion on desired performance is not reached* **do**
/* growth and shrinkage phases:
· generate positive random number r (through a discrete generation function from a normal distribution $d = ceiling(abs(\mathcal{N}(0, \sigma)))$, where σ is tuned according to desired variance of the number of features to be added)
· add r random features to the current set C, but if a feature to be added is already present, remove it from with probability p */
r=ceiling(abs(NormRand(0,σ)))
foreach *int* $i = 1$ *to* r **do**
j = new random uniform $\in \{1, \ldots, K\}$
if $c_j \in C$ **then**
u = new random uniform $\in (0, 1)$
if $u < p$ **then**
$C_{new} = C - \{c_j\}$
end
else
$C_{new} = C \cup \{c_j\}$
end
end
/* evaluate the performance of the current feature set $p(C_{new}, M)$ (using a robust in-sample or extra-sample performance estimator) and compare with p_B (with a statistical test) */
if $p(C_{new}, M)$ *is better than* p_B **then**
$C = C_{new}$
$p_B = p(C_{new}, M)$
end
end

8.3.5 Feature Extraction and Generation

In this section we cite some additional procedures for Feature Analysis: apart from PCA, HC and PC, used in chapter 5, the others were not applied in the experimental settings, but for completeness we group them together with the proper references.

Feature Extraction can be defined as any transformation applied to the original variable state with the double objective to reduce dimensionality (as for

Feature Selection) and to extract relevant characteristics. The difference with Feature Selection is that the input space is transformed in another space (of less dimensionality). There are currently several procedures that extract relevant characteristics from the input space: most of them can be viewed as a preliminary unsupervised learning. So far, we can cite:

- Principal Component Analysis PCA [15], Independent Component Analysis ICA [24] [66]
- Factor Analysis, Multiple Correspondence Analysis MCA [79]
- Wavelet Smoothing [66]
- Linear Vector Quantization [66]
- Clustering
  - Partitional Clustering PC (k-means, k-medoids, fuzzy c-means, fuzzy subtractive, gaussian mixture models GMM) [50]
  - Hierarchical Clustering HC (agglomerative, divisive) [68] [97]

Of note, in this chapter we did not reviewed any bayesian or maximum-likelihood method concerning the feature generation, which however deserves a citation [21] (in literature it is also known as semi-supervised learning).

Chapter 9

In-Vitro: Viral Tropism Assessment

HIV-1 shows a tremendous genetic variability, resulting from fast replication (as many as 10^{10} virions per day are generated) coupled with a high mutation rate (approximately 10^{-5} per nucleotide per replication cycle). Even a single infected patient harbours a swarm of related HIV-1 variants (quasispecies), rather than a distinct strain, at any time point. HIV-1 uses its surface protein gp120 (coded by the *env* gene) to bind T lymphocytes or macrophages through a specific interaction with the CD4 receptor. In 1996, Gallo et al [100] discovered that some proteins known as *chemokines* were inhibiting the virus entry into cells. Subsequent studies showed that HIV-1 additionally requires chemokine receptors when entering into CD4+ cells. The two most used *co-receptors* are *CXCR4* for viral T-lymphocyte-tropic (T-tropic or X4) strains [46] and *CCR5* for macrophage-tropic (M-tropic or R5) strains [59] [16] [60] [114]. Of note, the categorisation in R5 and X4 variants is highly correlated, but not identical, to M- or T-tropism; the same holds for the classification into nonsyncytium- versus syncytium-inducing viruses [33]. HIV-1 initially binds to CD4 through a specific portion of gp120. The following interaction between the coreceptor (CCR5 or CXCR4) and a different part of gp120 (mainly the hypervariable V3 loop region) induces a conformational change resulting in exposure of a hydrophobic stretch of a second viral surface protein (gp41). This mediates the virus-cell membrane fusion step leading to release of the virus genome into the target cell [32]. The critical importance of HIV-1 coreceptors in vivo was confirmed by the observation that host subjects homozygous for a genetic deletion compromising the CCR5 functionality are naturally resistant to the infection [114] [130]. Based on these evidences, coreceptor antagonists are under development, particularly to halt disease progression in the subset of patients where drugs of other classes are

not effective anymore [95]. Since CCR5 and CXCR4 antagonists can be toxic and are not cross-effective, it is essential to know the virus coreceptor tropism in patients candidate to treatments. Moreover, HIV-1 can evolve during treatment and switch co-receptor usage, escaping inhibition by the drug. Dual-tropic, or R5X4, viruses, targeting both CCR5 and CXCR4 expressing cells have also been reported. In vitro assays (*phenotypic* tests) developed so far for analysis of coreceptor tropism are time-consuming, costly and poorly standardised. Alternatively, the *env* virus sequence can be conveniently obtained through standard laboratory procedures and the virus coreceptor tropism has then to be inferred by using in-silico prediction systems.

9.1 State of the Art

The first model introduced was the *charge rule*, proposed by De Jong in 1992 [64], which analysed the gp120 V3 loop region. It was defined as follows:

```
IF there is a positively charged amino acid at position 11 or 25
(Lysine, Arginine or Histidine)
THEN predict CXCR4
ELSE predict CCR5.
```

Following this simple model, that actually performed with good specificity, subsequent works were due to Resch [129], Pillai [107], Jensen [84] and Sing [115]. Resch proposed a Neural Network method on 216 examples, validated through bootstrap: this resulted in a significant improvement with respect to the charge rule. Pillai tested different techniques, including Support Vector Machines and Decision Trees: the input space was defined extracting mutations from a multiple alignment of the V3 loop (271 sequences) and the methods yielded little performance improvements. This work lacked accuracy in terms of sequence manipulations, since alignments did not take into account ambiguous positions, insertions or deletions. Recently, Jensen introduced a method of scoring V3 amino acid sequences on the basis of position-specific scoring matrices (PSSM). Sing presented a rigorous general learning-validation methodology based on maximisation of AUC under ROC curves, using multiple learning techniques on a larger data set made of 1'110 sequences. There have been also a web-service built after Sing's analyses and on further improvements presented in [41], the geno2pheno[coreceptor] [54], that uses also additional variables as clinical markers.

The latest work published was again improvement of Sing's work and came from the same research group, lead by O. Sander [51]: the authors produced a hybrid SVM model based on sequence and structural analysis inputs of the V3 loop, with SVM parameter optimisation through bootstrap. Using multiple cross

validation and rank-sum tests, they showed that structural descriptors are able to improve significantly prediction performances. While the structural analysis was a promising step towards a more accurate coreceptor prediction, Sander's work had a few drawbacks: only the V3 loop region was analysed; some tricks had to be used when facing sequences with insertions or deletions and calculating structural descriptors; no feature selection was performed; domain coding of sequences was limited to dummy variables representative of amino acids in codons. Our paper has the aim to provide a more accurate domain coding of the sequences (through physicochemical properties), along with the evaluation of clinical markers and subtype. In addition, feature analysis and selection is extensively evaluated in conjunction with a wide set of machine learner comparisons. We'll show that further improvements in prediction performances can be achieved even without structural descriptors, gaining also more understandable models.

9.2 Data and Methods

9.2.1 Data Collection

Viral sequences were taken from the open access *Los Alamos* HIV repository [101]. The extraction criteria were the following:

- *organism*: HIV-1, any subtype, including recombinant forms, excluding problematic sequences (i.e. high content of non-ACTG characters, likely contamination with a laboratory strain, hyper-mutation, synthetic sequences, sequences containing an artefactual deletion)
- *genomic region*: at least V3 loop sequence in *env*
- *coreceptors*: at least CCR5 or CXCR4 reported

Additional data available for these sequences were also retrieved and stored for further analyses, including mode of HIV transmission, geographic region of isolation, subtype and clinical markers. Only clinical markers and subtype were used as additional information to be fed to the machine learners, because epidemiological variables were biased towards coreceptor usage. Table 9.1 reports data set sizes, compared with the ones used in previous works. Of note, Sander used a training set which was not allowing sequence replicates, while Sing allowed for replicates, even within the same patient, using a particular validation procedure for avoiding bias. The data used in the cited studies come from the Los Alamos repository, though the sample size increased through time. Sander and Sing used also an additional set of samples coming from different clinics, which however is copyrighted. While previous modelling approaches limited the analysis to the short V3 loop region (part of gp120), our investigation was extended to the whole viral genomic regions available, with gp120 and gp41 in *env*.

data set	no. of sequences	no. of patients	R5	X4	R5X4
Los Alamos	2896	593	2114	430	352
Sing [115]	1100	332	769	210	131
Sander [51]	1100 (514 after duplicate removal)	332	363	151	?
Resch [129]	216	177	169	18	29
Pillai [107]	271	?	168	82	21

Table 9.1: HIV-1 sequences - coreceptor usage pairs

Sequence Selection, Alignment and Mutation Extraction The sequence collection policy is essential in order to avoid positively biased results. We pointed out that previous works did not use the same policy.

The Los Alamos repository contains a considerable number of viral *clones* obtained from the same patient: a raw collection of sequences thus would lead to the presence of replicates in the data sets. The policy to keep all different sequences (restricting to a particular genomic region) however should be avoided as well, since the real population distribution would not be respected.

Two identical sequences can be considered as two different observations when coming from different patients, but two identical sequences coming from the same patient can be considered as replicates. Of note, the term *identical* is defined in the sense of an exact match between a pairwise alignment of two nucleotide sequences, cutting to the shorter sequence. This notion however is far too strict, since the variability of HIV-1 within the same individual is extremely high: another approach could be to consider the amino acidic translation, though losing the information about silent mutations and needing to handle ambiguous amino acidic codes.

The proposed policy is to collect all the viral sequences that contain at least the V3 loop, keeping just one sequence per patient (the largest found) except when in presence of different coreceptor usage. Possible replicates among different patients are allowed, in order to respect the population distribution. For the analyisis, dual tropic viruses are pooled to the X4 variant set, because the clinically relevant feature is the ability to use CXCR4, irrespective of combined CCR5 use.

The total number of sequence retrieved after this filter is in table 9.2. Of note, there were 244 sequences with no patient code, that were treated as coming from the same person.

no. of seqs.	no. of patients	only V3	env	R5	X4
659	593	390	269	507	152

Table 9.2: Filtered Data Set

As it concerns the sequence alignment and characterisation of input domain, so far the most used approach has been to execute a multiple alignment of all

the sequences and then record with dummy binary variables the amino acids at each codon, following the numbering of a consensus reference sequence (namely *HXB2* [72], whose clones are X4). This solution has the advantage to produce robust alignments, but – for instance – can produce also insertions, deletions and particularly frameshifts within the consensus reference, losing the standard reference numbering. In addition, the procedure is computationally intensive: this is not a problem since has to be done only one time for the majority of sequences, but every time a new prediction is needed, a new sequence must be aligned against the multiple alignment.

We propose here an alternative solution, which is faster and has some advantages: the idea is to align pairwisely each sequence to *HXB2*, extracting only the mutations (and insertions or deletions) with respect to the reference strain and the standard numbering; thus, for each sequence, the differences against the reference are recorded.

The Smith-Waterman-Gotoh [119] [93] local nucleotide pairwise alignment algorithm was used, with high gap penalty and EDNAFULL scoring matrix. Then aligned triplets were translated into amino acids using the correct coding frame. In addition, ad-hoc procedures for ambiguous nucleotides (R, Y, K, M, S, W, B, D, H, V, N codes*) interpretation, frameshift detection and correction (it can happen that some insertions or deletions are not in frame) were executed. In summary, for each sequence a mutation list was derived, as the following example, where the first letter represents the reference amino acid, the number is the codon and the following letter(s) are the substitutions found.

```
..S110?,L111?,W112?,D113?,S128T,T132S,D133N,K135R,S142N,S143N,
S144N,G145E,R146G,I148del,E153D,I154M,S162T,S164N,I165M,R166K,
G167D,K168R,F175L,Y177H,I182V,ins184N,ins184K,ins184K,ins184T,
D185N,N186K,D187N,T188I,S190N,K192I,T194I,V200T,S209T,T236K,E268K,
I272L,V275D,A281T,T290E,S291T,K305R,R308H,I309M,Q310del,R311del,
A316Y,F317V,V318H,ins318T,I320K,G321A,K322I,N325D,M326I,I333L,
R335S,A336S,A346V,S347T,G354del,N355K,I360V,Q363H..
```

If the sequence was shorter than the *env*, missing parts were treated as missing values "?". Also highly ambiguous configurations (like "NNN", where N indicates any of the four bases) were treated as missing information. All the amino acidic substitutions found for each codon in the data set were coded as dummy binary variables if their frequency was above 3%. In addition, for each codon a general binary variable was also set, defining the presence of "any" amino acidic substitution in the position, along with other two bits for the presence of insertions or deletions. Finally, subtype was also recorded as a nominal macro-indicator, using the standard Los Alamos nomenclature.

*R = A or G; Y = C or T; K = G or T; M = A or C; S = G or C; W = A or T; B = C or G or T; D = A or G or T; H = A or C or T; V = A or C or G; N = any base.

Physicochemical and Clinical Domain Coding The binary coding of mutations is a suitable input for machine learning. But amino acids can be also classifed with respect to their physicochemical properties, that take an important role in the corresponding protein structure. For this reason, knowing also that the *11/25* rule comes from the analysis of positively charged amino acids, an additional set of variables was derived, pooling the substitutions at each codon with respect to these properties. Amino acids were classified as: hydrophilic, hydrophobic, acidic or basic. Table 9.3 summarises the classification.

Another set of input variables was then generated pooling together clinical attributes, namely: Viral RNA Load (Log_{10} cp/mL), CD4+ and CD8+ counts (cells/mm^3). Sing [41] in fact reported that CD4+ are significantly associated with coreceptor usage.

amino acid	characteristic
C, N, Q, S, T, Y	hydrophilic
A, F, I, L, M, P, V, W	hydrophobic
H, K, R	basic
D, E	acidic

Table 9.3: Characterisation of amino acid properties

The final attribute space consisted of 74 binary attributes coding the single amino acidic mutations in V3 loop, 2215 binary attributes coding the whole set of mutations (any, amino-specific, ins-del, physicochemical) in *env*, plus attributes coding subtype (nominal), CD4+, CD8+ and Viral Load (numeric).

9.2.2 Statistical Learning Methods

Statistics and Clustering Univariable χ^2 [105] and rank-sum tests were executed between input domain variables and viral tropism: in order to assess robustness against multiple testing, all the tests were 10-fold cross validated. Then Unsupervised Learning was applied to analyse mutational covariation, using Hierachical Clustering (HC) [68] [97]: HC was executed using Jaccard coefficient and agglomerative average linkage, while p-values were assessed through multiscale bootstrap resampling [61].

Machine Learners The Supervised Learning technologies considered were:

- Support Vector Machines (SVM) [67]
- Decision Trees (DT) [23] [74]
- Random Forests (RF) [77]
- Rule Based Methods (RB) [132]

- Logistic Regression (LR) [106] [35]
- Instance Based Reasoning (IBR) [7]

Support Vector Machine parameters (either using linear or Radial Basis Function kernels) were optimised through a grid search on a bootstrap sample of the training set (cost $C \in [1, 5]$ and gamma $\gamma \in [0.0001, 0.5]$). Ridge parameter r of Logistic Regression was optimised with the same procedure ($r \in [0.0001, 12]$). Instance Based Reasoning used Euclidean distance, Nadaraya-Watson evaluation function and linear kernel. For any other general theoretical aspect of the learning technologies, we refer to the book by Hastie [66].

Feature Selection The huge cardinality of the input attribute set (more than 2'000 dimensions) requires a strong feature analysis. *Filter* methods were applied to reduce input space, using either χ^2 analysis or rank-sum or Correlation-based Feature Selection (CFS), through stepwise heuristics on the attribute power set [83]: CFS evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature, along with the degree of redundancy between them. *Embedded* methods were applied intrinsically by DT splitting and pruning procedures, rule finding policies for RB (Repeated Incremental Pruning to Produce Error Reduction (RIPPER) [132]) and (optimised) Ridge Shrinkage [106], LogitBoost [35] or stepwise selection based on Akaike Information Criterion (AIC) [58] for Logistic Regression.

Software The data mining suite *Weka* [112] and the mathematical programming language *R* [111] were used as base software for all analyses carried on.

Validation and Performance Assessment

As described in the methods section, *accuracy, AUC, sensitivity, specificity and f-measure*, will be evaluated, along with the usage of multiple cross validation for performance assessment and model selection.

9.3 Results

9.3.1 Univariable and Covariation Analysis

In this section univariable analysis will be carried on in order to discover variables significantly associated with coreceptor usage: these findings will be also used for feature selection in the machine learning phase. Then variable covariation will be investigated.

Table 9.4 show descriptive statistics for clinical markers and subtypes: under univariable rank-sum analysis (andalso chi-square on discretisation) low CD4+

counts were found to be significantly associated with X4 viruses ($p \leq 0.01$, also previously reported in [41]). Subtype was also significantly associated ($p \leq 0.01$) to coreceptor usage under chi-square test. In particular, subtypes B and D were associated with X4 viruses, whilst A, C and 02_AG with R5 (min $p \leq 0.05$ for 02_AG).

marker	average (st.dev)	% missing	median X4	median R5
CD4+ cells/mm^3	391 (366)	66	35	334
Viral Load Log_{10} cp/mL	4.898 (1.009)	67	-	-
CD8 cells/mm^3	1146 (910)	98	-	-
subtype	no. (tot=659)	prevalence	preval. X4	preval. R5
B	235	0.357	0.477	0.333
01_AE	39	0.059	-	-
C	199	0.302	0.174	0.352
02_AG	19	0.029	0.000	0.039
A1	23	0.035	-	-
D	32	0.048	0.121	0.028
A	36	0.056	0.001	0.073
other	58	0.008	-	-
missing	18	0.027	-	-

Table 9.4: Descriptive statistics for clinical markers and subtype attributes. Median values and prevalence were reported for X4/R5 groups only if the rank-sum or chi-square test was significant ($adj.p \leq 0.05$).

As it concerns mutations, there were no missing values in the V3 loop by definition, but ouside this region the missing values ranged from 20% to 70%, depending on the sequence lengths. Table 9.5 reports χ^2 values for the top ranked variables ($p < 0.01$), with standard deviations obtained from cross-validation. Figure 9.1 shows then prevalence for the mutations significantly associated with tropism. Out of 2215 mutation variables, 130 resulted significant: most of them were positions within the V3 loop, but a few positions in other *env* region were also detected (362_hydrophobic, 440_basic, 293_hydrophobic among the best ranked). The physicochemical coding and the "any" coding for mutations resulted to be highly discriminant when the corresponding positions were detected significant.

Regarding positions in V3 loop, as expected position 306 (corresponding to position *11* in the reference V3 numbering) plays a dominant role. Position 322 (i.e. *25*) is significant, but there are other more discriminant positions, namely 302(*7*), 303(*8*), 323(*26*), 301(*6*), 313(*16*), 321(*24*). Note that also some insertions and deletions are detected significant: 306_del, 321_ins, 307_del, plus other lower ranked. Jensen [84] reported that V3 sequences with ins-del are typically X4, while here the opposite holds, since the reference strain is actually X4. Ins-del can produce sensible conformational changes and mutant structures are difficult to be modelled with standard software. Sander [51] reported that structural descriptors are sufficiently robust to handle sequence variants containing ins-del (actually the designed structural descriptors ignore them), obtaining slightly dif-

ferences in prediction performances.

It's worth to point out here that, indeed, all the results are biased by the alignment procedure: if two different alignment schemes are used, this can produce a systematic error in mutation detection and coding; this is not a big problem in the modelling phase, but can lead to misunderstanding when comparing different analyses. For instance, the pairwise alignment (Smith-Waterman-Gotoh + frame corrections) algorithm here implemented not always reports the same set of mutations that the geno2pheno[coreceptor] [54] web-service gives.

In figure 9.2 Hierarchical Clustering to assess mutational covariation is depicted. Physicochemical indicators or "any" substitution codes were used only if no specific amino acids were found significant in the previous analysis. The covariation analysis was restricted to the sole V3 loop. Red boxes assess significance for branches ($p < 0.05$) From the graph, some defined associations can be easily found. At a threshold height of 0.35, the following clusters are identified: {317ins, 318A}; {311I, 308S, 306del, 307del}; {322I, 320_hydrophilic, 326I}. The associations probably have a role for the stabilisation of the mutant structure and for the tropism, but it seems that mutations positively associated with X4 viruses tend to behave more independently (see for example 306S, 303I and 308K, 300Y and 307T...). This could be explained by the fact that just a few changes are needed towards a coreceptor switch, along with the relative low frequency of X4 strains and related mutations.

9.3.2 Prediction Models

Assessing robustness and improvements in coreceptor usage prediction engines involves two scenarios: first, to compare different models and derive error estimation for the naive set of dummy variables in the V3 loop; second, verify whether the usage of additional descriptors (physicochemical, clinical) improves performances using different methodologies as well. The second point requires also accurate feature selection, since the input spaces grows dramatically (from 74 to 2000+ attributes). Moreover, different loss functions must be evaluated, since the class distribution in the data set is unbalanced.

Multiple Cross Validation is an effective solution for comparing model performances: in fact, the error estimation calculated from multiple CV runs yields Gaussian distributions that can be compared through a t-test, assessing p values for statistically significant differences.

Table 9.6 shows results for different models, trained using the sole V3 loop region, without physicochemical neither clinical indicators. As expected, best models are found to be SVM and Logistic, as also pointed out on [115] and [51]. Even if direct comparison with the performances published by other studies is not possible (due to slightly different data sets and sequence selection policies used), raw numbers are consistent with Sander's [51] results when SVM are applied

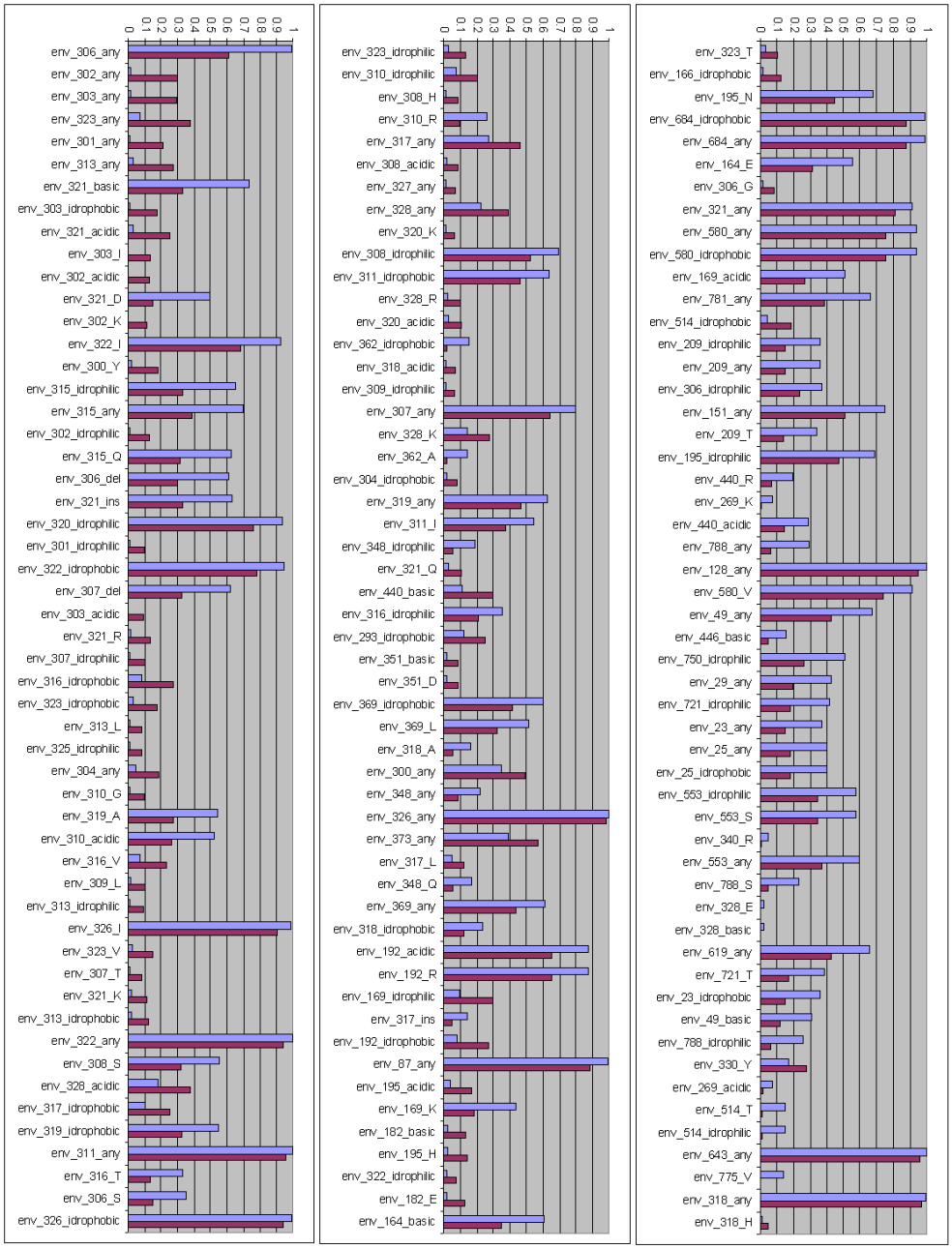

Figure 9.1: Frequency of amino acidic substitutions significantly associated with coreceptor usage ($p < 0.01$). Azure is for R5 and purple for X4.

χ^2	st.dev	variable
176.721	5.779	306_any
122.682	8.621	302_any
115.506	7.164	303_any
81.606	5.761	323_any
83.334	6.407	301_any
78.858	4.068	subtype
78.313	8.378	313_any
74.404	6.504	321_basic
72.098	4.527	303_hydrophobic
66.98	2.118	321_acidic
61.722	4.082	303_I
53.993	3.362	302_acidic
53.775	4.394	321_D
52.217	1.588	302_K
53.685	7.61	322_I
46.888	4.617	300_Y
45.22	4.114	315_hydrophilic
43.313	4.799	315_any
42.846	6.17	302_hydrophilic
42.989	4.201	315_Q
40.972	4.367	306_del
38.093	3.124	321_ins
37.639	3.654	320_hydrophilic
37.453	4.389	301_hydrophilic
36.876	6.26	322_hydrophobic
35.603	3.933	307_del
35.27	3.058	303_acidic
34.594	4.145	321_R
34.213	3.008	307_hydrophilic
34.125	4.773	316_hydrophobic
32.478	3.684	323_hydrophobic
31.398	4.388	313_L
31.41	5.724	325_hydrophilic
31.2	2.313	304_any
31.098	6.553	310_G
30.369	3.315	319_A
29.031	3.832	310_acidic
28.699	5.172	316_V
28.225	3.002	309_L
28.052	3.541	313_hydrophilic
28.124	3.819	326_I
27.859	3.651	323_V
25.363	3.082	307_T
24.505	2.412	321_K
23.448	4.229	313_hydrophobic
23.018	3.219	322_any
22.978	2.814	308_S
22.865	3.438	328_acidic
22.358	2.595	317_hydrophobic
21.969	2.862	319_hydrophobic
21.137	2.154	311_any
20.932	1.51	316_T
20.269	1.983	306_S
19.519	3.317	326_hydrophobic
19.153	3.911	323_hydrophilic
18.718	3.87	310_hydrophilic
18.531	3.567	308_H
17.481	2.05	310_R
16.69	1.974	317_any
16.71	3.16	308_acidic
15.321	2.425	327_any
15.007	2.487	328_any
14.881	3.205	320_K
14.362	1.865	308_hydrophilic
14.455	2.498	311_hydrophobic
13.835	2.983	328_R
13.468	2.75	320_acidic
13.264	1.259	CD4+
12.825	1.031	362_hydrophobic
13.708	3.55	318_acidic
13.065	2.847	309_hydrophilic
13.039	2.529	307_any
12.188	2.057	328_K
11.796	0.893	362_A
11.556	2.151	304_hydrophobic
11.563	1.634	319_any
11.875	1.788	311_I
11.327	1.791	348_hydrophilic
11.372	1.939	321_Q
10.97	1.936	440_basic
10.327	0.501	316_hydrophilic
10.613	2.216	293_hydrophobic
10.28	1.855	351_basic
10.28	1.855	351_D

Table 9.5: Cross validated chi-square test on variables and viral tropism cross tabulations (first 85 results shown)

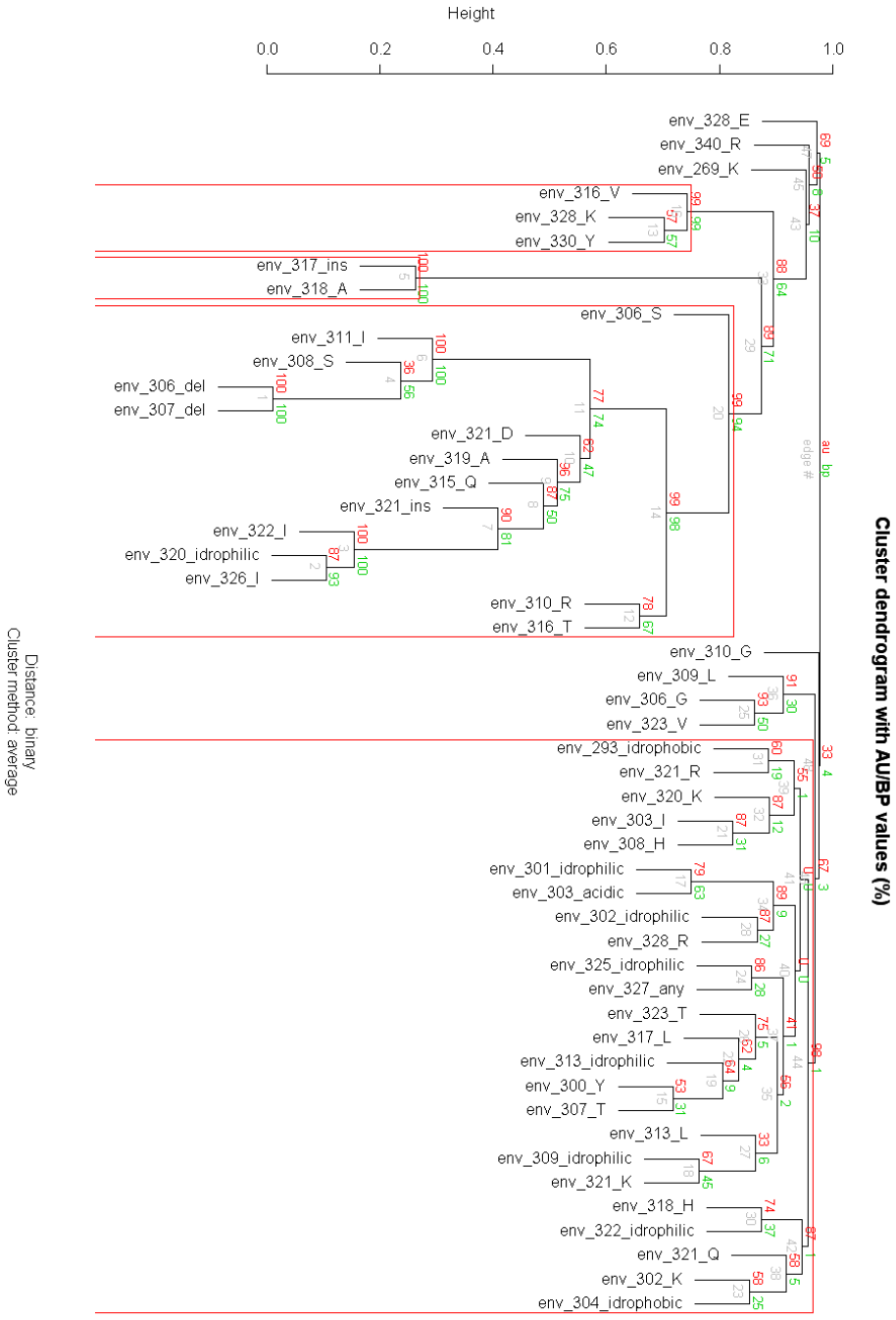

Figure 9.2: Bootstrapped hierarchical clustering of mutations significantly associated with coreceptor usage. Red boxes enclose branches with $p < 0.05$. Jaccard similarity coefficient and average linkage method used.

only V3 loop		multiple cross validation			
model	feature selection	accuracy % (st.dev)	AUC (st.dev)	sensitivity (st.dev)	f-measure (st.dev)
SVM lin	–	90.32 (3.35)	0.89 (0.06)	0.70 (0.11)	0.94 (0.02)
SVM RBF	–	89.59 (3.62)	0.91 (0.05)	0.69 (0.10)	0.93 (0.02)
Logistic	ridge shrinkage	88.77 (3.38)	0.89 (0.06)	0.68 (0.11)	0.93 (0.02)
Logistic	LogitBoost	88.51 (3.51)	0.89 (0.05)	0.66 (0.12)	0.93 (0.02)
RF	split/prune/rand	88.26 (3.65)*	0.88 (0.06)	0.63 (0.12)*	0.93 (0.02)
RB	RIPPER	87.10 (3.84)*	0.80 (0.06)*	0.64 (0.12)	0.92 (0.02)*
DT	split/prune	86.42 (3.83)*	0.81 (0.07)*	0.57 (0.13)*	0.92 (0.02)*
IBR	–	86.33 (3.64)*	0.82 (0.08)*	0.52 (0.14)*	0.92 (0.02)*
majority class	–	76.93 (0.60)*	0.50 (0.00)*	0.00 (0.00)*	0.87 (0.00)*

Table 9.6: Model comparison (10 independent runs of 10-fold CV) using only the V3 loop genomic region with naive amino acid encoding, without physicochemical or clinical attributes. Values with * are significantly worse than the best model.

without the addition of structural descriptors.

Sander reports 90.00% accuracy and 0.92 AUC, without giving st.dev; when structural descriptors are added, performances increase (91.56% accuracy, 0.93 AUC) and the improvement is proven to be statistically significant ($p = 0.002$). Sensitivity is reported at a fixed specificity 0.95 (0.73 and 0.80 without and with structural descriptors respectively), which is higher than the pure sensitivity: the choice is reasonable, since they accept a fixed number of false positives, but in this study we prefer to show the pure sensitivity and specificity rates. In summary, these first results obtained using the sole V3 loop, without additional indicators, seem comparable with Sander's naive model and seem inferior to his enhanced model, which was exactly the expectation. Of note, the Logistic model obtained through ridge shrinkage could be preferable to a SVM, since uses a small set of variables (whose importance and p values can be assessed multivariately looking at the odds ratios and standard errors) and is simple to understand.

Let's examine now how the models behave when the complete set of attributes (i.e. *env* region, physicochemical properties, clinical variables) is fed to them. Table 9.7 shows that again Logistic Regression and SVM are the best methods: moreover, these models are significantly better – wrt any loss function – than the best model obtained using the sole V3 loop, with naive variable encoding ($p < 0.05$). The reported values are also higher than the results presented by Sander using the structural descriptors (but still this is not a proper way to make comparisons).

Apart from the pure performance issues, let's note that the best performing

env region, physicochemical, clinical inputs		multiple cross validation			
model	feature selection	accuracy % (st.dev)	AUC (st.dev)	sensitivity (st.dev)	f-measure (st.dev)
Logistic	ridge shrinkage + CFS	92.76 (3.07)	0.93 (0.04)	0.76 (0.10)	0.95 (0.02)
SVM RBF	–	92.75 (3.19)	0.91 (0.06)	0.78 (0.10)	0.95 (0.02)
Logistic	ridge shrinkage	91.62 (3.34)	0.92 (0.04)	0.72 (0.12)	0.95 (0.02)
Logistic	LogitBoost + CFS	91.73 (3.13)	0.91 (0.05)*	0.73 (0.11)	0.95 (0.02)
Logistic	LogitBoost	90.87 (3.24)*	0.91 (0.05)*	0.72 (0.11)	0.94 (0.02)*
RF	split/prune/rand + CFS	89.41 (3.26)*	0.91 (0.05)*	0.68 (0.12)*	0.93 (0.02)*
RB	RIPPER	89.17 (3.60)*	0.82 (0.06)*	0.67 (0.12)*	0.93 (0.02)*
RF	split/prune/rand	88.73 (3.17)*	0.90 (0.05)*	0.60 (0.11)*	0.93 (0.02)*
DT	split/prune + CFS	88.30 (3.59)*	0.84 (0.06)*	0.62 (0.12)*	0.93 (0.02)*
RB	RIPPER + CFS	87.83 (3.81)*	0.80 (0.06)*	0.64 (0.12)*	0.92 (0.02)*
IBR	CFS	87.63 (3.52)*	0.84 (0.06)*	0.60 (0.12)*	0.92 (0.02)*
DT	split/prune	87.35 (3.45)*	0.83 (0.06)*	0.60 (0.11)*	0.92 (0.02)*
majority class	–	76.93 (0.60)*	0.50 (0.00)*	0.00 (0.00)*	0.87 (0.00)*

Table 9.7: Model comparison (10 independent runs of 10-fold CV) using the whole *env* region, with physicochemical encoding and clinical attributes. Values with * are significantly worse than the best model.

model is a Logistic Regression with a considerable low number of variables (applying first univariable + CFS filter). Physicochemical properties were included, along with a few positions in *env* outside the V3 loop region. Subtype A and CD4+ were also selected by univariable and by CFS filter, but their significance was lost in the multivariable analysis. After applying an additional stepwise selection using AIC, CD4+ was dropped from the model (in fact, the weight was also set close to zero with the ridge shrinkage). Table 9.8 shows the most compact model obtained, which is still among the best models in terms of accuracy and AUC.

9.4 Conclusions

Discussion In this paper a new domain coding for in-silico modelling of HIV tropism is proposed, that takes into account the whole *env* region, physicochemical properties of amino acids and clinical markers. Univariable analysis is carried

variable	estimate	std.error	z-value	Pr(> |z|)	sign.
(Intercept)	10.6713	16.9266	0.630	0.528402	
env_306_any	-3.7963	0.7195	-5.276	1.32e-07	***
env_302_any	1.6593	0.7548	2.198	0.027925	*
env_301_any	2.1777	1.0100	2.156	0.031071	*
env_313_any	1.6468	0.4943	3.331	0.000864	***
env_321_basic	-0.5858	0.3752	-1.561	0.118453	
env_321_acidic	1.3595	0.5635	2.413	0.015833	*
env_303_I	7.3132	9.9710	0.733	0.463285	
env_309_L	1.7320	0.9442	1.834	0.066591	.
env_311_any	-8.6866	16.9227	-0.513	0.607734	
env_302_K	5.3314	10.8174	0.493	0.622114	
env_192_hydrophobic	2.0997	0.6328	3.318	0.000907	***
env_315_any	-0.9634	0.3509	-2.746	0.006041	**
env_320_hydrophilic	-0.8753	0.5394	-1.623	0.104666	
env_310_G	2.6024	0.7615	3.417	0.000632	***
env_169_hydrophilic	1.1246	0.6414	1.753	0.079546	.
subtype A	-4.5072	4.7387	-0.951	0.341535	
env_328_acidic	0.7813	0.3935	1.986	0.047072	*
env_306_G	2.0825	0.6790	3.067	0.002164	**
env_293_hydrophobic	1.2269	0.4312	2.845	0.004442	**
env_195_acidic	1.3055	0.8207	1.591	0.111683	
env_209_hydrophilic	-1.1828	0.6293	-1.879	0.060188	.
env_325_hydrophilic	2.8179	1.2872	2.189	0.028579	*

Table 9.8: Logistic model built after χ^2 + CFS + stepwise AIC feature selection. Coefficients for X4 prediction. Null deviance: 711.80 on 658 degrees of freedom. Residual deviance: 266.05 on 636 degrees of freedom. AIC: 312.05. Significance p-values are: '***' $p \leq 0.001$; '**' $p \leq 0.01$; '*' $p \leq 0.05$; '.' $p \leq 0.1$.

on to select significant features and covariation of mutations is assessed through hierarchical clustering. A wide set of machine learning and feature selection methods is then applied and models are compared through t-tests on multiple validation runs. SVM and Logistic Regression (LR) models are found to be the best techniques, while the enhanced input domain coding ensures a performance increment which is statistically significant under different loss functions. Specifically, clinical markers (CD4+ cell counts) are found to be independent predictors of tropism, though in multivariable LR analysis the p-value increases and they can be dropped. The inclusion of mutations in the *env* region instead gives a few positions outside the V3 loop that are significantly associated (either under univariable or multivariable) with tropism, while the physicochemical coding takes the major role for the assessment of performance improvements. Finally, the best model turns out to be a LR that uses a considerable low number of variables: the importance of such a result is not only in the improvement of the state of the art for sequence-based learning methods, but also in the fact that the derived LR is compact and easy to interpret (and variable importance can be assessed multivariately, with odds ratios and standard errors).

Future Perspectives The improvement gained with the usage of physicochemical properties and clinical markers in the input attribute space, along with the analysis of the *env* region, could be easily coupled with the structural descriptors defined in [51], in order to see if further statistically significant improvements in the coreceptor prediction can be achieved. At the same time, additional information about positive-negative charge of amino acids can be added. Another interesting issue that should be investigated is the structural comparison between different chemokines and HIV V3 loop, with a proper mathematical description. Specific chemokines act as natural ligands for HIV coreceptors (particularly RANTES, MIP1-alpha, MIP1-beta for CCR5) and have been shown to block HIV replication in vitro [37] [39]. Finding an effective way to learn their structure-function correlation can provide additional information that can be modelled to improve the capability of predicting HIV coreceptor use from *env* sequence.

Chapter 10

In-Vivo: Therapy Optimisation and Follow Up Prediction

Predicting the actual viral load changes following treatment switches is a challenging task. The individual variability of immune response to infections adds considerable noise and the large number of possible drug combinations, along with mutational patterns, makes the problem complex. Other treatment-related factors – such as pharmacokinetics and patient adherence to therapy – play a crucial role in the control of virus replication and the development of resistance, but they're typically unknown.

10.1 State of the Art

Association of three or more antiretroviral drugs (HAART, Highly Active Anti Retroviral Therapy) has led to significant decreases in HIV-related morbidity and mortality by reducing often viral replication to undetectable levels [1]. However, complete eradication is not feasible with the current treatment armamentarium: drug-resistant variants can ultimately develop in patients, though most of them in high income countries can avoid resistance if adherence is good. These mutants display different degrees of decreasing susceptibility to the ongoing treatment regimen and often cross-resistance to other agents. This results in virologic rebound and eventually disease progression [82]. As long as drug resistance accumulates, it is critical to choose the appropriate drugs and build a "salvage" therapy resulting in the largest and longest-lasting viral load reduction. More than 100 mutations occurring in the HIV *pol*[*] gene have been found implicated in resistance to 20

[*]the *polymerase* gene is implicated in the viral reproduction and two major regions are targeted by current inhibitors: Reverse Transcriptase and Protease

available HIV-1 inhibitors.

Although robust (virtual) phenotyping and genotyping methods have been established, most experts eagerly awaits a support decision tool able to predict in-vivo response to treatment in terms of viral load as well as CD4+[†] cell count change, along with a rank of suitable therapies. This task could be pursued either by a naive learning procedure from genotype to treatment response, or adding information from in-vitro resistance, or deriving more discriminant indicators. Most of the currently available algorithms for genotype interpretation are Rule-Based (RB) [9] [8] [13]. Alternative approaches have been proposed, including Neural Networks (NN) [63] [40], Fuzzy-Rule-Based (FRB) systems [38] and DTs that take into account derived features through evolutionary pathways modelling [91] [109]. Overall, current models lack a rigorous input space modelling and performance evaluation, in terms of feature selection, feature extraction and error validation robustness assessment. For instance, RB methods ([8] [13]) were not properly validated, but only proven significantly associated with treatment outcome (through logistic multivariable analysis); in addition, they rely only on genotypic information. Neural Network based methods ([63] [40]), though explored a larger set of attributes, did not provide a clear data collection policy: this can yield positively biased results due the presence of replicates. Finally, the most promising approach based on DT and mutagenetic trees ([91] [109]) used non-standard data collection (now being re-considered), which demonstrated lower performances tested on the actual standard data.

10.2 Data Collection, Domain Coding and Methods

EuResist

The Eu*Resist* [98] [36] project is a STREP funded by the European Union within the Sixth framework Programme since January 2006, for the development and the management of a data warehouse, aiming at integrating virological and clinical information and developing antiretroviral (ARV) treatment optimisation tools. Several private and academic partners are working together collecting data (University of Siena, Ita; Karolinska Institute, Swe; Caesar, Ger), integrating data bases (IBM Haifa Research Lab, Isr), providing statistics (Kingston University, UK), building prediction tools through machine learning (Informa CRO, Ita; RMKI, Hun; Max Planck Institute, Ger; University of Roma TRE, Ita; IBM Haifa Research Lab, Isr) and providing web services (Informa CRO, Ita; IBM Haifa Research Lab, Isr). The work presented in this paper is at the base of the implementation of one out of four prediction engines that will be ultimately integrated as a public web service.

[†]measure of immune response

By now the integrated DB is composed mainly by three national data bases (Italian, German and Swedish) and smaller satellite sources. It is probably the largest data base in the world as it concerns the collection of clinical, genomic and epidemiological data, and its size is increasing through time with the spontaneous addition of new sources. Table 10.1 summarises the size of the principal information available, but also other tables are present, like real phenotypic data, HBV/HCV co-infections, CD8+ markers.

Even if this is out of the scope of this paper, it's worth to note that the DB is indeed a valuable source for all kind of statistical analyses, either concerning macro- (like phylogenetic analysis, epidemiology) or micro-objectives (analyses on selected therapy combinations, mutations). Moreover, the consortium agreements permit the collaboration with external partners when in presence of data sharing and related studies.

The Eu*Resist* Integrated DB

patients	17078
viral sequences	19444
therapies	59982
CD4+ isolates	279506
viral load RNA isolates	214516
Standard Datum instances	
with missing baseline info	2523
with complete baseline info	2176

Table 10.1: Number of instances for tables and standard data views on the integrated Eu*Resist* data base (July 2007 update).

Standard Datum Definition

A view on the Integrated DB was created following the *Standard Datum* definition proposed by the HIV Forum for Collaborative HIV Research [49] [26] and further implemented by the Eu*Resist* consortium.

In the Standard Data view, each instance is representing a Treatment Change Episode (TCE): a TCE corresponded to the real situation in which a patient starts a (new) anti retroviral therapy, either for the first time or after a previous failure. Usually the physicians decide a new combination therapy considering the patient's clinical background and actual condition (baseline), toxicity/adherence odds and eventually evaluating the report of an in-vitro resistance test. The patient is then followed up and the viral load or CD4+ cells are measured by subsequent analyses usually at 8, 12, 24, 48 weeks and more. The treatment is not interrupted and considered successful if the viral load is kept at undetectable levels, if the patient tolerates the therapy, if the CD4+ counts are high. When the viral load rebounds (due to lack of adherence or viral drug resistance rise) or

AIDS-defining events happen, usually the therapy (from one drug to the entire regimen) is changed, though this is not a necessary condition (a therapy change can be decided for other circumstances).

Figure 10.1 shows an ideal TCE as an instance of the view on EuResist DB: corresponding baseline variables (viral genotypes, clinical markers) were considered in a constrained time window around the new therapy start date ([-90,0] days). Patients' demographics and past treatments were collected when available. The aim was to predict a short-term (eight weeks, in a time window of [+4,+12])

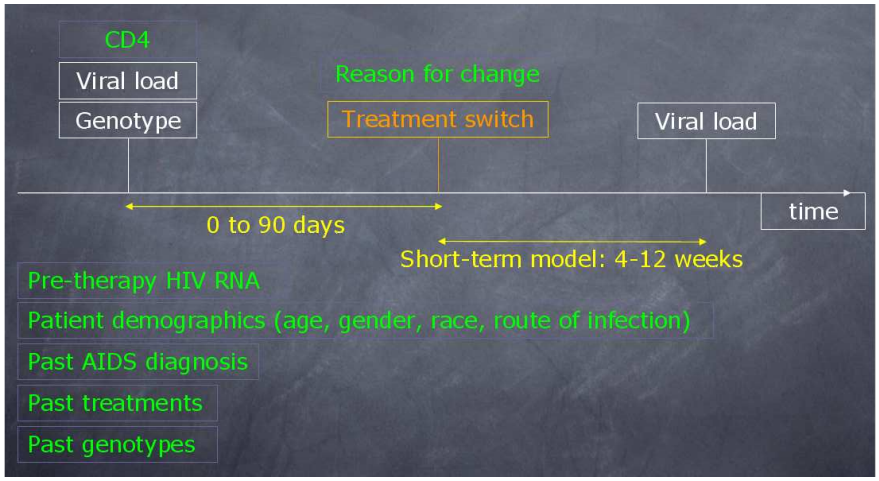

Figure 10.1: Standard Datum Instance

follow up outcome in terms of *virological success* – defined as the achievement of less than 500 copies of viral RNA load or decrease from the baseline value by two or more Logs –.

When in presence of multiple baseline or follow up data, the values closest to the therapy start date and 8th week follow up respectively were taken.

Additional minor constraints were also included, along with the exclusion of "suspicious" instances (for example the usage of Fusion Inhibitors as first line therapies): indeed, the Integrated DB is a retrospective cohort with data coming from heterogeneous and not equally reliable sources.

The genotypes collected were aligned through CLUSTALW [70] to subtype[‡] *B* consensus reference and mutations were extracted for each position in Reverse Transcriptase (RT) and Protease (PR), accounting also for ambiguous positions and missing values (i. e. shorter region sequenced), with a frequency threshold of

[‡]subtypes are the different genetic lineages of HIV-1, unevenly distributed in different geographic regions.

4% (more than 400 mutations were found). A typical genotype thus was a list of mutations of the form:

```
Protease (PR) mutations:
L10FIR,N37T,R41K,M46I,L63P,I66IV,A71V,V82T,I84V,L90M,I93L,C95F

Reverse Transcriptase (RT) mutations:
M41L,D67N,L74V,K103N,D123E,E138Q,I178L,M184VI,Q207E,L210S,T215Y
```

where the first letter coded the amino acid found in the consensus reference, the number was the codon position and the following letter(s) coded the substitution(s) present in the viral isolate. Genotypes were coded as real vectors in [0,1], where each element represented an amino acidic substitution in a codon position and the value was the fraction of the specific change observed in the sample[§].

No restrictions on therapies were contemplated, i. e. suboptimal treatment regimens made of less than three drugs were allowed: for this reason, in this paper the general *combined Anti Retroviral Therapy* (cART) term was used in place of the more specific HAART (defined as the use of at least three drugs belonging to more than one drug class). The following ARVs, all currently approved by the Food and Drug Administration (FDA) [48] and European Medicines Agency (EMEA) [6] were considered: nucleotide/nucleoside reverse transcriptase inhibitors AZT, D4T, DDC, ABC, 3TC, FTC, DDI, TDF (NRTIs); non-nucleoside reverse transcriptase inhibitors DLV, EFV, NVP (NNRTIs); protease inhibitors APV (or FPV), ATV, IDV, LPV, NFV, RTV, SQV, TPV, DRV, along with boosting RTV[¶] (PIs). DRV, DLV and TPV were excluded from the analysis due to low frequency.

Attribute Domains Table 10.2 summarises the attributes fed to the Machine Learners. Also derived features are included (described further). Sizes of selected training sets were showed in table 10.1.

As also described in figure 10.1, the input domain for a TCE consists of mandatory attributes which are the baseline (i.e. closest to the start date of the new therapy) viral sequence, with extracted mutations, and the new therapy (cART). The optional attributes consist of baseline markers (viral load and CD4+ cell counts), patients' demographics and epidemiological information (age, sex, ethinicity, mode of HIV transmission, viral subtype...), along with information about past treatments (total time exposure for each drug and time since the administration was stopped). Derived features consist of new features which can be calculated from the former attribute space, using different combining functions.

[§]in the text, mutations will be named only referring to the substitution in a specific position of the gene, i. e. PR_33_F will correspond to L33F in Protease.

[¶]a sub-therapeutic low dose of RTV used to enhance the activity of a companion PI through a favourable pharmacokinetic interaction

attribute	domain	type
cART	$\{0,1\}^d$	mandatory
baseline viral RNA load	$\mathbf{R}$	optional
baseline viral genotype mutations from consensus B	$[0,1]^n$	mandatory
baseline viral subtype	nominal	optional
baseline viral genotype consensus B match	$\mathbf{N}$	optional
baseline CD4+ cell counts	$\mathbf{R}$	optional
steady-state viral RNA load	$\mathbf{R}$	optional
steady-state CD4+ cell counts	$\mathbf{R}$	optional
risk	nominal	optional
ethnicity	nominal	optional
country of origin	nominal	optional
country of infection	nominal	optional
age	$\mathbf{N}$	optional
gender	nominal	optional
drug total time exposure	$\mathbf{R}^d$	optional
drug time since not used	$\mathbf{R}^d$	optional
drug history function	$[0,1]^d$	derived
NRTI previous usage	$\{0,1\}^d$	derived
NNRTI previous usage	$\{0,1\}^d$	derived
PI previous usage	$\{0,1\}^d$	derived
phenotypes	$\mathbf{R}^d$	derived
literature resistance associated mutation accumulation	$\mathbf{R}^d$	derived
number of drugs used	$\mathbf{N}$	derived
fuzzy phenotypic score (optimistic)	$\mathbf{R}$	derived
fuzzy phenotypic score (pessimistic)	$\mathbf{R}$	derived
fuzzy literature resistance rules score (optimistic)	$\mathbf{R}$	derived
fuzzy literature resistance rules score (pessimistic)	$\mathbf{R}$	derived
differential equation approx. (pessimistic)	$\mathbf{R}$	derived
differential approx. (optimistic)	$\mathbf{R}$	derived
second-order and third-order variable interactions	{0,1}	derived
8th week viral RNA load	$\mathbf{R}$	mandatory
8th week success	$\{0,1\}$	mandatory, derived

Table 10.2: Standard Datum Attribute Space.

For instance, the phenotype is calculated using the baseline viral sequence and gives information about the in-vitro resistance for a single drug pressure. The accumulation of mutations listed in literature counts specific mutational changes for each drug listed as resistant in literature guidelines. Fuzzy resistance scores use the former information of accumulation of mutations to calculate overall efficacy scores for a cART. Second- and third-order variable interactions set up additional dummy variables to account for mixed effects. Next section will describe each derived feature in detail.

Out of 2523 suitable instances (2176 with baseline viral RNA load), 2269 were used for training and validation, while 254 (244 with baseline info) instances were reserved as independent test set. Data set with multiple observation was composed by 3688 instances, with complete information.

Subtype assignment was calculated trough a BLAST search [110] on an updated reference subtype sequence data set from Los Alamos data base [101].

Machine Learners, Feature Selection, Losses, Validation Procedures The Supervised Learning technologies considered were:

- (Bagged) Decision Trees (DT) [74] [23] [76]
- Random Forests (RF) [77]
- Logistic Regression (LR) [106] [35]
- Instance Based Reasoning (IBR) [7]

Feature Selection was carried on using (i) filter methods based on univariate analysis and Correlation-Based Feature Selection (CFS) [83], (ii) embedded selection based on (boosted) tree splitting and ridge shrinkage for Logistic models [106] [35], (iii) wrappers with heuristics developed by the authors [81] were explored. Specifically, from the whole set of attributes, a filter based on non-parametric univariable analysis (Chi-Square and Wilcoxon rank-sum adjusted for multiple testing) selected attributes significantly associated with treatment response ($p \leq 0.05$): this selected set of features was fed to all the MLs. Then features were selected by the embedded methods specific of each ML: intrinsic splits and pruning on DTs or ridge shrinkage on Logistic Regression, plus the usage of CFS. References are also given in the introductory section 8.3.

Loss functions considered were AUC and *accuracy*. Multiple 10-fold CV was carried on for model selection and performance assessment.

10.3 Feature Derivation

All the features and additional indicators presented here have been derived *independently* from the training data, in order to avoid any bias in the performance

estimation. Each derived feature is calculable directly from the mandatory attributes of the Standard Datum: in this way no additional information is requested in the input. Experimental results will show that adding these features to the ML models significantly improves the results.

10.3.1 Phenotypes

One immediate idea is to add an estimation of in-vitro drug resistance to the input attribute set. The in vitro resistance of a given virus isolate (representative of patient's viral population), known as *phenotypic test*[‖], is done through cumbersome and time-consuming assays, performed at high cost by a few specialised companies. Alternatively, predicting the phenotype to single drugs from viral genotype has been accomplished with increasing accuracy. Multiple Linear Regressors (MLR), Decision Trees (DT) and Support Vector Machines (SVM) applied to genotype-phenotype pairs are able to perform predictions that explains correctly up to 80% of phenotypic variance [55] [90] [89].

Multiple Linear Regression (MLR) was shown to compete with more complex models and was the choice for this experiment. MLR models were trained and validated on a large set of genotype-phenotype pairs [126], obtaining Log Fold Change predictors for each drug compound. Thus, virtual phenotypes were calculated at any Treatment Change Episode (TCE) for each drug in the cART.

Feature selection was carried on selecting first variables significant under univariable chi-square analysis and then applying stepwise selection, assuming conditional independence on variables. The procedure resulted able to shrink the input space set from more than 400 variables (i.e. mutations) down to dozens. Table 10.3 reports cross-validation performances (Pearson's ρ correlation between predicted and observed outcome) for each drug.

Apart from the importance to have extremely reliable in-silico models to simulate laboratory experiments, the usefulness of these methods is in the sense that they can be applied to viral sequences associated to in-vivo data, in order to derive phenotypic values to be used as derived features for therapy optimisation. The results obtained ensure high performances, considering also the large training sample sizes. MLR models are competitive with SVM implemented in [55], but have the advantage to be more understandable: moreover, the feature selection techniques gave a reduced set of mutations in perfect agreement with previously reported statistics [121], which weights resemble also the resistance/susceptibility contributions (see table 10.4 for a few examples).

[‖] the phenotype is measured as a *fold change*, representing the drug dosage needed to inhibit the 50% of the viral mutant replication wrt wild type reference value, in laboratory cultures

NRTI & NNRTI	ρ	no. ex.
AZT	0.9408	939
ABC	0.9428	894
3TC	0.9735	934
D4T	0.9201	918
DDC	0.9398	891
DDI	0.8836	889
TDF	0.9148	770
DLV	0.9243	963
EFV	0.9309	957

PI	ρ	no. ex.
IDV	0.9669	936
NFV	0.9613	943
NVP	0.8784	972
RTV	0.9718	930
SQV	0.9702	954
APV	0.9679	944
LPV	0.9662	756
ATV	0.9572	342

Table 10.3: Phenotype Prediction: 10-fold cross validation ρ correlation

D4T Fold Change Log = 0.0594 · RT_44_A +0.0733 · RT_44_D +0.0242 · RT_62_V +0.0627 · RT_67_N +0.0713 · RT_69_D +0.0221 · RT_70_R +0.1096 · RT_75_M -0.0183 · RT_83_K +0.2528 · RT_116_Y +0.0222 · RT_118_I +0.7002 · RT_151_M -0.0831 · RT_184_V +0.1137 · RT_210_W -0.012 · RT_211_K -0.0226 · RT_214_L +0.1131 · RT_215_F +0.1255 · RT_215_Y +0.0231 · RT_219_Q -0.0442
APV Fold Change Log = -0.3956 · PR_88_S +0.1918 · PR_90_M +0.2225 · PR_54_V +0.2416 · PR_46_I +0.4812 · PR_84_V +0.0491 · PR_82_A +0.286 · PR_33_F +0.469 · PR_32_I +0.0993 · PR_10_I +0.1869 · PR_34_Q +0.0825 · PR_73_S +0.1598 · PR_47_V +0.1767 · PR_10_F +0.1263 · PR_24_I +0.7521 · PR_50_V +0.21 · PR_46_L +0.0712 · PR_55_R +0.065 ·PR_58_E +0.2508 ·PR_54_M +0.0781 ·PR_67_F +0.0893 ·PR_48_V -0.0401 · PR_82_T +0.0861 · PR_33_I -0.0236 · PR_88_D -0.2071

Table 10.4: Phenotype Prediction: MLRs for D4T and APV

10.3.2 Previous Class Exposure and Combined Drug History

A first rough indicator was defined: the previous exposure to NRTI, NNRTI and PI classes was recorded as a binary variable, calculating if before any TCE the patient experienced a drug pressure in a specific class for more than one year.

In a more detailed way, instead, total time exposure and time elapsed from the last exposure could be calculated for each drug at any TCE. A combined function was then defined as:

$$f(t,s) = e^{-k\frac{s}{t}} \in [0,1] \tag{10.1}$$

where k is tuning constant, t is the time of exposure and s is the time elapsed from the last exposure, calculated for each TCE. The effect of this function was to give a measure of the effect of a therapy not only based on the current drugs taken, but also to take into account for how long a compound was taken previously and the time since it was not used.

10.3.3 Derived Fuzzy Scores

In this section a set of Fuzzy scores that give indications about viral resistance – associating mutations and drugs – will be defined, translating the existing medical knowledge bases. In [121] a set of practical rules for the treatment of resistant viral strains is given, based on summary statistics taken from the literature studies published so far about HIV drug resistance. These "rules" usually are of the form:

```
IF the viral genotype shows
three or more mutations
in the set {M41L, D67N, K70R, L210W, T215Y, T215F, T219E}
THEN resistance to AZT is high
```

The same holds for phenotypic tests: depending on the resistance fold change values obtained from in-vitro tests, using a patient's viral isolate, a set of compounds can be discarded for the next therapy. The set of parameters used, along with choice of membership functions or logical, aggregation operators is arbitrary: it's an attempt to translate into Fuzzy formulae the medical protocols. Procedures towards the automatic derivation or improvement of these rules are not carried on in this study. However, it will be shown in the results that they are significantly associated with the treatment outcomes.

Phenotype-Based Resistance and Efficacy Scores It is possible to design a fuzzy Membership Function (MF) which calculates the resistance $r \in [0, 1]$ as a sigmoid function of the Log Fold Change for phenotype p and technical cutoff** c.

$$r = \frac{1}{1 + e^{-(p-c)}} \tag{10.2}$$

Obviously the shape of the function could be optimised (for example estimating an optimal separation between two Gaussian), but usually cutoffs reported in literature are a reliable source.

- *pessimistic* resistance score: suppose that the resistance to a single drug concurs to the total resistance of a combined Anti Retroviral Therapy (cART) under an unknown function. From a probabilistic point of view, if the drugs were acting independently, the probability of overall cART resistance would be the product of the single resistance values. However this not the case: in fact many resistances are indeed cross-resistances, drugs interact and moreover the virus mutates through time just in relation to drug pressure and efficacy. Visual inspection of a product-based resistance score did not gave satisfiable results. Instead, let's suppose that actually

**the cutoff is the fold change value from which a viral isolate is considered to be completely resistant to the drug

the single resistances will *sum* (it's unrealistic, but simulates the fact that through time the cross-resistant strains win). It can be set up a fuzzy $\bigvee$ that calculates this overall pessimistic resistance to a generic cART T.

$$\rho(T) = \bigvee_{\forall d \in T} (r_d) = \bigvee_{\forall d \in T} \left(\frac{1}{1 + e^{-(p_d - c_d)}} \right) \tag{10.3}$$

where $\bigvee$ is the probabilistic sum (i.e. $a \vee b = a + b - ab$) of each r_d for any drug in the cART. If there is a boosted PI, the corresponding r value is multiplied for 0.85 (accounting for the fact that RTV boosting dose helps the PI adsorption).

- *optimistic* score for efficacy. The above indicator considers ineffective even a therapy with just one drug for which the resistance is 1. In reality this is not the case: usually a cART can suppress viral load below 500 copies if it is composed by three or four (active) compounds in three different inhibition classes (NRTI, NNRTI, PI). A common quadruple therapy is made by 2 NRTIs, 1 NNRTI and 1 PI. It is possible then to use this information to design an efficacy estimation for each drug class and for a cART: assume that to have complete efficacy in NRTI class at least two drugs must be taken, while in order to have complete overall cART efficacy at least three drugs must be active. The corresponding fuzzy formulae are:

$$\eta_{\text{NRTI}}(T) = min \left\{ 1, \frac{\sum (1 - r_d)}{2} \right\}, \quad \forall d \in \text{NRTI}, \quad d \in T \tag{10.4}$$

$$\eta_{\text{NNRTI}}(T) = min \left\{ 1, \sum (1 - r_d) \right\}, \quad \forall d \in \text{NNRTI}, \quad d \in T \tag{10.5}$$

$$\eta_{\text{PI}}(T) = min \left\{ 1, \sum (1 - r_d) \right\}, \quad \forall d \in \text{PI}, \quad d \in T \tag{10.6}$$

If RTV booster is present, 0.15 is added to η_{PI}. Finally the total efficacy $\eta(T)$ will be

$$\eta(T) = min \left\{ 1, \frac{\eta_{NRTI} + \eta_{NNRTI} + \eta_{PI}}{3} \right\} \tag{10.7}$$

Scores Based on Literature Resistance Mutations The calculation is similar to the above scores, but it relies on the resistance hypotheses published on the international guidelines [121]. The fuzzy MF is related to how many resistance mutations accumulate for each drug, including cross-resistance clusters. The resistance for a single compound (o for a cross-resistant cluster) is

$$r_d = min \left\{ \frac{n_d}{N_d}, 1 \right\} \tag{10.8}$$

where n_d is the number of resistance mutations accumulated and N_d is the number needed to achieve complete resistance.

- *pessimistic* resistance score: calculated in the same way as for the phenotype-based score

$$\rho(\mathrm{T}) = \bigvee_{\forall d \in T} (r_d) = \bigvee_{\forall d \in T} \left(min \left\{ \frac{n_d}{N_d}, 1 \right\} \right) \tag{10.9}$$

- *optimistic* efficacy score: same as for phenotype-based score.

10.3.4 Simulation of Viral Replication through Time

Many papers have been published regarding the simulation of viral replication through time (Perelson [11] uses Ordinary Differential Equations, without accounting for viral evolution). Here a simple equation was used, parametrised on a constant (trough 8 weeks) viral resistance ρ (which can be one of the scores above defined, being $\rho = 1 - \eta$)

$$\frac{dV}{dt} = \rho P - cV = \rho c V_{equil} - cV \tag{10.10}$$

which solution is $V = V_0 e^{-ct} + \rho V_{equil}(1 - e^{-ct}) \rightarrow V_{8Weeks} = \rho V_{equil}$, Of note, it does not take into account CD4 or CD8, neither gives ρ a time dependency (like monotonic decreasing): the approximation is simply a percentage reduction from the steady state equilibrium.

10.3.5 Second- and Third-Order Variable Interactions

An interesting issue is to derive again from the EuResist data base a set of significant associations among mutations and drugs to see if they resemble (confirm or reject) medical hypotheses.

From a rigorous learning perspective, it can be considered "unfair" to use derived indicators from existing knowledge bases, even if they have been estimated from independent studies. The analysis of second- and third-order interactions between variables can point out combinations between drugs and mutations that are significantly associated with treatment response.

The codification for higher-order interactions can be made by dummy variables. Specifically, the following were considered:

- mutation × drug
- drug × drug
- drug × drug × drug

Note that they take into account only "and" associations: in fact, a combined dummy variable is non-zero iff all the single are non-zero. Thus, more complex logical associations like "variable a and not b" or "variable c or d" are not modelled.

The space set generated, though, yielded an extremely large feature set (more than 12'000 not-null variables), due to the fact that in the Standard Data hundreds of mutations are observed with a frequency $> 4\%$ and the number of available drugs is 20 circa.

10.3.6 Distance Measures for HIV Data

Since the input space is not a simple vector in $\mathbf{R}^n$, in order to apply Instance Based Reasoning, a set of new ad-hoc distances was designed for each different input domain.
For vectors in $\mathbf{R}^n$ and binary (i.e. therapies, phenotypes and resistance vectors obtained by fuzzy functions) let's introduce the *Tanimoto* (or *cosine*) coefficient, which for binary vectors is also called *Jaccard*:

$$Tanimoto(a, b) = \frac{a \cdot b}{\|a\|^2 + \|b\|^2 - a \cdot b} \in [0, 1] \tag{10.11}$$

If the complementary of Tanimoto similarity is defined as $1 - Tanimoto(\cdot)$, a distance measure is obtained: it has been shown that this has higher discriminant power under text classification.
For variables in $\mathbf{R}$ (i.e. viral load Log, CD4+ counts) a Gaussian function is defined:

$$d(a, b) = e^{-\frac{|a-b|^2}{2\sigma^2}} \tag{10.12}$$

where σ is the standard deviation observed in the population.
For nominal variables (risk, gender, country of infection...) the following criteria is chosen:

$$d(A, B) = \begin{cases} 1 & \text{if } (A \neq B) \wedge (A \neq ?) \wedge (B \neq ?) \\ 0.5 & \text{if } (A = ?) \vee (B = ?) \\ 0 & \text{if } (A = B) \wedge (A \neq ?) \wedge (B \neq ?) \end{cases} \tag{10.13}$$

where ? stands for missing or unknown attribute value.

All the single distances are combined using the overall sum.

There is need to point out here that no theoretical justification for the choice of these distance function is here given. The choice of the Tanimoto coefficient rises from the fact that for high dimensional vectors the Euclidean distance suffers more the curse of dimensionality. The usage of a Gaussian function for the real variables tries to take into account the natural variability of clinical measures, which can be tuned through the σ value.

10.4 Results

10.4.1 Univariable Filter: Selection of Statistically Significant Features

A preliminary filter was applied to the whole data to select for significant features under univariable non-parametric analysis: higher-order interactions were considered present in the derived features or were directly coded into dummy variables (see section 10.3.5). Drug history and Fuzzy scores were all detected significant (the one based on phenotypes and the one from literature guidelines) along with almost all mutations listed in [121] when in presence of a corresponding drug known to be not active ($p < 0.05$). Without considering higher-order dummy variables (treated separately), from more than 400 variables, ˜200 were detected significant and fed to the machine learners. Table 10.5 reports first 60 highest results.

Higher-order interactions (drug × drug, drug × mutation, drug × drug × drug) generated more than 12'000 variables: of these, a thousand circa have been detected significant and were passed to embedded selection.

Different methods were experimented, using different feature sets, increasing the feature set space from a base model to more complex. Parameter tuning and model selection was made through multiple CV (10 runs).

10.4.2 Classification of 8th Week Follow Up Virological Success

Tables 10.6 and 10.7 show prediction performances for Logistic, Decision Trees, Random Forests and Instance Based Reasoning (best models under different scenarios) varying feature spaces.

The base models that included only genotypes and cARTs (or worse only cARTs) gave little improvements compared to a null model. The inclusion of baseline markers (viral RNA load and CD4+ cell counts) improved significantly the performances, but was still significantly worse than a model trained using derived features. Other attributes like epidemiological indicators (risk factors, route of infections) and drug histories were detected significant as well and improved slightly the performances. In fact, the presence of macro-indicators like the country of infection or the ethnicity was consistent with the source heterogeneity of the integrated DB.

Among the models that used derived features, relying only on one Fuzzy score (the one based on resistant mutations listed in literature) gave a compact model which was equally powerful as a model which included all significant attributes selected by univariable analysis. Regarding machine learning model choice, Decision Trees were not as powerful as Logistic Regression or Random Forests: due to its easier understandability, then, Logistic Regression could be preferred.

The local IBR behaves similarly: the wrapper feature selection method and ad-hoc distances are able to yield a better local approximation than the base (cART + mutations) model. Moreover, in the next paragraph it will be shown that performance rate can be enhanced using an enlarged Case Base.

Table 10.8 instead presents results obtained using dummy variables, coding higher-order interactions between drugs and mutations. No derived features were included in the attribute space, since the objective was to discover again discriminant associations. Accuracy and AUC were even higher than the values obtained using the Fuzzy scores, moreover the feature selection yielded a compact model. This was an extremely interesting result, that demonstrated how an accurate investigation of the space state can lead to optimal results.

For completeness, two compact Logistic models are reported in tables 10.9, 10.11 and 10.10: they used Fuzzy scores and higher-order interactions respectively. Significance p_values are shown under multivariate analysis. As expected, Fuzzy score and baseline markers are significant also in the multivariate. The coefficients reported for drugs don't have to be misinterpreted, because of either the intercept value or the fact that these dummy variables are not calculated with respect to any therapeutic backbone. Actually, immediate interpretation of variable importance is not easy: for example, mutation PR_33_F takes always a role in the second model, whereas its function would be intrinsically dependent on the usage of a PI. In the model that uses the Fuzzy resistance score, other variables can be viewed as adjustments, while in the mixed effect model the interpretation is tougher.

Table 10.12 finally gives error estimation on the test set. Results resemble the errors estimated through CV, thus proving the robustness of parameter optimisation techniques and validation procedures.

Classification of 8th Week Follow Up Virological Success using Multiple Observations and Adjusted Validation

As described in section 10.2, a view that includes multiple observation for a TCE can assure a more confident estimation of the 8th week follow up, partially solving the viral "blips" problems. This is useful for the usage of a local estimator. In order to avoid positively biased results, coming from the classification of a TCE with itself, the cross validation procedure has been corrected, avoiding the presence of the same TCE either in the fold-test or the training set. Moreover, all the TCE in the test set have been completely removed from the reference case base, providing the same test domain as for the previous models. Table 10.13 shows results coming from the wrapper optimisation via Bubble selection heuristic. With multiple observations, IBR reached the best ranked Logistic model.

10.4.3 Regression of 8th Week Follw Up Viral RNA Load

It has to be pointed out here that the regression problem is far more complex than the classification one, due to the high variability of RNA measures within individuals (± 0.5 Logs even in a small time window) and to detection limits of instruments (some have a resolution of 500 cp/ml, some other of 50 cp/ml). While the classification problem divides between detectable levels of HIV-RNA and undetectable levels (so, below higher detection limit), the regression problem faces the actual viral RNA charge prediction, with the bias of saturating all the undetectable instances (i.e. all the successes) at 500 copies. These premises reduce the chances to achieve satisfying results, even with an accurate application of the ML techniques described so far.

Attribute selection and feature derivation has been carried on in the same way as for classification problem. Bagged Decision Trees and IBR resulted to be the best models. The considerations about feature sets are the same as described for the classification problem: derived features improve indeed prediction performances significantly, but the correlation values cannot go above 0.6.

Table 10.14 show results for Bagged DTs, while table 10.15 for IBR.

10.5 Conclusions

Discussion In this study a wide range of Feature Selection and Feature Extraction/Derivation techniques was applied to real clinical and biological HIV data. Different ML methods (local and global) were applied and statistically compared in order to build prediction models either for in-vitro or in-vivo problems.

For the specific problem of *phenotype* prediction, MLR models were proven to be competitive with the State of the Art (SoA) techniques, but preferable due to their compactness (achieved thanks to Feature Selection) and understandability.

Regarding *in-vivo* therapy success classification and eager learning, it was shown that baseline markers, epidemiological and historical information do improve performances over the naive cART + genotype model (at now SoA). Actually, the inclusion of derived features lifts further performances, still significantly ($p < 0.05$). Of note, one fuzzy score is based on the existing medical knowledge and another one on the in-vitro phenotypes: this means that at now the rules used by physicians are indeed effective and that the phenotypic testing (or virtual phenotyping through genotype-based models) is a valid help in clinical practice. The analysis of higher-order interactions among variables provided then completely data-driven models able to compete with the Fuzzy-enhanced ones.
Lazy IBR, a completely non-linear local approximation, showed equally good behaviour with the designed distance functions, better with the usage of multiple observations. Considering that the Eu*Resist* integrated DB is continuously increasing in size, yielding a denser space, IBR is an acceptable alternative if there

is no need to discover an approximating function (i.e. a model).

It's important also to address the impact of the models presented for the scientific community: phenotype predictor is not a new discovery – geno2pheno [55] is working good since a couple of years – but the simplicity of MLR is a preferable alternative. The most important results are the models for in-vivo follow up classification. In a therapy optimisation perspective, the model can be run with a set of allowed cARTs (presumably HAARTs) and give a *rank* of suitable therapies with the corresponding probability of success. Moreover, this study is investigating one out of four prediction engines that are being designed at the same time under the Eu*Resist* project, with a further aim of model integration.

From a pure analysis of model power, accuracy of ˜76% and correlation of ˜0.6 can be not satisfactory: however, these results are the best ever obtained in the field (using the approved standard TCE datum policy) and are supported by the largest data base in the world. Plasma analyses are just an indirect way to observe the reality and the variability of the body response adds noise to the system (until it remains unknown). In this study the whole set of clinical, biological and epidemiological attributes currently retrievable has been explored (except for HLA information, i.e. the human genome...), so it could be that these performances are really an upper bound: there is the chance to design more efficient features or to model better the unobserved variables that at now act as stochastic fluctuations.

Future Perspectives Since the fuzzy indicators derived from literature guidelines showed to be significantly associated with prediction improvement, the next step can be to use the existing decision algorithms (REGA, Stanford [8] [13]) as additional indicators and fed them to the MLs. In addition, the importance of higher-order dummy variables, explaining conditional dependence between variables, suggests that automated reasoning (like association rule discovery) could provide new findings: if it is performed on an independent set, these rules can be added as well to the feature space.

avg-chi-square	st.dev	var.name
296.259	7.95	fuzzyLiteratureResistanceMutScore optimistic
291.145	10.754	diffEquaSolution fuzzyLiteratureResistanceMutScore pessimistic
255.836	8.056	LPV pheno logFC
249.535	6.336	literatureResMut LPV
243.712	10.444	RTV pheno logFC
236.108	7.943	literatureResMut APV
232.1	7.57	APV pheno logFC
231.898	7.646	literatureResMut RTV
221.861	10.196	IDV pheno logFC
218.863	11.915	ATV pheno logFC
212.272	8.229	literatureResMut IDV
204.644	14.473	D4T pheno logFC
198.759	8.302	SQV pheno logFC
192.762	9.492	diffEquaSolution fuzzyLiteratureResistanceMutScore optimistic
190.274	7.095	TPV pheno logFC
184.509	8.375	fuzzyLiteratureResistanceMutScore pessimistic
183.171	8.959	literatureResMut ATV
180.86	14.426	fuzzyPhenoResistanceScore optimistic
179.869	12.001	RT 69 INS Complex
179.405	11.748	AZT pheno logFC
173.229	5.561	ABC pheno logFC
167.79	7.516	literatureResMut SQV
167.4	7.65	literatureResMut NFV
166.387	7.372	TDF pheno logFC
158.203	14.23	TAMs
158.203	14.23	literatureResMut D4T
158.203	14.23	literatureResMut AZT
155.202	7.04	diffEquaSolution fuzzyPhenoResistanceScore pessimistic
152.811	9.364	AZT combinedUsage
151.366	6.248	DDI pheno logFC
149.529	6.751	NFV pheno logFC
148.915	8.057	AZT totalTimeExposure
148.322	8.78	NRTI experienced
145.173	9.129	AZT sinceNotUsed
142.947	8.442	fuzzyPhenoResistanceScore pessimistic
142.487	8.525	SQV totalTimeExposure
141.581	9.596	drugHistoryScore
134.62	12.707	FTC pheno logFC
132.388	8.204	SQV combinedUsage
131.665	18.016	3TC pheno logFC
130.71	9.036	SQV sinceNotUsed
126.221	7.919	DDI totalTimeExposure
122.145	8.323	DDI sinceNotUsed
122.061	7.731	NVP pheno logFC
120.81	8.399	literatureResMut TPV
117.247	6.421	3TC totalTimeExposure
116.695	5.804	D4T combinedUsage
114.937	4.573	D4T sinceNotUsed
114.661	5.377	D4T totalTimeExposure
113.592	6.785	3TC sinceNotUsed
112.228	7.11	diffEquaSolution fuzzyPhenoResistanceScore optimistic
111.792	7.854	DDC sinceNotUsed
111.611	7.974	DDC totalTimeExposure
110.773	6.352	3TC combinedUsage
109.375	6.606	DDC pheno logFC
108.697	7.35	RNA equil log
106.418	8.417	DDI combinedUsage
105.804	7.624	RTV totalTimeExposure
103.621	7.633	PI experienced
102.576	6.828	DDC combinedUsage

Table 10.5: univariable chi-square analysis

model	feature set	accuracy % (st.dev)	AUC of ROC (st.dev)
majority class	none	67.08 (0.18)*	0.50 (0.02)*
Logistic	cART	67.45 (1.89)*	0.65 (0.05)*
Logistic	cART + IAS mutations	72.90 (2.51)*	0.71 (0.04)*
Logistic	cART + IAS mutations + baseline markers	72.94 (2.53)*	0.75 (0.04)*
Logistic	cART + fuzzy resistance scores + baseline markers + drug class exposure + epidemiological	75.45 (2.89)	0.77 (0.05)
Decision Tree	all filtered attributes (including derived)	72.64 (2.82)*	0.68 (0.04)*
Logistic	all filtered attributes (including derived)	75.13 (2.16)	0.77 (0.03)
Random Forest (30 trees)	all filtered attributes (including derived)	75.44 (2.26)	0.77 (0.03)

Table 10.6: Classification of 8th week follow up success - Multiple CV Results for Logistic, DT and RF (10 independent runs). Values with * are significantly worse than best models ($p < 0.05$)

feature set	$\hat{k}$	kernel	accuracy % (st.dev)
cART	8	TriCube	66.18 (0.36)*
cART + IAS mutations	16	Linear	70.88 (0.23)*
cART + IAS mutations + baseline markers	16	Linear	72.98 (0.18)*
gender, subtype, consensusB match, no of drugs, mutations, cART, time exposure to drugs, phenotype, literature resistance mutation clusters, baseline markers, equilibrium markers	30	Linear	75.99 (0.14)

Table 10.7: Classification of 8th week follow up success - Multiple CV Results for Instance Based Reasoning (10 independent runs). Kernel and $\hat{k}$ optimised by multiple CV accuracy maximisation; feature sets manually tuned and selected by *Bubble* search. Values with * are significantly worse than best models ($p < 0.05$)

model	feature set	accuracy % (st.dev)	AUC of ROC (st.dev)
Logistic	all filtered attributes (higher-order interactions)	76.61 (2.40)	0.79 (0.03)

Table 10.8: Classification of 8th week follow up success - Multiple CV Results for Logistic (10 independent runs) using higher-order dummy variables.

variable	estimate	std.err	z.value	Pr(> |z|)	signif.
(Intercept)	6.888	1.151	5.984	0.000	***
consensusB match	0.001	0.000	1.506	0.132	
age	0.006	0.009	0.636	0.525	
risk = vertical transmission	-2.019	0.663	-3.046	0.002	**
country of origin = ITA	-0.249	0.178	-1.398	0.162	
subtype = D	1.710	1.085	1.577	0.115	
3TC	0.310	0.219	1.413	0.158	
AZT	0.706	0.213	3.312	0.001	***
RTV booster	0.793	0.250	3.169	0.002	**
D4T	-0.145	0.211	-0.685	0.493	
DDC	0.490	0.959	0.512	0.609	
DDI	0.440	0.222	1.984	0.047	*
FTC	0.565	0.382	1.479	0.139	
LPV	-0.099	0.195	-0.507	0.612	
SQV	-0.993	0.323	-3.077	0.002	**
NRTI experienced	-0.557	0.243	-2.287	0.022	*
NNRTI experienced	-0.358	0.233	-1.540	0.124	
PI experienced	-0.222	0.235	-0.945	0.345	
no. of drugs	-0.096	0.153	-0.625	0.532	
RNA equilibrium Log	-0.745	0.138	-5.397	0.000	***
CD4 equilibrium	0.001	0.000	3.905	0.000	***
RNA baseline Log	-0.218	0.104	-2.094	0.036	*
Fuzzy score on resist. mutations	-4.710	0.509	-9.254	0.000	***

Table 10.9: Logistic model built after univariable filter, CFS and ridge shrinkage, with derived Fuzzy scores. Coefficients for the success prediction. Null deviance: 1398.0 on 1181 degrees of freedom. Residual deviance: 1042.0 on 1159 degrees of freedom. AIC: 1088. Signif. codes: '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

variable	estimate	std.err	z.value	Pr(> |z|)	signif.
(Intercept)	2.287	0.692	3.305	0.001	***
PR 54 V	-0.442	0.330	-1.340	0.180	
NRTI experienced	-0.303	0.169	-1.795	0.073	.
PR 82 A	-0.461	0.275	-1.672	0.094	.
RNA equil log	-0.601	0.091	-6.589	0.000	***
PR 10 I	-0.304	0.160	-1.901	0.057	.
PR 84 V	-0.620	0.273	-2.276	0.023	*
no. of drugs	0.582	0.080	7.291	0.000	***
RNA baseline Log	-0.201	0.070	-2.883	0.004	**
PR 33 F	-1.235	0.395	-3.127	0.002	**
SQV and PR 54 V	-2.020	1.138	-1.775	0.076	.
PI experienced	-0.248	0.167	-1.482	0.138	
RT 210 W	-0.164	0.239	-0.683	0.494	
PR 90 M	0.144	0.188	0.765	0.444	
PR 46 I	0.003	0.221	0.016	0.988	
age	0.014	0.006	2.183	0.029	*
DDI and ABC and TDF	-2.739	0.860	-3.184	0.001	**
RT 215 Y	-0.224	0.225	-0.993	0.321	
CD4 equilibrium	0.000	0.000	3.238	0.001	**
subtype = D	1.858	0.644	2.883	0.004	**
AZT and LPV	0.232	0.425	0.547	0.584	
APV and PR 10 I	-1.443	0.456	-3.164	0.002	**
D4T and RT 215 Y	0.001	0.320	0.003	0.998	
D4T and DDI	-0.271	0.185	-1.467	0.143	
AZT and 3TC and LPV	0.479	0.487	0.983	0.326	
NVP and RT 184 V	-0.842	0.303	-2.780	0.005	**
RT 41 L	-0.076	0.187	-0.405	0.686	
RT 67 N	0.013	0.171	0.076	0.940	

Table 10.10: Logistic model built after univariable filter and CFS. Coefficients for the success prediction. Null deviance: 2875.4 on 2268 degrees of freedom. Residual deviance: 2185.8 on 2214 degrees of freedom. AIC: 2295.8. Signif. codes: '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

variable	estimate	std.err	z.value	Pr(> |z|)	signif.
D4T and RT 210 W	-0.316	0.381	-0.830	0.407	
LPV and PR 54 V	-0.620	0.406	-1.528	0.126	
APV and PR 84 V	-1.852	1.172	-1.581	0.114	
NFV and PR 82 A	-1.584	1.122	-1.411	0.158	
RTV and PR 64 L	-2.281	1.219	-1.872	0.061	.
consensusB match	0.001	0.000	1.685	0.092	.
PR 10 F	-0.786	0.349	-2.253	0.024	*
IDV and PR 90 M	-1.267	0.449	-2.824	0.005	**
DDC and RT 215 Y	-1.002	0.846	-1.185	0.236	
NFV and PR 46 I	-1.921	1.123	-1.711	0.087	.
DDC and 3TC	-13.130	1455.000	-0.009	0.993	
DDC and 3TC and NVP	-2.772	1672.000	-0.002	0.999	
D4T and ABC and NVP	-15.510	784.500	-0.020	0.984	
APV and PR 16 E	14.420	397.800	0.036	0.971	
NNRTI experienced	-0.312	0.165	-1.895	0.058	.
D4T and RT 67 N	0.036	0.269	0.133	0.894	
RT 184 V	-0.313	0.123	-2.539	0.011	*
NVP and RT 215 Y	-0.080	0.337	-0.238	0.812	
FTC and TDF	0.345	0.261	1.321	0.186	
RTV and PR 54 V	0.369	1.217	0.303	0.762	
SQV and PR 62 V	-0.792	0.455	-1.740	0.082	.
DDI and RT 181 C	-0.753	0.264	-2.850	0.004	**
EFV and RT 103 N	-1.527	0.542	-2.816	0.005	**
AZT and RT 70 R	-1.441	0.315	-4.578	0.000	***
LPV and PR 30 N	1.103	0.584	1.888	0.059	.
D4T and DDI and RTV	-15.030	585.900	-0.026	0.980	
D4T and DDI and APV	-1.950	1.014	-1.924	0.054	.

Table 10.11: Logistic model built after univariable filter and CFS using higher-order interactions (*continued*). Coefficients for the success prediction. Null deviance: 2875.4 on 2268 degrees of freedom. Residual deviance: 2185.8 on 2214 degrees of freedom. AIC: 2295.8. Signif. codes: '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

model	feature set	accuracy %	AUC of ROC
majority class	none	66.93	0.50
Logistic	cART	68.50	0.60
Logistic	cART + IAS mutations	73.62	0.70
Logistic	cART + IAS mutations + baseline markers	72.83	0.72
Logistic	cART + fuzzy resistance scores + baseline markers + drug class exposure + epidemiological	78.35	0.77
Decision Tree	all filtered attributes (including derived)	71.26	0.74
Logistic	all filtered attributes (including derived)	79.13	0.79
Random Forest (30 trees)	all filtered attributes (including derived)	76.77	0.75
Logistic	all filtered attributes (higher-order interactions)	76.77	0.77
IBR ($\hat{k} = 8$, TriCube ker.)	cART	65.86	not.av
IBR ($\hat{k} = 16$, Linear ker.)	cART + IAS mutations	68.67	not.av
IBR ($\hat{k} = 16$, Linear ker.)	cART + IAS mutations + baseline markers	71.08	not.av
IBR ($\hat{k} = 30$, Linear ker.)	gender, subtype, consensusB match, no of drugs, mutations, cART, time exposure to drugs, phenotype, literature resistance mutation clusters, baseline markers, equilibrium markers	75.98	0.73

Table 10.12: Classification of 8th week follow up success - Test Set Error for Models (n=254)

IBR			Multiple CV	Test Set	
feature set	$\hat{k}$	kernel	accuracy % (st.dev)	accuracy %	AUC
country of infection, ethnicity, mutations, IAS mutations, cART, compound time exposure, time elapsed since last compound usage, phenotype, baseline markers, equilibrium markers	30	Linear	76.52 (0.53)	77.11	0.75

Table 10.13: Classification of virologic success at week 8 - Multiple Observations data set - Adjusted Multiple CV Results for Instance Based Reasoning (10 independent runs) and Test Set evaluation. Kernel and $\hat{k}$ optimised by multiple CV accuracy maximisation; best feature set selected by *Bubble* search.

75 Bagged Decision Trees		
Multiple 10-fold CV (10 independent runs)		
feature set	correlation (st.dev)	RMSE (st.dev)
cART + IAS mutations + baseline markers	0.56 (0.06)*	0.67 (0.05)*
all filtered attributes (including derived)	0.55 (0.07)*	0.68 (0.05)*
fuzzy scores + baseline markers + epidemiological	0.57 (0.07)	0.66 (0.05)
Test Set Evaluation		
feature set	correlation	RMSE
cART + IAS mutations + baseline markers	0.5356	0.684
all filtered attributes (including derived)	0.4967	0.7017
fuzzy scores + baseline markers + epidemiological	0.5648	0.6675

Table 10.14: Regression of actual viral load at week 8 - Results for Bagged DTs. Values with * are significantly worse than best model ($p < 0.05$).

IBR			Multiple CV	Test Set
feature set	$\hat{k}$	kernel	correlation (st.dev)	correlation
cART	16	Linear	0.159 (0.007)	0.288
cART + IAS mutations	32	TriCube	0.326 (0.007)	0.365
cART + IAS mutations + baseline markers	32	Linear	0.537 (0.003)	0.587
age, ethnicity, subtype, consensusB match, no. of drugs, mutations, IAS mutations, cART, time elapsed since last compound usage, phenotype, baseline markers, equilibrium markers	28	Linear	0.590 (0.003)	0.569

Table 10.15: Regression of actual viral load at week 8 - Multiple CV Results for Instance Based Reasoning (10 independent runs) and Test Set evaluation (n=249). Kernel and $\hat{k}$ optimised by multiple CV correlation maximisation; feature sets manually tuned and selected by *Bubble* search.

Bibliography

[1] Mocroft A. Changing patterns of mortality across europe in patients infected with hiv-1 eurosida study group. *Lancet*, 352:1725–1730, 1998. [cited at p. 139]

[2] Prosperi MCF Zazzi M Ulivi G Di Giambenedetto S De Luca A. Modelling in-vivo hiv evolutionary mutational pathways under azt-3tc regimen through markov chains. BioSapiens-viRgil Workshop on Bioinformatics for Viral Infections, Caesar Bonn, Germany, September 2005. [cited at p. 73]

[3] Desper R Jiang F Kallioniemi OP Moch H Papadimitriou CH Shaffer AA. Inferring tree models for oncogenesis from comparative genome hybridization data. *J Comp Biol*, 6(1):37–51, 1999. [cited at p. 85]

[4] H. Wu A.A. Ding. Population hiv-1 dynamics in vivo: Applicable models and inferential tools for virological data from aids clinical trials. *Biometrics*, 55(2):410–418, 1999. [cited at p. 31]

[5] H. Wu A.A. Ding. A comparison study on models and fitting procedures for biphasic viral dynamics in hiv-1 infected patients treated with antiviral therapies. *Biometrics*, 56(1):293–300, 2000. [cited at p. 31]

[6] European Medicines Agency:. http://www.emea.europa.eu/. [cited at p. 143]

[7] D Aha and D Kibler. Instance-based learning algorithms. *Machine Learning*, 6:37–66, 1991. [cited at p. 114, 129, 145]

[8] Van Laethem K De Luca A Antinori A Cingolani A Perno CF Vandamme AM. A genotypic drug resistance interpretation algorithm that significantly predicts therapy response in hiv-1-infected patients. *Antivir Ther*, 7(2):123–129, Jun 2002. [cited at p. 140, 155]

[9] ANRS:. http://www.hivfrenchresistance.org/. [cited at p. 140]

[10] Pozniak A.L. Gallant J.E. Dejesus E. Gazzard B. Campo R.E. Chen S.S. McColl D. Holmes C.B. Enejosa J. Toole J.J. Cheng A.K. Arribas, J.R. Tenofovir disoproxil fumarate, emtricitabine, and efavirenz compared with zidovudine/lamivudine and efavirenz in treatment-naive patients: 144-week analysis. *J Acquir Immune Defic Syndr*, 47(1):74–8, Jan 1 2008. [cited at p. 39]

[11] Perelson AS and Nelson PW. Mathematical analysis of hiv-1 dynamics in vivo. *SIAM Review*, 41(1):3–44, 1999. [cited at p. 31, 34, 150]

[12] Kuroda MJ Schmitz JE Santra S Peyerl FW Krivulka GR Beaudry K Lifton MA Gorgone DA Montefiori DC Lewis MG Wolinsky SM Letvin NL Barouch DH, Kunstman J. Eventual aids vaccine failure in a rhesus monkey by viral escape from cytotoxic t lymphocytes. *Nature*, 415(6869):335–9, Jan 17 2002. [cited at p. 20]

[13] Stanford HIV Drug Resistance Data Base:. http://hivdb.stanford.edu/. [cited at p. 140, 155]

[14] Sabin CA Portsmouth S Hill T Johnson M Gilson R Easterbrook P Gazzard B Fisher M Orkin C Dunn D Delpech V Taylor GP Walsh JC Phillips AN Benzie AA, Bansi LK. Increased duration of viral suppression is associated with lower viral rebound rates in patients with previous treatment failures. *AIDS*, 21(11):1423–30, Jul 11 2007. [cited at p. 23]

[15] KV Mardia JT Kent JM Bibby. *Multivariate Analysis*. Academic Press, London, 1979. [cited at p. 47, 122]

[16] Doranz BJ. A dual-tropic primary hiv-1 isolate that uses fusion and the βchemokine receptors ckr-5, ckr-3 and ckr-2b as fusion cofactors. *Cell*, 85(7):1149–1158, 1996. [cited at p. 123]

[17] F. Castiglione and M. Bernaschi. Hiv-1 strategies of immune evasion. *Int J Mod Phys C*, 16(12):1869–1879, 2006. [cited at p. 32, 42]

[18] Poccia F. D'Offizi G. Castiglione, F. and M. Bernaschi. Mutation, fitness, viral diversity and predictive markers of disease progression in a computational model of hiv-1 infection. *AIDS Research and Human Retrovirus*, 20(12):1316–1325, 2004. [cited at p. 32, 34, 42]

[19] Kirkpatrick S Gelatt CD and Vecchi MP. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983. [cited at p. 120]

[20] Ruiz L Mocroft A Phillips AN Olsen CH Gatell JM Gunthard HF Reiss P Perno CF Clotet B Lundgren JD Ceccherini-Silberstein F, Cozzi-Lepri A. Impact of hiv-1 reverse transcriptase polymorphism f214l on virological response to thymidine analogue-based regimens in antiretroviral therapy (art)-naive and art-experienced patients. *J Infect Dis*, 196(8):1180–90, Oct 15 2007. Epub 2007 Sep 19. [cited at p. 60]

[21] B. Schlkopf Chapelle, O. and A. Zien. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006. [cited at p. 122]

[22] M. Chen. Modelling and simulation of metabolic networks: Petri nets approach and perspective. In Amorski K. et al., editor, *Modelling and Simulation: ESM2002*, pages 441–444, Darmstadt, 2002. [cited at p. 80]

[23] Breiman L Friedman JH Olshen RA Stone CJ. *Classification and Regression Trees*. Belmont, California, wadsworth and brooks international group edition, 1984. [cited at p. 114, 119, 128, 145]

[24] P Comon. Independent component analysis, a new concept? *Elsevier Signal Processing*, 36(3):287–314, April 1994. Special issue on Higher-Order Statistics. [cited at p. 122]

[25] T.F. Cox and M.A.A. Cox. *Multidimensional Scaling*. Chapman & Hall, 1994, 2001. [cited at p. 47, 50]

[26] A. Cozzi-Lepri. Initiatives for developing and comparing genotype interpretation systems: external validation of existing rule-based interpretation systems for abacavir against virological response. *HIV Medicine*, 2007. in press. [cited at p. 141]

[27] Loveday C Phillips AN Clotet B Reiss P Ledergerber B Holkmann C Staszewski S Lundgren JD Cozzi-Lepri A, Ruiz L. Thymidine analogue mutation profiles: factors associated with acquiring specific profiles and their impact on the virological response to therapy. *Antivir Ther*, 10(7):791–802, 2005. [cited at p. 50]

[28] Dempster A Laird N Rubin D. Maximum likelihood from incomplete data via the em algorithm. *J R Statist Soc B*, 39:1–38, 1977. [cited at p. 87]

[29] The AREVIR data base:. http://www.mpi-inf.mpg.de/∼niko/arevir/. [cited at p. 92]

[30] Cozzi-Lepri A. Perno C.F. Balotta C. Di Giambendetto S. Orani A. Mussini C. Toti M. D'Arminio Monforte A. De Luca, A. The prognostic value to predict virological outcomes of 14 distinct systems used to interpret the results of genotypic hiv-1 drug resistance testing in untreated patients starting their first haart. *Proc. of the 1st European HIV Drug Resistance Workshop, HIV Medicine*, 4:20, 2003. [cited at p. 39]

[31] Ellmeier W et al. Deng H, Liu R. Identification of a major co-receptor for primary isolates of hiv-1. *Nature*, 381:661–6, 1996. [cited at p. 16]

[32] Eckert DM. Mechanisms of viral membrane fusion and its inhibition. *Annu Rev Biochem*, (70):777–810, 2001. [cited at p. 123]

[33] Berger EA Doms RW Fenyo EM Korber BT Littman DR. A new classification for hiv-1. *Nature*, 391:290, January 1998. doi:10.1038/34571. [cited at p. 123]

[34] Jang JSR Sun CT Mizutani E. *Neuro-Fuzzy and Soft Computing*. Prentice Hall, 1997. [cited at p. 120]

[35] Landwehr N Hall M Frank E. *Logistic Model Trees*. Kluwer Academic Publishers, Netherlands, 2004. [cited at p. 114, 119, 129, 145]

[36] G. Borgulya R. D'Autilia F. Incardona R. Kaiser C. Kent T. Lengauer H. Neuvirth Y. Peres A. Petroczi M.C.F. Prosperi M. Rosenzvi E. Schulter T. Sing A. Sonnenborg R. Thompson M. Zazzi E. Aharoni, A. Altmann. Integration of viral genomics with clinical data to predict response to anti-hiv treatment. IST-Africa 2007 Conference & Exhibition. May, 9–11, Maputo, Mozambique, 2007. [cited at p. 140]

[37] Alkhatib G et al. Cc ckr5: A rantes, mip-1a, mip-1b receptor as a fusion cofactor for macrophage-tropic hiv-1. *Science*, 272:1955–1958, June 28 1996. [cited at p. 138]

[38] DeLuca A Ulivi G et al. Construction, training and clinical validation of an interpretation system for genotypic hiv-1 drug resistance based on fuzzy rules revised by virological outcomes. *Antiviral Therapy*, 9(4), 2004. [cited at p. 140]

[39] F Arenzana-Seisdedos et al. Hiv blocked by chemokine antagonist. *Nature*, 383:400, October 3 1996. [cited at p. 138]

[40] Montaner J et al. Anns are more accurate predictors of virologic response to arv therapy than rules-based genotype interpretation systems. CROI, Denver, CO, USA, February 2006. [cited at p. 140]

[41] Sing T et al. Immunologic markers improve genotypic prediction of hiv-1 coreceptor usage. 4th European HIV Drug Resistance Workshop, Monaco, Monte Carlo, 2006. [cited at p. 124, 128, 130]

[42] Molecular Evolution and Phylogenetics. *Masatoshi Nei, Sudhir Kumar*. Oxford University Press, Jun 2000. ISBN13: 978-0-19-513585-5, ISBN10: 0-19-513585-7. [cited at p. 27]

[43] MM Santoro M Prosperi F Forbici A Bertoli C Mussini A De Luca G Palamara P Narciso A Antinori C Balotta A dArminio Monforte F Ceccherini-Silberstein, V Svicher and CF Perno. Impact of hiv-1 reverse transcriptase polymorphism r83k on virological response in drug-naive patients starting thymidine-analogue-containing haart. XVI International HIV Drug Resistance Workshop 12-16 June 2007, Barbados - Antiviral Therapy 2007; 12:S77. [cited at p. 51]

[44] Ge D Colombo S Ledergerber B Weale M Zhang K Gumbs C Castagna A Cossarizza A Cozzi-Lepri A De Luca A Easterbrook P Francioli P Mallal S Martinez-Picado J Miro JM Obel N Smith JP Wyniger J Descombes P Antonarakis SE Letvin NL McMichael AJ Haynes BF Telenti A Goldstein DB Fellay J, Shianna KV. A whole-genome association study of major determinants for host control of hiv-1. *Science*, 318(5849):390–391, Oct 19 2007. [cited at p. 11, 17]

[45] J Felsenstein. *Inferring Phylogenies*. Sinauer, 2004. [cited at p. 27]

[46] Y Feng. Hiv-1 entry cofactor: Functional cdna cloning of a seven-transmembrane, g protein-coupled receptor. *Science*, 272(5263):809–810, 1996. [cited at p. 123]

[47] Collier A.C. Mukherjee A.L. Feinberg J.E. Demeter L.M. Tebas P. Giuliano M. Dehlinger M. Garren K. Brizz B. Bassett-R. Fischl, M.A. Randomized open-label trial of two simplified, class-sparing regimens following a first suppressive three or four-drug regimen. *AIDS*, 21(3):325–33, Jan 30 2007. [cited at p. 40]

[48] U.S. Food and Drug Administration:. http://www.fda.gov/oashi/aids/virals.html. [cited at p. 143]

[49] HIV forum:. www.hivforum.org. [cited at p. 141]

[50] C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97:611–631, 2002. [cited at p. 47, 122]

[51] Sander O Sing T Sommer I Low AJ Cheung PK Harrigan PR Lengauer T Domingues FS. Structural descriptors of gp120 v3 loop for the prediction of hiv-1 coreceptor usage. *PLoS Computational Biology*, 3(3):e58, March 2007. [cited at p. 124, 126, 130, 131, 138]

[52] Yun-Xin Fu. Estimating mutation rate and generation time from longitudinal samples of dna sequences. *Molecular Biology and Evolution*, 18:620–626, 2001. [cited at p. 37]

[53] Schwarz G. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978. [cited at p. 89, 115, 116]

[54] geno2pheno[coreceptor] web-service tool:. http://coreceptor.bioinf.mpi-inf.mpg.de/. [cited at p. 124, 131]

[55] The geno2pheno[resistance] web-service tool:. http://www.geno2pheno.org/cgi-bin/geno2pheno.pl. [cited at p. 146, 155]

[56] Kohavi R John GH. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, 1997. [cited at p. 117]

[57] S.M. Goodreau. Assessing the effects of human mixing patterns on human immunodeficiency virus-1 interhost phylogenetics through social network simulation. *Genetics*, 172:20332045, April 2006. DOI: 10.1534/genetics.103.024612. [cited at p. 27]

[58] Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716723, 1974. [cited at p. 115, 129]

[59] Choe H. The β-chemokine receptors ccr3 and ccr5 facilitate infection by primary hiv-1 isolates. *Cell*, 85(7):1135–1148, 1996. [cited at p. 123]

[60] Deng H. Identification of a major coreceptor for primary isolates of hiv-1. *Nature*, 381(6584):661–666, 1996. [cited at p. 123]

[61] Shimodaira H. Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling. *Annals of Statistics*, 32:2616–2641, 2004. [cited at p. 47, 128]

[62] Peter J. Haas. *Stochastic Petri Nets: Modelling, Stability, Simulation*. operation research. Springer, 2002. [cited at p. 80]

[63] RDI HIV Resistance Response Database Initiative:. http://www.hivrdi.org/. [cited at p. 140]

[64] DeJong J. Human immunodeficiency virus type 1 clones chimeric for the envelope v3 domain differ in syncytium formation and replication capacity. *J Virol*, 66(2):757765, 1992. [cited at p. 124]

[65] Edmonds J. Optimum branchings. *J Research of the National Bureau of Standards*, 71(B):233–240, 1967. [cited at p. 85]

[66] Hastie T Tibshirani R Friedman J. *The Elements of Statistical Learning*. Springer, New York, 2001. [cited at p. 46, 47, 62, 88, 89, 114, 115, 116, 119, 120, 122, 129]

[67] Platt J. Fast training of support vector machines using sequential minimal optimization. In Schoelkopf B Burges C and Smola A, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1998. [cited at p. 114, 128]

[68] Hartigan JA. *Clustering Algorithms*. Wiley, New York, 1975. [cited at p. 47, 122, 128]

[69] MacQueen JB. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, Berkeley, 1967. University of California Press. [cited at p. 46]

[70] Thompson JD. Clustal w: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 1994. [cited at p. 142]

[71] K. Jensen. Coloured petri nets. basic concepts, analysis methods and practical use. In Springer-Verlag, editor, *Monographs in Theoretical Computer Science, vol. 1*. 1997. ISBN: 3-540-60943-1. [cited at p. 80]

[72] Korber B Foley BT Kuiken C Pillai SK Sodroski JG. Numbering positions in hiv relative to hxb2cg. *Human Retroviruses and AIDS*, III:102–111, 1998. [cited at p. 127]

[73] Jefferys WH Berger JO. Ockham's razor and bayesian analysis. *American Scientist*, 80:64–72, 1992. [cited at p. 90]

[74] Quinlan JR. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1993. [cited at p. 114, 119, 128, 145]

[75] Brettler DB Sullivan JL Desrosiers RC Kirchhoff F, Greenough TC. Brief report: Absence of intact nef sequences in a long-term survivor with nonprogressive hiv-1 infection. *N Engl J Med*, 332:228–232, 1995. [cited at p. 11]

[76] Breiman L. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996. [cited at p. 114, 145]

[77] Breiman L. Random forests. *Machine Learning*, 45(1):5–32, October 2001. [cited at p. 114, 128, 145]

[78] Mazzoldi A. Andreoni C. Bianchi M. Cavallini A. Laurino M. Ricotti L. Iuliano R. Matteoli B. Landi, A. and L. Ceccherini-Nelli. Modelling and control of hiv dynamics. *Comput. Methods Prog. Biomed.*, 89:162–168, Feb 2 2008. DOI = http://dx.doi.org/10.1016/j.cmpb.2007.08.003. [cited at p. 31]

[79] D. N. Lawley and A. E. Maxwell. *Factor Analysis as a Statistical Method*. Butterworths, second ed. edition, 1971. [cited at p. 122]

[80] Phillips AN Youle M Johnson M and Loveday C. Use of a stochastic model to develop understanding of the impact of different patterns of antiretroviral drug use on resistance development. *AIDS*, 15:2211–2220, 2001. [cited at p. 32, 39]

[81] Prosperi MCF Ulivi G Zazzi M. Statistical comparison of machine learning techniques for treatment optimisation of drug-resistant hiv-1. In *IEEE Computer Based Medical Systems*, pages 427–432, 2007. [cited at p. 120, 145]

[82] Zaccarelli M. Multiple drug class-wide resist. associated with poorer survival after treatment failure in a cohort of hiv-infected patients. *AIDS*, 19(10):1081–1089, 2005. [cited at p. 139]

[83] Hall MA. *Correlation-based Feature Selection for Machine Learning*. PhD thesis, Waikato University, Department of Computer Science, Hamilton, NZ, 1998. [cited at p. 118, 129, 145]

[84] Jensen MA. 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. *Journal of Virology*, 77(24):13376–13388, 2003. [cited at p. 124, 130]

[85] Nowak M.A. and R.M. May. *Virus Dynamics: Mathematical Principles of Immunology and Virology*. Oxford University Press, 2000. [cited at p. 31]

[86] L.M. Mansky and H.M. Temin. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *J. Virol.*, 69:5087–5094, 1995. [cited at p. 37]

[87] Prosperi MCF and Ulivi G. Evolutionary fuzzy modelling for drug resistant hiv-1. In Abraham A et al Pedrycz W, editor, *Enginnering Evolutionary Intelligent Systems*, chapter 10. Springer. in press. [cited at p. 119, 120]

[88] Lathrop R Pazzani MJ. Combinatorial optimization in rapidly mutating drug-resistant viruses. *Journal of Combinatorial Optimization, Kluwer Academic Publishers*, 3:301–320, 1999. [cited at p. 9]

[89] Beerenwinkel N. Geno2pheno: Interpreting genotypic hiv drug resistance tests. *IEEE Intelligent Systems in Biology*, 16(6):35–41, 2001. [cited at p. 146]

[90] Beerenwinkel N. Geno2pheno: estimating phenotypic drug resistance from hiv-1 genotypes. *Nucleic Acids Research*, 31(13):3850–3855, 2003. [cited at p. 35, 36, 146]

[91] Beerenwinkel N. Tree models for the evolution of drug resistance. *Antiviral Therapy*, 8(S107), 2003. [cited at p. 140]

[92] Perelson A.S. Nelson P.W., Murray J.D. A model of hiv-1 pathogenesis that includes an intracellular delay. *Mathematical Biosciences*, 163(2):201–215, 2000. [cited at p. 31]

[93] Gotoh O. An improved algorithm for matching biological sequences. *Journal of Molecular Biology*, 162:705–708, 1982. [cited at p. 127]

[94] Brunak S Baldi P. *Bioinformatics: the Machine Learning Approach*. MIT Press, 2001. [cited at p. 10]

[95] Dorr P. 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. *Antimicrobial Agents and Chemotherapy*, 49(11):4721–4732, 2005. [cited at p. 124]

[96] C.A. Petri. *Kommunikation mit Automaten*. PhD thesis, University of Bonn, 1962. [cited at p. 79]

[97] Kaufman L Rousseeuw PJ. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York, 1990. [cited at p. 47, 122, 128]

[98] The EuResist project:. www.euresist.org. [cited at p. 52, 140]

[99] J.R. Quinlan. Learning with continuous classes. *In: 5th Australian Joint Conference on Artificial Intelligence, Singapore*, pages 343–348, 1992. [cited at p. 36]

[100] Garzino-Demo A Gallo R. Spontaneous and antigen-induced production of hiv-inhibitory β-chemokines are associated with aids-free status. *Proc Natl Acad Sci, USA*, 96(21):11986–91, 1999. [cited at p. 123]

[101] Los Alamos HIV repositories:. http://www.hiv.lanl.gov/content/index. [cited at p. 125, 145]

[102] The ARCA repository:. https://www.hivarca.net. [cited at p. 74, 92]

[103] S. Bonhoeffer R.M. Ribeiro. A stochastic model for primary hiv infection: optimal timing of therapy. *AIDS*, 13(351-357), 1999. [cited at p. 32]

[104] S. Bonhoeffer R.M. Ribeiro. Production of resistant hiv mutants during antiretroviral therapy. *PNAS*, 97(14):7681–7686, 2000. [cited at p. 32]

[105] Johnson NL Kotz S and Balakrishnan N. *Continuous Univariate Distributions*, volume 1-2, chapter 18-29. Wiley, New York, 1995. [cited at p. 47, 128]

[106] LeCessie S and VanHouwelingen JC. Ridge estimators in logistic regression. *Applied Statistics*, 41(1):191–201, 1992. [cited at p. 114, 119, 129, 145]

[107] Pillai S. A new perspective on v3 phenotype prediction. *AIDS Res Hum Retroviruses*, 19(2):145–149, 2003. [cited at p. 124, 126]

[108] J. W. Sammon. A non-linear mapping for data structure analysis. *IEEE Trans. Comput.*, C-18:401409, 1969. [cited at p. 47, 50]

[109] I. Savenkov. Probabilistic modeling of genetic barriers enables reliable prediction of haart outcome at below 15% error rate. BioSapiens-viRgil Workshop on Bioinformatics for Infectious Diseases, Casear Bonn, Germany, September 2005. [cited at p. 140]

[110] Altschul SF. Basic local alignment search tool. *J Mol Biol*, 215:403–410, 1990. [cited at p. 145]

[111] R statistical programming language:. http://www.r-project.org/. [cited at p. 129]

[112] Weka Data Mining Suite. http://www.cs.waikato.ac.nz/ml/weka/. [cited at p. 129]

[113] Beerenwinkel N Rahnefuhrer J Daumer M Hoffman D Kaiser R Selbig J Lengauer T. Learning multiple evolutionary pathways from cross-sectional data. *RECOMB*, pages 27–31, March 2004. [cited at p. 91, 97]

[114] Dragic T. Hiv-1 entry into cd4+ cells is mediated by the chemokine receptor cc-ckr-5. *Nature*, 381(6584):667–673, 1996. [cited at p. 123]

[115] Sing T. Learning mixtures of localized rules by maximizing the area under the roc curve. 16th European Conference on Artificial Intelligence (ECAI), Workshop on ROC Analysis in AI, 2004. [cited at p. 124, 126, 131]

[116] Sing T Svicher V Perno CF Lengauer T. Characterization of novel hiv drug resistance mutations using clustering, multidimensional scaling and svm-based feature ranking. *PKDD - LNAI*, 3721:285–296, 2005. [cited at p. 49, 51, 55]

[117] Yin J Beerenwinkel N Rahnefuhrer J Lengauer T. Model selection for mixtures of mutagenetic trees. *Statistical Applications in Genetics and Molecular Biology*, submitted, 2005. [cited at p. 90]

[118] W.Y. Tan. A stochastic model for drug resistance in aids chemotherapy and the hiv incubation distribution. *Statistics & Probability Letters*, 25:289–299, 1995. [cited at p. 32]

[119] Smith TF and Waterman MS. Comparison of biosequences. *Advan Appl Math*, 2:482–489, 1981. [cited at p. 127]

[120] Routy JP Petrella M Wainberg MA Turner D, Brenner BG. Rationale for maintenance of the m184v resistance mutation in human immunodeficiency virus type 1 reverse transcriptase in treatment experienced patients. *New Microbiol*, 27(2 Suppl 1):31–9, Apr 2004. [cited at p. 75]

[121] International AIDS Society USA. http://www.iasusa.org/. [cited at p. 48, 49, 54, 55, 74, 99, 146, 148, 149, 152]

[122] Foulkes AS De Gruttola V. Characterizing the progression of viral mutants over time. *Journal of the American Statistical Association*, 98(464):859–867, 2003. [cited at p. 73, 74]

[123] Weston J Mukherjee S Chapelle O Pontil M Poggio T Vapnik V. Feature selection for svms. *Advances in Neural Information Processing Systems*, 12:526–532, 2000. Cambridge, MA, USA, MIT Press. [cited at p. 117]

[124] M Salemi A Vandamme. *The Phylogenetic Handbook: A Practical Approach to DNA and Protein Phylogeny.* Cambridge University Press, 2003. ISBN 052180390X. [cited at p. 27]

[125] P.-F. Verhulst. Recherches mathmatiques sur la loi d'accroissement de la population. *Nouv. mm. de l'Academie Royale des Sci. et Belles-Lettres de Bruxelles*, 18:1–41, 1845. [cited at p. 32]

[126] Virco:. www.vircolab.com. [cited at p. 146]

[127] V. Volterra. Variations and fluctuations of the number of individuals in animal species living together. In *Animal Ecology.* McGraw-Hill, 1931. [cited at p. 28, 31]

[128] vv.aa. *HIV Medicine.* Hoffmann, Rockstroh, Kamps eds, Flying, 2006. ISBN 3-924774-50-1. [cited at p. 10]

[129] Resch W. 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, 2001. [cited at p. 124, 126]

[130] Paxton WA. Relative resistance to hiv-1 infection of cd4 lymphocytes from persons who remain uninfected despite multiple high-risk sexual exposure. *Nat Med*, 2(4):412–417, 1996. [cited at p. 123]

[131] Kuritzkes D.R. Wu, H. and D.R. et al. McClernon. Characterization of viral dynamics in human immunodeficiency virus type 1-infected patients treated with combination antiretroviral therapy: Relationships to host factors, cellular restoration and virological endpoints. *Journal of Infectious Diseases*, 179(4):799–807, 1999. [cited at p. 31]

[132] Cohen WW. Fast effective rule induction. In *Twelfth International Conference on Machine Learning (ML95)*, pages 115–123, 1995. [cited at p. 114, 128, 129]

[133] Bengio Y. Inference for the generalization error. *Machine Learning*, 52:239–281, 2003. [cited at p. 116]

[134] Ghahramani Z. Unsupervised learning. In Bousquet O Raetsch G Von Luxburg U, editor, *Advanced Lectures on Machine Learning - LNAI 3176*, chapter 5. Spriger-Verlag, Heidelberg, 2004. [cited at p. 46, 88]