THE USE OF MACHINE LEARNING TO IMPROVE THE EFFECTIVENESS OF ANRS IN PREDICTING HIV DRUG RESISTANCE

BY

ABHISHEK SHRIVASTAVA

Submitted in partial fulfilment of the requirements for the Qualification of Master of Medical Informatics (MMedSci-MI)

In the Discipline of Telehealth,

School of Nursing and Public Health, College of Health Sciences,

University of KwaZulu-Natal, Medical School Campus

Supervisor

Dr Yashik Singh

2016

PREFACE

The study described in this dissertation was carried out by Mr. Abhishek Shrivastava and has not been submitted in any other form to another University. This study was carried out in the School of Nursing and Public Health, College of Health Sciences, University of KwaZulu-Natal, Durban, South Africa under the supervision of Dr. Yashik Singh.

Abhishek Shrivastava

Student  Date 04-04-2017

Dr Yashik Singh

Supervisor _____ Date_____

DECLARATION

I, Abhishek Shrivastava, student number: 983193775 hereby declare that the research reported in this dissertation/thesis titled: '**The use of machine learning to improve the effectiveness of ANRS in predicting HIV drug resistance'** except where otherwise indicated, is the result of my own research. This dissertation has not been submitted in part or full for any other degree or to any other University or Tertiary Institution.

This dissertation does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons. This dissertation does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers.

Where other written sources have been quoted, then:

a) their words have been re-written but the general information attributed to them has been referenced;

b) where their exact words have been used, their writing has been placed inside quotation marks, and referenced.

Where use was made of the work of others, it is duly acknowledged.

The research conducted in this study was carried out under the supervision of Dr. Yashik Singh.

Abhishek Shrivastava

Signature 

Date ……..

Supervisor Dr. Yashik Singh

Signature ……..

Date 04-04-2017

ACKNOWLEDGEMENTS

I would like to express my gratitude to the following people

- Professor M. Mars for his guidance and support in the completion of this study.
- Dr. Yashik Singh, also my Supervisor, for helping me in every step of my study. I would be nothing without his invaluable guidance and support.
- My father, Dr. C.P. Shrivastava, there's never a time when you have never been there for me. Thank you for always giving me direction, and for constantly reminding me of all the possibilities. Thank you for always believing in me in my rough times.
- My mother, Mamta Shrivastava; you have moulded, chiselled and polished me.
- My wife , Dr. Sonal Shrivastava ;you always stood by me. Thank you.

TABLE OF CONTENTS

LIST OF TABLES

CONFERENCE PRESENTATIONS

Shrivastava A, Singh Y. A pilot study to determine if machine learning can improve the accuracy of the ANRS gold standard in predicting HIV drug resistance. Oral presentation at ICT4 Health Conference 11-13 September,2013.

Shrivastava A, Singh Y. A pilot study to determine if machine learning can improve the accuracy of the ANRS gold standard in predicting HIV drug resistance. Oral presentation at ICT4 Health Conference Including Global Telehealth,10-11 November,2014.

ABBREVIATIONS

AI	Artificial Intelligence
AIDS	Acquired Immunodeficiency Syndrome
ANRS	National Agency for AIDS Research
ARV	Anti- retroviral therapy
CD4	Cluster of Differentiation 4
CFS	Corelation Based Feature Selection
DDI	Diadonosine
DNA	Deoxyribonulceic acid
DTL	Decision Tree Learning
FCBF	Fast Corelation Based Filter
FI	Fusion Inhibitors
HAART	Highly active anti-retroviral therapy
HIV	Human Immunodeficiency Virus
IC50	Half Maximal Inhibitory Concentration
LPV	Lopinavir
MLP	Multi-layer perceptron
MODTREE	Multivalued Oblivious Decision tree
NFV	Nelfinavir
NN	Neural Network
NNRTI	Non-nulceoside reverse transcriptase Inhibitors
NRTI	Nucleoside reverse transcriptase Inhibitors
NVP	Nevirapine
PI	Protease Inhibitors
PR	Protease
RT	Reverse Transcriptase
RTI	Reverse transcriptase Inhibitors
SIM	Simian Immunodeficiency Virus
SIV	Simian Immunodeficieny virus
SQV	Sequanavir
TI	Transcriptase Inhibitors
TDF	Tenofovir

ABSTRACT

BACKGROUD

HIV has placed a large burden of disease in developing countries. HIV drug resistance is inevitable due to selective pressure. Computer algorithms have been proven to help in determining optimal treatment for HIV drug resistance patients. One such algorithm is the ANRS gold standard interpretation algorithm developed by the French National Agency for AIDS Research AC11 Resistance group.

OBJECTIVES

The aim of this study is to investigate the possibility of improving the accuracy of the ANRS gold standard in predicting HIV drug resistance.

METHODS

Data consisting of genome sequence and a HIV drug resistance measure was obtained from the Stanford HIV database. Machine learning factor analysis was performed to determine sequence positions where mutations lead to drug resistance. Sequence positions not found in ANRS were added to the ANRS rules and accuracy was recalculated.

RESULTS

The machine learning algorithm did find sequence positions, not associated with ANRS, but the model suggests they are important in the prediction of HIV drug resistance. Preliminary results show that for IDV 10 sequence positions where found that were not associated with ANRS rules, 4 for LPV, and 8 for NFV. For NFV, ANRS misclassified 74 resistant profiles as being susceptible to the ARV. Sixty eight of the 74 sequences (92%) were classified as resistance with the inclusion of the eight new sequence positions. No change was found for LPV and a 78% improvement was associated with IDV.

CONCLUSION

The study shows that there is a possibility of improving ANRS accuracy.

CHAPTER ONE

INTRODUCTION

This chapter deals with the background of the Human Immunodeficiency virus (HIV) and the need for the machine learning algorithms to detect HIV drug resistance in developing countries

INTRODUCTION

1.1 Introduction to HIV worldwide and South Africa

HIV continues to be the major health concern. Worldwide, there is an estimated 36.9 Million (including 2.6 million children) that are infected with HIV with a global prevalence of 0.8% (UNAIDS 2015) many still living with AIDS (Health Systems Trust, 2011b). In 2014, there were roughly 2 million new HIV infections, of which 220,000 were children. Most of these children live in sub-Saharan Africa and were infected via maternal to child transmission (UNAIDS 2015). HIV/AIDS places a major burden on the resources of developing countries (Health Systems Trust, 2011a). It is the leading cause of death in sub-Saharan Africa (Campbell et al., 2008) and is one the major epidemics in South Africa. Among adults aged 15–49, HIV prevalence for women is over 23% compared to 13% among men, a gender difference of more than 10%. Women aged 15 to 24 have an HIV incidence rate more than four times that of men, with a 4.5% incidence of HIV among Black African women (HSRC 2012).

There are currently almost 5.6 million infected with HIV in South Africa, which is approximately 11% of the South African population (AIDS Committee of Actuarial Society of South Africa, 2008). It is also estimated that there are almost 500 000 patients who exhibit AIDS defining conditions (Health Systems Trust, 2011a). There are an estimated 24.7 million (23.5–26.1 million) people living with HIV in sub-Saharan Africa, nearly 71% of the global total. Ten countries— Ethiopia, Kenya, Malawi, Mozambique, Nigeria, South Africa, Uganda, the United Republic of Tanzania, Zambia and Zimbabwe— account for 81% of all people living with HIV in the region and half of those are in only two countries— Nigeria and South Africa (HIV/AIDS, 2014).

South Africa has the biggest and most high profile HIV epidemic in the world, with an estimated 6.3 million people living with HIV in 2013. In the same year, there were 330,000 new infections while 200,000 South Africans died from AIDS related illnesses (Simbayi et al., 2014). Province wise, Kwazulu-Natal has been reported the highest percentage of HIV prevalence. The rates of HIV infection is shown in the figure table below;

PROVINCE	HIV PREVALENCE RATE
Eastern Cape	11.6
Free State	14.0
Gauteng	12.4
KwaZulu-Natal	16.9
Limpopo	9.2
Mpumalanga	14.1
Northern Cape	7.4
North West	13.3
Western Cape	5.0

Table 1: shows the prevalence rate of HIV in different provinces in South Africa (South African National HIV Prevalence, Incidence and Behaviour Survey, 2012)

1.2 The Human Immunodeficiency virus

Human Immunodeficiency virus (HIV) belongs to the family retroviridae and genus lentivirus. These are positive-stranded RNA viruses. Following infection, the single-stranded RNA genome is reverse-transcribed into double-stranded DNA into the host cell (Cann and Karn, 1989). The resulting double-stranded DNA is subsequently irreversibly integrated into the host genome. These HIV virions have spherical shape morphology of 100-120nm. The virus is composed of lipid bilayer membrane which envelopes a dense nucleocapsid (Cann and Karn, 1989). This nucleocapsid is formed by capsid protein p24 and contains two single stranded RNA copies, the reverse transcriptase, the viral protease and the other accessory and regulatory protein and also some cellular factors.These destroy the human immune system over a long period of time by invading T

helper cells. There are two types of HIV: HIV I and HIV II, both of which believed to be originated in Africa (Wilson et al., 2002). These were at first only found in primates in the form of Simian Immunodeficiency virus but zoonotic infections, through infected meat, caused it to transfer to humans (Wilson et al., 2002).

1.2.1 HIV I

HIV I was first detected in the late 1970s (Gallo and Montagnier 2003)HIV I is believed to be originated from Central Africa, and 95% of all HIV infection can be attributed to HIV I (Quinn, 1998). HIV-1 is subdivided into four groups representing four separate introductions of simian immunodeficiency virus (Akhilesh et al.) into humans: M (major), N (non-major), O (outliers) and P (most similar to SIV) (Faria et al., 2014)

Currently group M is subdivided into nine subtypes or clades: A, B, C, D, F, G, H, J, K based on variations in genetic sequence characteristics (Kanki et al., 1999, Robertson et al., 2000). It is, however, possible for viruses from different subtypes to form mosaic genomes called circulation recombinant forms (CRFs) (Abecasis et al., 2007)

Most African patients are infected with M type with subtype C, whereas the European patients are infected with M type with subtype B (Robertson et al., 2000).

1.2.2 HIV II

HIV II arose from Western Africa and it was first identified in the mid-1980s (De Cock and Brun-Vézinet, 1989) Until 2000 it was only confined to West Africa, but due to immigration of people all-round the world it is found elsewhere as well. HIV II is also further divided into 5 types from A to E. It has a slower infectivity and progression rate compared to HIV-I (De Cock and Brun-Vézinet, 1989) The patients who progress slowly and has more indolent course can be deciding factor to monitor for HIV-2–infection (Campbell-Yesufu and Gandhi, 2011).

1.2.3 Mode of infection

There are different modes of HIV transmission, blood transmission being the most common. Table 2 shows that probability of being infected with HIV-1 depending on the means of infection.

EXPOSURE	PROBABILITYOF INFECTION
1. Vaginal Intercourse	0.1%
2. Anal Intercourse	1%
3. Percutaneous exposure	0.3%
4. Needle Sharing	1%
5. Mother-to-child	20%-40%
6. Blood transfusion	100%

Table 2: Shows the type of exposure and the associated probability of infection (Adapted from 1.(Boily et al., 2009), 2. (Vitinghoff et al., 1999), 3. (Del Romero et al., 2002), 4. (Baggaley et al., 2006), 5. (Townsend et al., 2008), 6. (Baggaley et al., 2006).

1.2.4 HIV Management

There is no cure for HIV, but HIV infection can be effectively managed with antiretroviral (Marques et al.) drugs usually in the form of highly active antiretroviral therapy (HAART), which comprises of a regimen of three drugs from at least two of the following five drug classes (Mitton, 2000a, Bartlett and others, 2004, Pierret, 2007, Beerenwinkel and others, 2002). Reverse transcriptase inhibitors (RTI) which consists of Nucleoside Reverse Transcriptase Inhibitors (NRTI) and Non-Nucleoside Reverse Transcriptase Inhibitors (NNRTI), Protease inhibitors (PI), Integrase

inhibitors (Tomimatsu et al.), Fusion inhibitors (FI) and Entry Inhibitors (EI) (Arts and Hazuda, 2012)

1.2.5 Mechanism of action of Anti-retroviral drugs

Nucleoside transcriptase inhibitors (NRTI's) work by competing with the nucleoside for active binding sites. If the drug successfully binds to an active site, then the transcription or copying of the sequence is hindered (Fowler et al., 2014) Non- nucleoside reverse transcriptase inhibitors NNRTI's inhibits the transcription in a non-competitive manner. It binds to any location other than active sites in an attempt to stereotypically obstruct the sequence from passing over the active sites of the reverse transcriptase enzyme (Usach et al., 2013)

Protease inhibitors (PI's) work by inhibiting the action of the protease enzyme. It impedes the cleavage of the large gap-pol polyprotein into active virion by binding to the active pocket-like site of the protease enzyme and this causes the formation of a non-infectious and immature virion (Ghosh et al., 2016)

Integrase Inhibitors inhibit the action of the integrase enzyme. It prevents the virus from inserting it's vDNA into the hosts cell DNA. This is accomplished by blocking the active sites required by the integrase enzyme, thus preventing the insertion (Mesplède et al., 2014)

Fusion Inhibitors (FI) prevent the replication of the HIV from outside the host cell by blocking fusion. Fusion inhibitors work by preventing the hair-pin bend action that allows the HIV to be closer to the host cell. Thus because the HIV cannot get close to the host cell, no fusion can take place (Chong et al., 2016)

Entry Inhibitors (EI's) work by halting the HIV from binding to the gp120 receptor by competing for this binding site. This prevents the HIV from binding to the co-receptor (gp41). Without the virus binding onto the receptors, entry into the host cell is impossible (Wilson D et al., 2002).

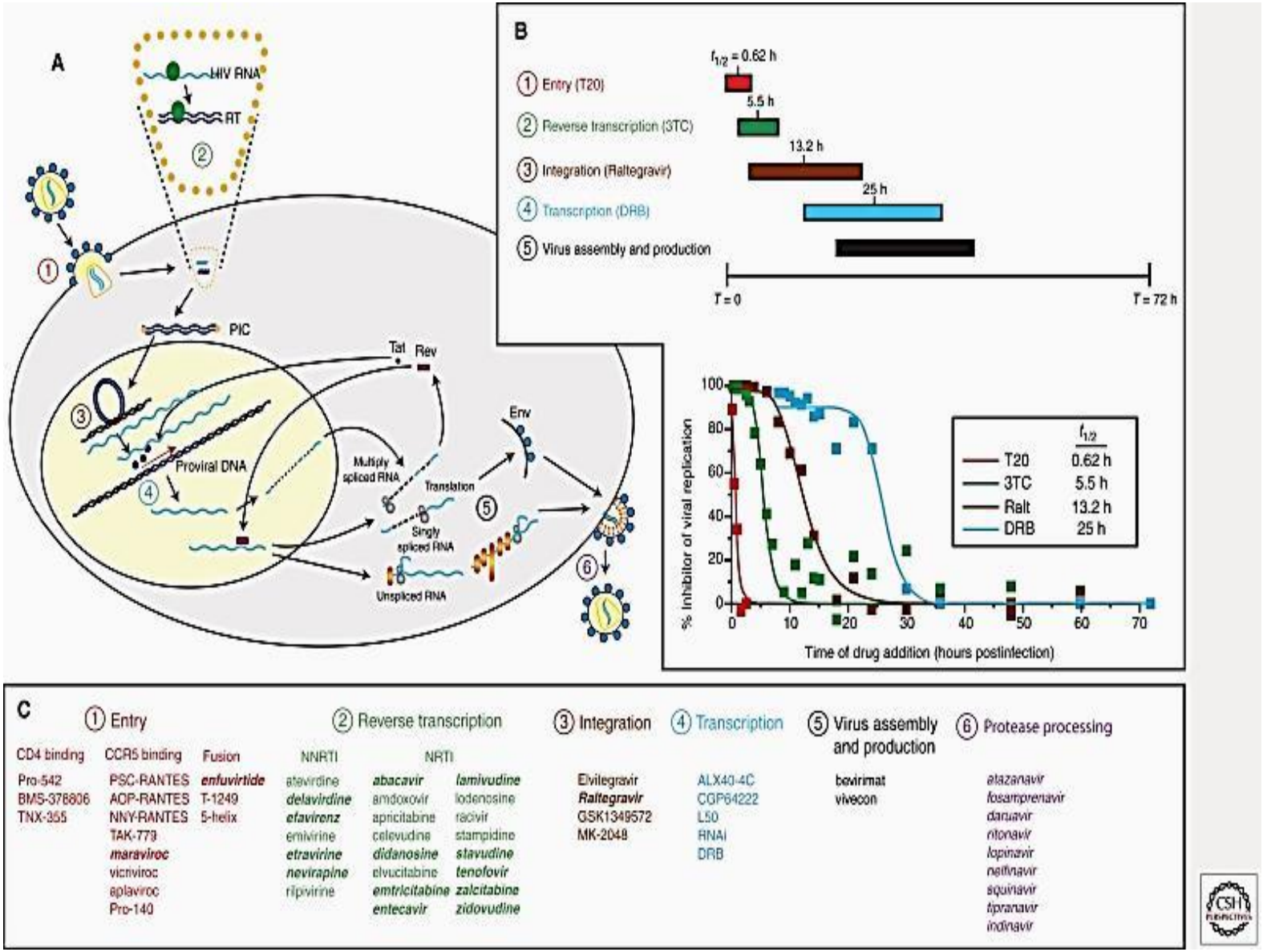
A
HIV RNA
RT
PIC
Tat
Rev
Env
Proviral DNA
Multiply spliced RNA
Translation
Singly spliced RNA
Unspliced RNA
B
$t_{1/2}$ = 0.62 h
1 Entry (T20)
5.5 h
2 Reverse transcription (3TC)
13.2 h
3 Integration (Raltegravir)
25 h
4 Transcription (DRB)
5 Virus assembly and production
T = 0
T = 72 h
% Inhibitor of viral replication
Time of drug addition (hours postinfection)
T20 0.62 h
3TC 5.5 h
Ralt 13.2 h
DRB 25 h
C
1 Entry
2 Reverse transcription
3 Integration
4 Transcription
5 Virus assembly and production
6 Protease processing
CD4 binding
CCR5 binding
Fusion
NNRTI
NRTI
Pro-542
BMS-378806
TNX-355
PSC-RANTES
AOP-RANTES
NNY-RANTES
TAK-779
maraviroc
vicriviroc
aplaviroc
Pro-140
enfuvirtide
T-1249
5-helix
atevirdine
delavirdine
efavirenz
emivirine
etravirine
nevirapine
rilpivirine
abacavir
amdoxovir
apricitabine
calevudine
didanosine
elvucitabine
emtricitabine
entecavir
lamivudine
lodenosine
racivir
stampidine
stavudine
tenofovir
zalcitabine
zidovudine
Elvitegravir
Raltegravir
GSK1349572
MK-2048
ALX40-4C
CGP64222
L50
RNAi
DRB
bevirimat
vivecon
atazanavir
fosamprenavir
darunavir
ritonavir
lopinavir
nelfinavir
saquinavir
tipranavir
indinavir

Figure 1;The above figure shows the potential or current target for anti-retroviral drugs in HIV life cycle. (*A*) Schematic of the HIV-1 life cycle in a susceptible $CD4^+$ cell. (*B*) Time frame for antiretroviral drug action during a single-cycle HIV-1 replication. Adapted from (Arts and Hazuda, 2012).

1.2.6 HIV drug Resistance

Various factors influence the management of HIV/AIDS. Some factors include poor treatment regimen prescribed by the physician, WHO stage and progression of the disease, levels of drug concentration achieved, patient adherence to the treatment, drug resistance, and toxic effects of the drug (Clavel and Hance 2004) Drug resistance is arguably the most critical aspect of treatment and three common reasons that lead to the development of HIV antiretroviral drug resistance are high replication rates, selective pressure and initial infection by resistant strains of HIV. Thus it is inevitable that drug resistance will become a reality in most patient's treatment (Vercauteren and Vandamme, 2006).

There are various means of measuring HIV drug resistance. The use of computer algorithms is one of the methods which is gaining popularity in developing countries. These algorithms take various attributes as inputs, and predict a measure of resistance to HIV drugs. ANRS is one example of such algorithms (Gilks et al., 2006)

1.3 ANRS Algorithm

One widely used computer based interpretation algorithm was built by the French National Agency for AIDS Research AC11 Resistance group and is called the Nationale de Recherches sur le SIDA (ANRS) (AC11 Resistance group, 2013). ANRS is seen as a gold standard in interpreting HIV drug resistance using mutations in the genomes. ANRS classifies ARV resistance according to three levels: susceptible, intermediate, and resistant. "Susceptible" indicates that a particular ARV drug will be effective against HIV; "intermediate" indicates that the ARV drug is partially effective; and if the ARV is not effective at all, it is classified "resistant" (Liu et al., 2008)

This algorithm is mainly based on correlation between drug resistance and virological outcome of patients who fail ARV treatment. The algorithm is based on a linear combination of mutations. If a particular mutation or a group of mutations are is present in the genome, the algorithm returns a resistance profile applicable to that particular sequence e.g. if the mutation A98S is present in the

genome, resistance to the NVP is deduced, were as a E138K mutation will indicate intermediate resistance to the NVP drug (Meynard et al., 2002b)

Compounded to the burden of disease, healthcare is moving towards personalized medicine (Ginsburg and McCarthy, 2001) This paradigm of healthcare consists of services provided based on individual criteria rather than a set of generalized symptoms. The management of HIV/AIDS is also becoming personalized (De Luca et al., 2004, Frentz et al., 2010, Yashik and Maurice, 2012). Thus the added burden of timely access to information, representative computer algorithms, shortage of adequately trained staff etc. becomes more evident in developing countries.

The use of computer algorithms in personalized medicine is becoming more popular (Vercauteren and Vandamme, 2006, Pasomsub et al., 2010). However, in many cases, like ANRS these algorithms are created in silos, uses the knowledge and perceptions of a few specialists in the field, and in specific environments. There are thus many opportunities to improve these algorithms (Snoeck et al., 2006a, Vergne et al., 2006).

Given the background for the study, the aim of this study is to improve on the HIV drug resistance prediction ability of the ANRS algorithm using machine learning techniques. The research questions of the study and to answer the questions theobjectives made are as follows:

I. Is it possible to determine important HIV drug resistant associated mutations using machine learning?

Objective I. . Create machine learning algorithms to determine the relationship between mutations and the resistance in the genome of HIV patients i.e. create an association matrix.

II. Is it possible to improve the current accuracy of the ANRS gold standard?

Objective II. Compare the association matrix mutations to the matrix associated with resistance as per ANRS gold standard rules.

III. Is it possible to improve the current sensitivity, specificity, positive predictive value and negative value of the ANRS gold standard?

Objective III. Determine if the ANRS gold standard can be improved with the incorporation of the associated matrix

1.4 Conclusion

Computer based algorithms are usually based on an experts' understanding of the domain, available datasets that are used for machine learning and understanding of published literature. This has led to the creation of many different interpretation algorithms, which produce different resistance measures even if applied to the same resistance profile. The main aim of this study is to improve the efficiency of predicting HIV drug resistance by ANRS. This chapter described the background required for the thesis. The next chapter describes the area of HIV drug resistance

CHAPTER TWO

LITERATURE REVIEW

LITERATURE REVIEW

The literature review explores the three dominant themes of research questions i.e. the mechanism of drug resistance in HIV treatment, possibility to determine important HIV drug resistant associated mutations using machine learning and improving the current sensitivity, specificity, positive predictive value and negative value of the ANRS gold standard. Major scientific databases like PubMed, google scholar and science direct were used for doing the literature review.

2.1 Introduction

Although combination therapy for HIV infection represents a triumph for modern medicine, eradication for HIV is still not possible due to the latency of the virus (Katlama et al., 2013). Both preventive and therapeutic vaccines are currently being tested. HIV is effectively controlled by antiretroviral which is introduced at some point between clinical latency and progression to AIDS. Several factors are considered before starting ARV like the CD_4 counts. Prophylactic treatment can be done in post exposure; this reduces the chances of the virus progression. Also these drugs are used to prevent mother to child transmission of the virus. ARV are managed in highly anti-retroviral therapy (HAART), the drugs are usually a combination of three drugs from the class of possible five drugs (Beerenwinkel et al., 2005, Mitton, 2000b, Tang and Shafer, 2012).

Factors that influence treatment of HIV/AIDS with antiretroviral include poor treatment regimen prescribed by the physician, WHO stage of the disease which is related to the progression of the disease, levels of drug concentration achieved, how strictly the patient adheres to the regimen, drug resistance, and the toxic effects of the drug (Richman and Staszewski, 2000). Drug resistance is arguably the most critical aspect of treatment.

HIV drug resistance is caused inevitably due to selective pressure (viruses that happen to survive the drug are favoured, and resistant virus strains evolve within the patient) which can happen sometimes in just few weeks as ARV is introduced. Not only ARV's but other things like pesticides, industrialization and antibiotics also confer selective pressure. HIV virus represents an interesting case as its speed of evolution has changed considerably over time (Feder et al., 2015).When the HIV virus becomes tolerant to the antiretroviral treatment it causes HIV drug resistance.

The ability of HIV to continue to mutate and reproduce even during the treatment is called HIV drug resistance. Three common reasons that lead to the development of HIV antiretroviral drug resistance are high replication rates, selective pressure and initial infection by resistant strains of HIV. Thus it is inevitable that drug resistance will become a reality in most patient's treatment (Vercauteren and Vandamme, 2006).

2.2 Factors that lead to the development of HIV drug resistance

2.2.1 High replication rate

The HIV-replicates itself 1010 times in 24 hours (Moodley, 2006). The minimum duration of the HIV-1 life cycle in vivo is 1.2 days on average, and that the average HIV-1 generation time, defined as the time from release of a virions until it infects another cell and causes the release of a new generation of viral particles, is 2.6 days (Perelson et al., 1996). It is estimated that in an individual with the active disease, all possible combinations of nucleotides in the genome are generated in a single day. Due to a high replication rate and lack of error checking in the polymerase enzyme, the HIV has the highest known mutation rate of any organism. It has been shown that HIV-1 virus has extremely high mutation rate of $(4.1 \pm 1.7) \times 10^{-3}$ per base per cell, the highest reported for any biological entity (Cuevas et al., 2015b).

Mutations occur when there is a change in the nucleotide sequence of the genome of any organism. It is the result of any unrepaired damage which has been there in either DNA or the RNA of the cell. These are usually caused by radiation or chemical reactions. Mutations may or may not produce observable changes in the characteristics (Lodish et al., 2000).

Mutations can result in several different types of changes in the sequences. It can either have effect in the gene, where it doesn't allow the gene to function properly or it can have absolutely no effect. Mutations can cause a lot of DNA and RNA to duplicate, which causes evolving of new genes. There are two main kinds of mutations. Small scale mutations are a kind of mutations where a small gene in one or few nucleotide is affected. Small scale mutation is further classified into point mutation, insertion and deletions. Point mutation can further be broken into silent, missense and nonsense mutations (Lodish et al., 2000)

The other major mutation type is the large scale mutation, where the chromosome structure is affected. Large scale mutation is further divided into amplification, deletion, mutation which juxtaposes already separated DNA and loss of heterozygosity (Lodish et al., 2000)The virus mutates by adding, subtracting or substituting one or more nucleotides at various positions along the genome sequence. The high mutation rate (4.1 ± 1.7) × 10-3 per base per cell (results in the formation of quasi-species/variants and minority sub-populations (Cuevas et al., 2015a) Similar to other organisms, the virus's main purpose is to survive. It does this by encouraging the survival of the mutated virus that will survive in patient's body and pays less attention to those that will not survive (survival of the fittest) (Wilson D et al., 2002).

2.2.2 Selective pressure

The rate of resistance caused by the selective pressure is determined by residual replication and the genetic barrier of the drug (David, 2009) The first factor, residual replication, refers to the potency of the drug regimen. When under the selective pressure of a single drug regimen, it stands to reason that, a large number of mutants will escapes from the ARV and will replicate (http://i-base.info/category/publications/) However, if a two drug regimen is used i.e. a slightly increased drug regimen potency, the rapid evolution of the drug resistant variants is initiated. This results in more drug resistant variants escaping and replicating. Fortunately, under very high drug regimen potency (three or more drugs from at least two different drug classes) viral replication is minimized to such an extent that the likelihood of resistance emerging is diminished. Such drug combinations increase the second selective pressure factor called the generic barrier (Feder et al., 2015).

The second factor is due to the nature of the drug itself. If the genetic barrier is low it signifies that very few mutations are "sufficient to cause resistance". Examples of such drugs are lamivudine and NNRTI's. Other drugs have a high genetic barrier, implying that many mutations are required to induce resistance. Indinavir and Abacavir are examples of drugs with a high genetic barrier (Moutouh et al., 1996) In a high genetic barrier environment, variants that do escape have a very small advantage over the wild-type virus because it must "accumulate multiple additional mutations before out -growth can occur". This implies that a person under ARV treatment for extended periods will eventually develop drug resistance. HAART treatment is lifelong and adherence of around 95% is required in order to maintain low levels of virus and reduce the development of drug resistant quasispecies (Carvalho and Pinto, 2016)

2.2.3 Initial infection

Individuals may also acquire drug resistance during initial infection as a result of primary transmission. In developed countries, it is estimated that 10% of the circulating viruses contain at least one drug resistance mutation (Clavel and Hance 2004)

It was reported in 2009, that 37% of patients that required ARV treatment actually received ARV drugs in South Africa (Adam and Johnson, 2009). Due to the high replication rates of the virus, selective pressure caused by the ARV drugs and initial infection by resistant strains of HIV, it is inevitable that drug resistance will become a concern in the treatment of HIV/AIDS infected patients.

2.3 Measuring HIV drug resistance

HIV drug resistance is normally measured by phenotypic or genotypic testing. Both these are usually wet chemistry tests that comprise of isolating the HIV from an infected individual and then using standard wet-chemistry techniques to measure and represent concentration or resistance directly (Hirsch et al., 1998)

2.3.1 Phenotypic testing

Phenotypic assays use direct in-vitro method of predicting resistance which is based on IC_{50} score (it is the concentration of drugs which is required to kill half the virus of the sample virus population which is wild type). It also requires growing HIV in a sterile and controlled laboratory environment. Here ARV is used to kill the 50% of the virus and then the two are compared. These comparisons are used in predicting the resistance (John et al., 2000)

$$\text{Resistance} = \begin{cases} \text{Not Resistant}, & IC_{50} < \text{susceptible score}; \\ \text{Susceptible resistance}, & \text{susceptible score} < IC_{50} < \text{resistant score} \\ \text{Resistant}, & IC_{50} > \text{resistant score} \end{cases}$$

2.3.2 Genotypic testing

HIV virus genome when mutated causes chemical and stereotypical changes, which cause the ARV to be less effective. Genotypic assays use chemicals to test for resistance by determining the

presence of mutation. This array requires complex mutation and genotypic variation patterns. Genotypic assay is a kind of resistance test which uses chemical methods to get the presence of any mutation or variation of mutations. Therefore, this kind of mutation requires interpretation of complex mutation and genotypic patterns. This also requires inference of phenotype, using rules and algorithms. These rules and algorithms mainly apply to subtype prevalent (Durant et al., 1999)

Phenotypic assays are time consuming, expensive and each test is done with just single ARV. Therefore, genotypic assays are more used as it is much cheaper and faster (Jiamsakul et al., 2016).

Interpreting the results of the mutations present found in the genome sequence is difficult and is a source of contention. Computer based interpretation algorithms may be created to determine HIV drug resistance using mutations in the genome. There are generally two types of these interpretation algorithms (Yashik and Maurice, 2012) One consists of predetermined rules, created by human experts, which associate particular mutations to resistance profiles or measures. For example, a typical expert rule is that a mutation in the 215^{th} amino acid in the reverse transcriptase sequence, where tyrosine replaces threonine, causes resistance to ARV Zidovudine. The other consists of using artificial intelligence or also called machine learning, to determine associations between mutation in the genome and resistance profiles based of large amounts of data (Hales et al., 2006)

There are two groups in which the interpretation algorithm has been classified (Yashik and Maurice, 2012)

1. One based on the domain knowledge, which means it is based on the fact that certain combinations of known mutations cause unequivocal resistance.

2. The other is not based on the predefined domain knowledge; it includes machine learning and statistical methods.

2.4 Interpretation algorithm based on domain knowledge

This interpretation algorithm is accepted by most of the scientific community. All the computational decisions concerning resistance are based on known mutation-resistance rules which are already published in scientific literature. REGA, ANRS Stanford's HIV db and RetroCram are the examples of such algorithms (de Oliveira et al., 2005).

REGA and ANRS have 3 different levels of resistance classification, resistance, intermediate and susceptible (Singh and Mars, 2014). Susceptible means that a particular ARV will be effective against HIV. Intermediate means the particular drug will be partially effective against HIV. Resistant means that the particular drug will be ineffective and lead to virological failure (Singh and Mars, 2014).On the other hand Stanford's HIV db output consists of (AC11 Resistance group) a list of penalty scores for each antiretroviral (Marques et al.) resistance mutation in a submitted sequence, (2) estimates of decreased NRTI, NNRTI, protease and integrase inhibitor susceptibility, and (3) comments about each ARV resistance mutation in the submitted sequence (Tang et al., 2012).

ANRS is a French National Agency for AIDS Research, which has an algorithm which is used to check the HIV drug resistance. ANRS is an interpretation algorithm which is based domain knowledge. This is much more accepted by the scientific community. ANRS has been regarded as the gold standard among all algorithm used for HIV drug resistance (de Oliveira et al., 2005).

2.5 Interpretations algorithms not based on known domain knowledge

There are many machine learning algorithms applied to finding a predictable correlation between genotyping and phenotypic data. These data are known as virtual phenotyping.
Virtual phenotyping is getting more popular and Kuritzkes supports the virtual phenotype as a tool for viral genotype (Singh and Mars, 2014) Here are few successful algorithms -Neural Network
-Support Vectar Machines
-Linear regression models
-Decision Tree
-Ridge Regression.
The above mentioned algorithms were primarily developed using different sets of datasets, subtypes and from non-treated (drug naïve) and treated patients (Singh and Mars, 2014).
Interpretation algorithm based on domain knowledge is regarded as gold standard and widely used by the scientific communities as they are based on scientific and published interaction between certain mutation or combination of mutations (Singh and Mars, 2014).
The table below summarises some of the previously done studies;

Published algorithm against gold standard	Creator	Accuracy
AntiRetroScan	Zazzi M. (2008)	89.4
Associative classification	Srisawat A and Kijsirikul B (2008)	84.1
MLP(ΔEnergy)	Bonet I et al (2007)	85.8
MLP (Energy)	Bonet I et al (2007)	88.6
RBNN	Bonet I et al (2007)	93.6
Committee Neural Network	Draghici S and Potter R (2003)	78
Decision tree	James R (2004)	76
Dr Seqan	Garriga C and Menendez A (2006)	43
Geno2Pheno	Beernwinkel N. et al (2003)	
KNN	James R (2004)	49
RegaInst	Garriga C and Menendez A (2006)	46

Retrogram	Garriga C and Menendez A (2006)	61
Stanford HIV-db algorithm	authors implementation	
Stanford HIV-db algorithm	Garriga C and Menendez A (2006)	29
Stanford HIV-db algorithm	Zazzi M. (2008)	84.3
Visible Genetics/Bayer Diagnostics Guidelines 6.0	Zazzi M. (2008)	75

Table 3: The summary of the comparisons of the results from the previous studies.

The above table also indicates that although interpretation algorithms have been researched for a few years, improvement is possible. There may be a time when these electronic interpretation algorithms may completely replace the laboratory phenotyping testing

Drug resistance is a very important factor influencing the failure of current HIV therapies. The ability to predict the drug resistance of HIV protease mutants may be useful in developing more effective and longer lasting treatment regimens. Drăghici and Potter best combination yielded an average of 85% coverage and 78% accuracy on previously unseen data. This was more than two times better than the 33% accuracy expected from a random classifier. DR_SEQAN which is an easy to use off-line application that provides expert advice on HIV genotypic resistance interpretation with an improved reported accuracy of 43%.

Previous studies have shown a maximum accuracy of 93.6 % and the lowest as 29%.Hence, it clearly shows possibility of using machine learning imprvovment in the accuaracies of ANRS in predicting HIV drug resistance.

CHAPTER THREE

MANUSCRIPT

THE USE OF MACHINE LEARNING TO IMPROVE THE EFFECTIVENESS OF ANRS IN PREDICTING HIV DRUG RESISTANCE

THE USE OF MACHINE LEARNING TO IMPROVE THE EFFECTIVENESS OF ANRS IN PREDICTING HIV DRUG RESISTANCE

3.1 Manuscript format

This chapter consists of a paper that has been formatted according to 'African Journal of Biomedical Research'.

Shrivastava A, and Singh Y *The use of machine learning to improve the effectiveness of ANRS in predicting HIV drug resistance.* Original Research paper

3.2 Introduction

This chapter describes the use of machine learning to improve the effectiveness of ANRS in drug resistance. Genomic mutations can be used in computer based algorithms to detect HIV drug resistance. Computer based algorithms are generally based on available datasets and published literature. Therefore, there are always chances of improvement.

3.3. Materials and Methods

Data consisting of genome sequence and a HIV drug resistance measure was obtained from the Stanford HIV database. Machine learning factor analysis was performed to determine sequence positions where mutations lead to drug resistance.

3.4 Results

The use of machine learning showed improvements in ANRS in prediction of HIV drug resistance.

The use of machine learning to improve the effectiveness of ANRS in predicting HIV drug resistance

Shrivastava A[1] and Singh Y[1]

[1]Department of Telehealth, Nelson R Mandela School of medicine, University of KwaZulu Natal, South Africa

Corresponding author; singhyashik@gmail.com

First author; Abhishek@jmh.co.za

Abstract

HIV drug resistance is inevitable due to selective pressure. Computer based interpretation algorithm, built by the French National Agency for AIDS Research AC11 Resistance group, is called the Nationale de Recherches sur le SIDA (ANRS) and is the gold standard in interpreting HIV drug resistance using mutations in the genomes. Computer based gold standard interpretation algorithms are usually based on an experts' understanding of the domain and available datasets and so there may be discrepancies and areas of improvement. The aim of this study is to investigate the possibility of improving the accuracy of the ANRS gold standard in predicting HIV drug resistance. Genome sequence and a HIV drug resistance measure were obtained from the Stanford HIV database (http://hivdb.stanford.edu/). Feature selection was used to determine the most important mutations associated with resistance prediction (association matrix). These mutations were added to the ANRS rules and difference in prediction ability between ANRS and ANRS with the association matrix was measured. Preliminary results show that for IDV 10 sequence positions where found that were not associated with ANRS rules, 4 for LPV, and 8 for NFV. For NFV, ANRS misclassified 74 resistant profiles as being susceptible to the ARV. Sixty eight of the 74 sequences (92%) were classified as resistance with the inclusion of the eight new sequence positions. No change was found for LPV and a 78% improvement was associated with IDV. The above study shows that there is significant improvement in the prediction ability of ANRS gold standard.

Key words: ANRS, Artificial Intelligence, interpretation algorithms, ARV resistance prediction, machine learning

Running title

Machine learning and prediction of HIV drug resistance

Introduction

Developing countries are characterized by poor infrastructure and limited resources. The World Health organization has indicated that the current financing strategy, in many developing countries, does not meet the requirements for universal health care coverage. Thus, developing countries struggle under the burden of human immunodeficiency virus (HIV), Tuberculosis and Malaria (Yashik and Maurice, 2012). HIV is an incurable disease which affects the functioning of the immune system of a human over a long period of time. There are two known strains of the HIV, i.e. HIV-1 and HIV-2 (Wilson et al., 2002, The EuroGuidelines Group for HIV Resistance, 2003). The majority of HIV infection is due to HIV-1.

There are currently almost 6.4 million infected with HIV in South Africa, which is approximately 12.2% of the South African population (AIDS Committee of Actuarial Society of South Africa, 2011, Simbayi et al., 2014). Swaziland has a HIV infection prevalence of 26%, Botswana 25% and Lesotho 24% (Central Intelligence Agency, 2012). The burden of HIV in Africa is seen in the contrast with the prevalence of HIV in developed countries. France and Spain has a prevalence of 0.4% while Netherland has a prevalence of 0.2%. It is also estimated that there are almost 500 000 patients who exhibit AIDS defining conditions in South Africa (Health Systems Trust, 2011b).

HIV is managed by highly active antiretroviral therapy, which comprises of antiretroviral (Marques et al.) drugs from protease inhibitors, reverse transcriptase inhibitors; integrate inhibitors, fusion inhibitors, and entry inhibitors. However the success of managing HIV with ARV's is dependent on the actual treatment, stage of the disease, drug potency, patient adherence, achievable drug concentrations, drug resistance and toxic effects of the drugs (Richman and others, 2000, Singh and Mars, 2012b). Of these factors drug resistance is crucial, and is defined as the diminished ability of antiretroviral drugs to reduce the HIV viral load adequately (Singh and Mars, 2012b)

HIV drug resistance is inevitable due to selective pressure facilitated by the presence of ARVs during the management of HIV, high replication errors of the virus and initial infection (The EuroGuidelines Group for HIV Resistance, 2003) . Thus the ability to easily determine drug resistance is vital in the treatment of the HIV positive patients. HIV drug resistance is normally tested using phenotypic test (Meynard et al., 2002b).

In brief, phenotypic tests work by analyzing the concentration of ARV that is required to reduce the reproduction of a laboratory grown sample of the HIV that has infected a specific patient by 50%. The ratio of this concentration over the concentration required when using the wild type (original) HIV virus is called the IC50. The IC50 score is compared to cutoff values obtained from literature and is thus characterized as being either resistant to ARV drugs, susceptible to ARV drugs or intermediate resistance to the ARV drugs. Although the IC50 score is seen as the absolute measurement, laboratory based tests are relatively expensive, time consuming, susceptible to error and each test detects

resistance to a single drug and thus many tests are required to determine multiple drug resistances (Bartlett and others, 2004).

Electronic computerized algorithms (Toor et al., 2011) may also be used to determine ARV drug resistance, and have many advantages over phenotype testing. Computer based genotype interpretation algorithms usually determine mutations in the patient's pol gene and uses this information to determine which ARV drugs the patients are resistant to. Literature has associated mutations with particular resistance profiles. These computer based tests are faster and cheaper than phenotypic tests.

One widely used computer based interpretation algorithm was built by the French National Agency for AIDS Research AC11 Resistance group and is called the Nationale de Recherches sur le SIDA (ANRS). ANRS is seen as a gold standard in interpreting HIV drug resistance using mutations in the genomes. ANRS classifies ARV resistance according to three levels: susceptible, intermediate, and resistant. "Susceptible" indicates that a particular ARV drug will be effective against HIV; "intermediate" indicates that the ARV drug is partially effective; and if the ARV is not effective at all, it is classified "resistant".

. This algorithm is mainly based on correlation between drug resistance and virological outcome of patient who fail ARV treatment. The algorithm is based on a linear combination of mutations. If a particular mutation or a group of mutations are is present in the genome, the algorithm returns a resistance profile applicable to that particular sequence e.g.if the mutation A98S is present in the genome, resistance to the NVP is deduced, were as a E138K mutation will indicate intermediate resistance to the NVP drug (Meynard et al., 2002a)

Each rule consists of a Boolean expression. For example, an ANRS rule for abacavir (version 13, July 2005) states: "If there are five or more of the following RT mutations (M41L, D67N, L74V, M184V/I, L210W, T215Y/F), report resistance to abacavir." Both systems contain interpretations for all available antiretroviral drugs, including the fusion inhibitor (Liu and Shafer, 2006).

The ANRS system bases its interpretations almost entirely on genotype outcome studies and the ANRS has published a large proportion of the studies linking genotype to virological outcome, including studies on the genotypic predictors of response to abacavir (Katlama et al., 2000),tenofovir(Katlama et al., 2000), didanosine (Masquelier et al., 2004)and many more.

Computer based gold standard interpretation algorithms are usually based on an experts' understanding of the domain, available datasets and understanding of published literature. There therefore may be discrepancies and areas of improvement.

The average accuracy for ANRS was 59%,HIV-db 59% and REGA 61% (Singh and Mars, 2012a).It has been shown that there is no difference between the three algorithms in the

accuracies obtained (Poonpiriya et al., 2008). Various other studies (Snoeck et al., 2006b, Yebra et al., 2010) have shown there is no difference between accuracies, however, accuracy cannot be the sole factor for determining discrepancy between the three algorithms .Thus the aim of this study is to use machine learning to see if there can be improvements in the effectiveness of ANRS in predicting HIV drug resistance.

Materials and Methods

Development of the association matrix

Machine learning is literally an adaptive process whereby computers can improve by experience and analogy. Feature Selection ReliefF, MODTREE Filtering, FCBF Filtering and CFS Filtering were used to determine HIV drug resistance.

Feature Selection Relief is a supervised based on RELIEFF principle. This approach does not take into consideration the redundancy of the input attributes. Fast Correlation based Filter (FCBF) is a supervised feature selection algorithm based upon a filtering approach i.e processes the selection independently from the learning algorithm. This algorithm, unlike the ranking approaches takes into consideration the redundancy of the input attribute (Bouhamed et al., 2012). Correlation based feature Selection is also a supervised feature selection algorithms based upon a filtering approach i.e. processed the selection independently from the learning algorithm. The Multivalued Oblivious Decision tree (MODTREE) method is based on the notions of relevance and redundancy, through it does not use the same correlation measure, as it rests on the principle of pair-wise comparison. Lallich and Rakotomala(1999-2002)designed this (MODTREE) feature selection. The calculation is linear in number of observations n, even if the criterion is based on the principle of pair-wise comparisons. This makes it operational for processing databases involving a large number of lines. The partial correlation is applied to achieve the step-by-step "Forward" selection. It measures the correlation degree between two variables X and Y by subtracting the effect of a third variable Z (Rakotomalala and Lallich, 2002).

The open source software used to perform machine learning was Tanagra. Tanagra is a data mining suite build around graphical user interface. Tanagra is particularly strong in statistics, offering a wide range of uni- and multivariate parametric and nonparametric tests (Rajkumar and Reena, 2010).

The methods used in this paper are divided into three parts: data-processing, development of an association matrix, and the determination of the effectiveness of ANRS with the incorporation of the association matrix.

Data Processing

Free publically available and de-identified Genotype-Phenotype datasets that consisted of approximately 23000 protease gene and 23000 reverse transcriptase gene sequences were obtained from the Stanford HIV drug resistance database (http://hivdb.stanford.edu/).

These datasets were fed in the ANRS algorithm and a resistance measure was obtained for each sequence. The ANRS result was then compared to the known resistance measure obtained from laboratory test for each sequence, and the accuracy of the ANRS algorithm was calculated.

Effectiveness of ANRS with the incorporation of the association matrix

The association matrix for each drug was added to the rules of the ANRS algorithm. This new model was then applied to the testing dataset and the changes in the ability to predict HIV drug resistance was analyzed.

Results

The study showed considerable improvement in predicting HIV drug resistance using machine learning against gold standard ANRS. Table 1 shows the important mutations that contribute to resistance for each of the ARV's. Table 2 shows the correctly classified sequences out of the total sequences and the accuracy in percentage of all the PR ARVs. *RS = the sequence was incorrectly classified as resistant instead of susceptible. SR =the sequence was incorrectly classified as susceptible instead of resistant. Table3 shows the correctly classified sequences out of the total sequences and the accuracy in percentage of all the RT ARVs. *RS = the sequence was incorrectly classified as resistant instead of susceptible. SR = the sequence was incorrectly classified as susceptible instead of resistant.

Discussion

Results show that the ANRS gold standard can be improved in predicting HIV drug resistance in all ten ARV drugs tested. Table one shows the mutations not present in the ANRS algorithm for each of the 10 ARV drugs. Some of the major contributors to predicting HIV drug resistance for protease ARV drugs, using the feature selection algorithms, were P63, P57, P82, and P69. However, the HIVdb drug resistance database has only P82 in its major mutation list. It clearly shows machine learning also picked up those mutations that were not listed in the HIVdb database. Similarly P30, P35, P142, and P83 were identified as important mutations for RT ARV drugs. The ten most ANRS algorithm in order to determine if there is any change to the ability of ANRS to predict susceptibility and resistance to ARV drugs.

An average of 85 PI sequences that were supposed to be interpreted as resistance, were classified as susceptible according to ANRS. Adding the rules derived from the machine

learning algorithm, results in an 88 ± 7.1% improvement in the overall accuracy. A T-test was performed to determine if the improvement was due to random chance, and a $p < 0.001$ was obtained. This indicates that there is a statistically significant improvement in the prediction of susceptibility measures for the five protease drugs.

An average of 31 PI sequences was wrongly classified as susceptible instead of resistant. Adding the rules derived from the machine learning algorithm, results in an 83 ± 12.1% improvement in the overall accuracy. A T-test was performed to determine if the improvement was due to random chance, and a $p < 0.004$ was obtained. This indicates that there is a statistically significant improvement in the prediction of resistance measures for the five protease drugs.

Nearly 130 sequences were wrongly put under susceptible, which were actually resistant. Using the machine learning algorithm rules, results in a 69 ± 10.9% improvement in the overall accuracy. A T-test was performed to determine if the improvement was due to random chance, and a p value of 0.004 was obtained. This indicates that there is a statistically significant improvement in the prediction of resistance measures for the five RT drugs

An average of 147 PI sequences was wrongly classified as susceptible instead of resistant. Adding the rules derived from the machine learning algorithm, results in an 80 ± 5.9% improvement in the overall accuracy. A T-test was performed to determine if the improvement was due to random chance, and a $p < 0.004$ was obtained. This indicates that there is a statistically significant improvement in the prediction of resistance measures for the five RT drugs. Table 4 shows the positive predictive value (PPV) and negative predictive value (NPV) of all the PR ARVs used in the study. * Z-Score is > 1.98, indicating there is a statically significant difference when adding the association matrix. Table 5 shows the PPV and NPV of predicting HIV resistance for PR ARV drugs for both the ANRS algorithm alone and when the machine learning mutations are incorporated into them. The PPV improved by 27% while the NPV improved by 16%. These results show that the incorporation of the machine learning mutation does positively influence the ability of ANRS to predict RT ARV drug resistance. * Z-Score is > 1.98, indicating there is a statically significant difference when adding the association matrix

Conclusions

The above study shows that there is significant improvement in the prediction ability of ANRS gold standard. On average the ARNS algorithm was improved by 79% ±6.6. The positive predictive value improved by 28% and the negative predicative value improved by 10%. Some of the major contributors to predicting HIV drug resistance for protease ARV drugs, using the feature selection algorithms, were P63, P57, P82, and P69. Similarly P30, P35, P142, and P83 were identified as important mutations for RT ARV drugs. These indicate that the ANRS gold standard has its limitations, which can be improved. Future studies may include using other machine learning algorithms like support vector machines and Bayesian networks. A larger dataset will be of benefit.

Acknowledgements

This work was supported by the National Institute of Health Fogarty International Centre (grant number: 4D43TW007004-13).

Conflict of interest statement

The authors declare no conflict of interest.

References

AIDS Committee of Actuarial Society of South Africa. (2011). ASSA2008 Model: ProvOutput. Retrieved from http://aids.actuarialsociety.org.za/ASSA2008-Model-3480.htm

Bartlett, J. C., & others. (2004). Medical Management of HIV infection: John Hopkins University School of Medicine.

Central Intelligence Agency. (2012). HIV/AIDS - adult prevalence rate The World Factbook

Health Systems Trust. (2011). Percentage of deaths due to AIDS Health Indicators, Statistics South Africa: Statistical release P0302 Mid-year estimates. from http://indicators.hst.org.za/indicators/StatsSA/

Katlama, C., Clotet, B., Plettenberg, A., Jost, J., Arasteh, K., Bernasconi, E., Jeantils, V., Cutrell, A., Stone, C. & Ait-khaled, M. 2000. The role of abacavir (ABC, 1592) in antiretroviral therapy-experienced patients: results from a randomized, double-blind, trial. Aids, 14**,** 781-789.

Liu, T. F. & Shafer, R. W. 2006. Web resources for HIV type 1 genotypic-resistance test interpretation. Clinical infectious diseases, 42**,** 1608-1618

Meynard, J. L., Vray, M., Morand-Joubert, L., Race, E., Descamps, D., Peytavin, G., Matheron, Sophie; Lamotte, Claire; Guiramand, Sonia; Costagliola, Dominique; Brun-Vézinet, Françoise; Clavel, François Girard, P. M. (2002). Phenotypic or genotypic resistance testing for choosing antiretroviral therapy after treatment failure: a randomized trial. AIDS, 16(5), 727-736.

Rajkumar, A. & Reena, G. S. 2010. Diagnosis of heart disease using datamining algorithm. Global journal of computer science and technology, 10**,** 38-43.

Richman, D., & others. (2000). A practical guide to HIV drug resistance and its implications for antiretroviral treatment strategies: International Medical Press.

SimbayI, L., Shisana, O., Rehle, T., Onoya, D., Jooste, S., Zungu, N. &

Zuma, K. 2014. South African national HIV prevalence, incidence and behaviour survey, 2012. Pretoria: Human Sciences Research Council.

Singh, Y., & Mars, M. (2012). Predicting a single HIV drug resistance measure from three international interpretation gold standards. Asian Pacific Journal of Tropical Medicine, 5(7), 566-572. doi: http://dx.doi.org/10.1016/S1995-7645(12)60100-X

The EuroGuidelines Group for HIV Resistance. (2003). Clinical and laboratory guidelines for the use of HIV-1 drug resistance testing as part of treatment management: recommendations for the European setting. AIDS, 15(3), 309-320.

Toor, J. S., Sharma, A., Kumar, R., Gupta, P., Garg, P., & Arora, S. K. (2011). Prediction of drug-resistance in HIV-1 subtype C based on protease sequences from ART naive and first-line treatment failures in North India using genotypic and docking analysis. Antiviral Research, 92(2), 213-218. doi: 10.1016/j.antiviral.2011.08.005

Wilson, D., Naidoo, S., Bekker, l., Cotton, M., & Maartens, G. (2002). Handbook of HIV Medicine. Southern Africa: University Press

Yashik, S., & Maurice, M. (2012). Predicting a single HIV drug resistance measure from three international interpretation gold standards. Asian Pacific Journal of Tropical Medicine, 5(7), 566-572. doi: 10.1016/s1995-7645(12)60100-x

Tables

IDV	LPV	NFV	SQV	TPV	ABC	DDI	EFV	NVP	TDF
P 10	P 83	P 62	P 71	P 84	P 184	P 34	P 231	P 41	P 33
P 63		P 63	P 54	P 54	P 231	P 17	P 215	P 74	P 25
P 57		P 20	P 46	P 33	P 103	P 26	P 184	P 221	P 26
P 88		P 69	P 63	P 47	P 38	P 116	P 41	P 108	P 34
P 69		P 14	P 36	P 46	P 83	P 15	P 35	P 214	P 37
P 30		P 12	P 50	P 10	P 211	P 12	P 67	P 219	P 30
P 14			P 57	P 71	P 214	P 37	P 247	P 184	P 19
P 70			P 69	P 13	P 232	P 27	P 102	P 35	P 122
P 50			P 59	P 82	P 135	P 67	P 214	P 135	P 118
P 78			P 14	P 63	P 177	P 74	P 133	P 36	P 142

Table 1: Shows the important mutations that contribute to resistance for each of the ARVs.

Drug	Classification error*	No.of incorrectly classified sequences using ANRS	No. of incorrectly classified sequences using ANRS and Association matrix	% Improvement
IPV	RS	14	1	92.85
IPV	SR	32	3	90.62
IDV	RS	201	47	76.6
IDV	SR	9	3	66.6
NFV	RS	74	6	91.8
NFV	SR	25	6	76
SQV	RS	128	8	93.7
SQV	SR	68	18	83.3
TPV	RS	7	1	85.7
TPV	SR	22	0	100

Table 2: The above table shows the correctly classified sequences out of the total sequences and the accuracy in percentage of all the PR ARVs. *RS = the sequence was incorrectly classified as resistant instead of susceptible. SR =the sequence was incorrectly classified as susceptible instead of resistant.

Drug	Classification error*	No. of incorrectly classified sequences using ANRS	No. of incorrectly classified sequences using ANRS and Association matrix	% Improvement
ABC	RS	55	13	76.36
ABC	SR	206	28	86.4
DDI	RS	165	29	82.42
DDI	SR	165	44	73.33
EFV	RS	233	96	58.79
EFV	SR	131	20	84.73
NVP	RS	170	73	57.05
NVP	SR	177	45	74.57
TDF	RS	25	8	68
TDF	SR	55	10	81.81

Table 3: The above table shows the correctly classified sequences out of the total sequences and the accuracy in percentage of all the RT ARVs. *RS = the sequence was incorrectly classified as resistant instead of susceptible. SR = the sequence was incorrectly classified as susceptible instead of resistant.

ARV	INTIAL % PPV	% PPV WITH ASSOSIATION MATRIX	NPV	NPV WITH ASSOCIATION MATRIX
IDV	70	92*	98	99*
LPV	97	100*	93	99*
NFV	88	99*	96	99*
SQV	74	98*	94	99*
TPV	0	86*	95	100*

Table 4: The above table shows the positive predictive value (PPV) and negative predictive value (NPV) of all the PR ARVs used in the study. * Z-Score is > 1.98, indicating there is a statically significant difference when adding the association matrix

.

ARV	INITIAL % PPV	% PPV WITH ASSOSIATION MATRIX	NPV	NPV WITH ASSOCIATION MATRIX
ABC	76	94*	74	96*
DDI	42	85*	81	95*
EFC	66	71*	82	97*
NPV	80	97*	78	94*
IDF	24	75*	84	97*

Table 5 shows the PPV and NPV of predicting HIV resistance for PR ARV drugs for both the ANRS algorithm alone and when the machine learning mutations are incorporated into them. The PPV improved by 27% while the NPV improved by 16%. These results show that the incorporation of the machine learning mutation does positively influence the ability of ANRS to predict RT ARV drug resistance. * Z-Score is > 1.98, indicating there is a statically significant difference when adding the association matrix

.

CHAPTER FOUR

SYNTHESIS, CONCLUSION AND RECOMMENATION

SYNTHESIS, CONCLUSION AND RECOMMENDATION

The high mutation rate of HIV drug under selective pressure leads to drug resistance and treatment failure (Cuevas et al., 2015a) The dissertation provides following important points

I. The use of current high speed technologies and computational methods has recently contributed in predicting HIV drug resistance.

II. The interpretation algorithms like REGA, HIV-db and Agence Nationale de Recherches sur le SIDA (ANRS) are logic or decision tree based, however, these were developed using different data subsets on experienced and drug niave patients (Yashik and Maurice, 2012).

III. The key findings of our study are that there are chances of improvement in the gold standard in predicting HIV drug resistance, almost an 80 ± 5.9% improvement in the overall accuracy.

The above study shows that there is significant improvement in the prediction ability of ANRS gold standard. On average the ARNS algorithm was improved by 79% ±6.6. The positive predictive value improved by 28% and the negative predicative value improved by 10%. Some of the major contributors to predicting HIV drug resistance for protease ARV drugs, using the feature selection algorithms, were P63, P57, P82, and P69. Similarly P30, P35, P142, and P83 were identified as important mutations for RT ARV drugs. These indicate that the ANRS gold standard has its limitations, which can be improved. Future studies may include using other machine learning algorithms like support vector machines and Bayesian networks. A larger dataset will be of benefit.

Conclusion

The study shows that there are great amount of improvement which can be done in the ANRS. Although ANRS is regarded as the gold standard, there are certain rules which can still be improved.

There were quite a few sequences which were note correctly noted in ANRS which can be changed to improve the testing of HIV drug resistance.

This study can help increase the effectiveness of the HIV treatment throughout the world, especially in developing countries. When the resistance is determined beforehand it will certainly give doctors a better option for treatment.

The most challenging portion was data processing and analyzing the results of the machine learning algorithms. The sheer volume of the neucliotites that were analysed in order to create the input space possed a lenge. Trying to map a large unprocessed input space into a smaller input space created many issues. The smaller input space was required for the machine learning algroithm to better map the learning process. However, if the input space was too small, it would have been so simplistic to represent the large unprocessed input space. Hence there would have been valuable connections and better data points that would have been lost.

Generally, computer science articles do not process the results of the learning algorithms statistically. This was a challenge as very few papers explain how to statistically compare results of machine learning.

This study gave me an appreciation of the statistical components of research, and also allowed me to understand the impact of data, data processing and data analysis from a machine learning point of view.

Limitations

Among the different subtypes, subtype C is predominant in developing countries, however due to the availability of data subtype B is most present in developed counties was used in the study. Validations of the algorithms were done using publicly available data from Standford University HIV db.

Future recommendations

Machine learning requires large dataset. As such a larger dataset, particularly with a greater number of subtype C will be beneficial. Further investigation is required using other machine learning techniques e.g. support vector machine, gene expression, random forest tree. Interpretation may be improved by adding interpretation from other algorithms.

This will play a huge part economically. As countries with limited resources will not have to waste the ARVs. It will help in the treatment of the patient, as it will help the doctors to choose the drugs in advance and not the current way of testing different drugs on single patient.

The most important thing is that developing and economically weak nations will have the benefit of utilizing their medication properly, which will play a huge part in HIV management.

CHAPTER FIVE

REFERENCES

REFERENCES

ABECASIS, A. B., LEMEY, P., VIDAL, N., DE OLIVEIRA, T., PEETERS, M., CAMACHO, R., SHAPIRO, B., RAMBAUT, A. & VANDAMME, A.-M. 2007. Recombination confounds the early evolutionary history of human immunodeficiency virus type 1: subtype G is a circulating recombinant form. *Journal of virology,* 81**,** 8543-8551.

AC11 RESISTANCE GROUP. 2013. *ANRS: Table of Rules* [Online]. France: National Agency for AIDS Research. Available: http://www.hivfrenchresistance.org/2013/Algo-sep-2013.pdf [Accessed September 2014].

ADAM, M. A. & JOHNSON, L. F. 2009. Estimation of adult antiretroviral treatment coverage in South Africa. . *South African Medical Journal* 99**,** 6.

AIDS COMMITTEE OF ACTUARIAL SOCIETY OF SOUTH AFRICA 2008. ASSA2008 Model: ProvOutput.

AIDS COMMITTEE OF ACTUARIAL SOCIETY OF SOUTH AFRICA 2011. ASSA2008 Model: ProvOutput.

ARTS, E. J. & HAZUDA, D. J. 2012. HIV-1 antiretroviral drug therapy. *Cold Spring Harbor perspectives in medicine,* 2**,** a007161.

BAGGALEY, R. F., BOILY, M.-C., WHITE, R. G. & ALARY, M. 2006. Risk of HIV-1 transmission for parenteral exposure and blood transfusion: a systematic review and meta-analysis. *Aids,* 20**,** 805-812.

BARTLETT, J. C. & OTHERS 2004. *Medical Management of HIV infection,* Johm Hopkins University School of Medicine.

BEERENWINKEL, N. & OTHERS 2002. Diversity and complexity of HIV-1 drug resistance: A bioinformatics approach to predicting phenotype from genotype. *National Academy of Science.*

BEERENWINKEL, N., SING, T., LENGAUER, T., RAHNENFÜHRER, J., ROOMP, K., SAVENKOV, I., FISCHER, R., HOFFMANN, D., SELBIG, J. & KORN, K. 2005. Computational methods for the design of effective therapies against drug resistant HIV strains. *Bioinformatics,* 21**,** 3943-3950.

BOILY, M.-C., BAGGALEY, R. F., WANG, L., MASSE, B., WHITE, R. G., HAYES, R. J. & ALARY, M. 2009. Heterosexual risk of HIV-1 infection per sexual act: systematic review and meta-analysis of observational studies. *The Lancet infectious diseases,* 9**,** 118-129.

BONET, I., GARCÍA, M. M., SAEYS, Y., VAN DE PEER, Y. & GRAU, R. 2007. Predicting Human Immunodeficiency Virus (HIV) drug resistance using recurrent neural networks. *Bio-inspired Modeling of Cognitive Tasks.* Springer.

BOUHAMED, H., LECROQ, T. & REBAÏ, A. 2012. New Filter method for categorical variables' selection. *International Journal of Computer Science Issues,* 9**,** 10-19.

CAMPBELL-YESUFU, O. T. & GANDHI, R. T. 2011. Update on human immunodeficiency virus (HIV)-2 infection. *Clinical infectious diseases,* 52**,** 780-787.

CAMPBELL, C., NAIR, Y., MAIMANE, S. & SIBIYA, Z. 2008. Supporting people with AIDS and their carers in rural South Africa: possibilities and challenges. *Health Place,* 14**,** 507-18.

CANN, A. J. & KARN, J. 1989. Molecular biology of HIV: new insights into the virus life-cycle. *Aids,* 3**,** S19-34.

CARVALHO, A. R. & PINTO, C. M. 2016. Emergence of drug-resistance in HIV dynamics under distinct HAART regimes. *Communications in Nonlinear Science and Numerical Simulation,* 30**,** 207-226.

CENTRAL INTELLIGENCE AGENCY 2012. HIV/AIDS - adult prevalence rate *The World Factbook*

CHONG, H., WU, X., SU, Y. & HE, Y. 2016. Development of potent and long-acting HIV-1 fusion inhibitors. *Aids,* 30**,** 1187-1196.

CLAVEL , F. & HANCE , A. J. 2004. HIV Drug Resistance. *New England Journal of Medicine,* 350**,** 1023-1035.

CUEVAS, J. M., GELLER, R., GARIJO, R., LOPEZ-ALDEGUER, J. & SANJUAN, R. 2015a. Extremely High Mutation Rate of HIV-1 In Vivo. *PLoS Biol,* 13**,** e1002251.

DAVID, R. 2009. HIV: Selective pressure. *Nat Rev Immunol,* 9**,** 459-459.

DE COCK, K. M. & BRUN-VÉZINET, F. 1989. Epidemiology of HIV-2 infection. *AIDS,* 3**,** S89-96.

DE LUCA, A., COZZI-LEPRI, A. & PERNO, C. F. 2004. Variability in the interpretation of transmitted genotypic HIV-1 drug resistance and prediction of virological outcomes of the initial HAART by distinct systems. *Antiretroviral Therapy,* 9**,** 743-752.

DE OLIVEIRA, T., DEFORCHE, K., CASSOL, S., SALMINEN, M., PARASKEVIS, D., SEEBREGTS, C., SNOECK, J., VAN RENSBURG, E. J., WENSING, A. M., VAN DE VIJVER, D. A., BOUCHER, C. A., CAMACHO, R. & VANDAMME, A. M. 2005. An automated genotyping system for analysis of HIV-1 and other microbial sequences. *Bioinformatics,* 21**,** 3797-800.

DEL ROMERO, J., MARINCOVICH, B., CASTILLA, J., GARCÍA, S., CAMPO, J., HERNANDO, V. & RODRÍGUEZ, C. 2002. Evaluating the risk of HIV transmission through unprotected orogenital sex. *Aids,* 16**,** 1296-1297.

DRĂGHICI, S. & POTTER, R. B. 2003. Predicting HIV drug resistance with neural networks. *Bioinformatics,* 19**,** 98-107.

DURANT, J., CLEVENBERGH, P., HALFON, P., DELGIUDICE, P., PORSIN, S., SIMONET, P., MONTAGNE, N., BOUCHER, C. A. B., SCHAPIRO, J. M. & DELLAMONICA, P. 1999. Drug-resistance genotyping in HIV-1 therapy: the VIRAD APT randomi sed controlled trial. *The Lancet,* 353**,** 2195-2199.

FARIA, N. R., RAMBAUT, A., SUCHARD, M. A., BAELE, G., BEDFORD, T., WARD, M. J., TATEM, A. J., SOUSA, J. D., ARINAMINPATHY, N., PÉPIN, J., POSADA, D., PEETERS, M., PYBUS, O. G. & LEMEY, P. 2014. The early spread and epidemic ignition of HIV-1 in human populations. *Science,* 346**,** 56-61.

FEDER, A. F., RHEE, S.-Y., SHAFER, R. W., PETROV, D. A. & PENNINGS, P. S. 2015. More efficacious drugs lead to harder selective sweeps in the evolution of drug resistance in HIV-1. *bioRxiv*.

FOWLER, B. J., GELFAND, B. D., KIM, Y., KERUR, N., TARALLO, V., HIRANO, Y., AMARNATH, S., FOWLER, D. H., RADWAN, M. & YOUNG, M. T. 2014. Nucleoside reverse transcriptase inhibitors possess intrinsic anti-inflammatory activity. *Science,* 346**,** 1000-1003.

FRENTZ, D., BOUCHER, C. A. B., ASSEL, M., DE LUCA, A., FABBIANI, M., INCARDONA, F., LIBIN, P., MANCA, N., MÜLLER, V., NUALLÁIN, B. Ó., PAREDES, R., PROSPERI, M., QUIROS-ROLDAN, E., RUIZ, L., SLOOT, P. M. A., TORTI, C., VANDAMME, A.-M., VAN LAETHEM, K., ZAZZI, M. & VAN DE VIJVER, D. A. M. C. 2010. Comparison of HIV-1 Genotypic Resistance Test Interpretation Systems in Predicting Virological Outcomes Over Time. *PLoS ONE,* 5**,** e11505.

GALLO , R. C. & MONTAGNIER , L. 2003. The Discovery of HIV as the Cause of AIDS. *New England Journal of Medicine,* 349**,** 2283-2285.

GARRIGA, C. & MENÉNDEZ-ARIAS, L. 2006. DR_SEQAN: a PC/Windows-based software to evaluate drug resistance using human immunodeficiency virus type 1 genotypes. *BMC infectious diseases,* 6**,** 44.

GHOSH, A. K., OSSWALD, H. L. & PRATO, G. 2016. Recent progress in the development of HIV-1 protease inhibitors for the treatment of HIV/AIDS. *Journal of medicinal chemistry,* 59**,** 5172-5208.

GILKS, C. F., CROWLEY, S., EKPINI, R., GOVE, S., PERRIENS, J., SOUTEYRAND, Y., SUTHERLAND, D., VITORIA, M., GUERMA, T. & DE COCK, K. 2006. The WHO public-health approach to antiretroviral treatment against HIV in resource-limited settings. *The Lancet,* 368**,** 505-510.

GINSBURG, G. S. & MCCARTHY, J. J. 2001. Personalized medicine: revolutionizing drug discovery and patient care. *Trends in Biotechnology,* 19**,** 491-496.

HALES, G., BIRCH, C., CROWE, S., WORKMAN, C., HOY, J. F., LAW, M. G., KELLEHER, A. D., LINCOLN, D. & EMERY, S. 2006. A randomised trial comparing genotypic and virtual phenotypic interpretation of HIV drug resistance: the CREST study. *PLoS Clin Trials,* 1**,** e18.

HEALTH SYSTEMS TRUST. 2011a. *Aids sick: Indicator data* [Online]. Available: http://indicators.hst.org.za/healthstats/85/data [Accessed October 2011.

HEALTH SYSTEMS TRUST. 2011b. *Percentage of deaths due to AIDS* [Online]. Available: http://indicators.hst.org.za/indicators/StatsSA/.

HIRSCH, M. S., CONWAY, B., D'AQUILA, R. T. & ET AL. 1998. Antiretroviral drug resistance testing in adults with hiv infection: Implications for clinical management. *JAMA,* 279**,** 1984-1991.

HIV/AIDS, J. U. N. P. O. 2014. The Gap Report 2014: people living with HIV. *Retrieved November,* 28**,** 2014.

HTTP://I-BASE.INFO/CATEGORY/PUBLICATIONS/.

JIAMSAKUL, A., CHAIWARITH, R., DURIER, N., SIRIVICHAYAKUL, S., KIERTIBURANAKUL, S., VAN DEN EEDE, P., DITANGCO, R., KAMARULZAMAN, A., LI, P. C. & RATANASUWAN, W. 2016. Comparison of genotypic and virtual phenotypic drug resistance interpretations with laboratory-based phenotypes among CRF01_AE and subtype B HIV-infected individuals. *Journal of medical virology,* 88**,** 234-243.

JOHN, G., BARTLETT, M. & JOEL, E. 2000. Medical management of HIV infection. *John Hopkins University***,** 4-5.

KANKI, P. J., HAMEL, D. J., SANKALE, J. L., HSIEH, C., THIOR, I., BARIN, F., WOODCOCK, S. A., GUEYE-NDIAYE, A., ZHANG, E., MONTANO, M., SIBY, T., MARLINK, R., I, N. D., ESSEX, M. E. & S, M. B. 1999. Human immunodeficiency virus type 1 subtypes differ in disease progression. *J Infect Dis,* 179**,** 68-73.

KATLAMA, C., CLOTET, B., PLETTENBERG, A., JOST, J., ARASTEH, K., BERNASCONI, E., JEANTILS, V., CUTRELL, A., STONE, C. & AIT-KHALED, M. 2000. The role of abacavir (ABC, 1592) in antiretroviral therapy-experienced patients: results from a randomized, double-blind, trial. *Aids,* 14**,** 781-789.

KATLAMA, C., DEEKS, S. G., AUTRAN, B., MARTINEZ-PICADO, J., VAN LUNZEN, J., ROUZIOUX, C., MILLER, M., VELLA, S., SCHMITZ, J. E., AHLERS, J., RICHMAN, D. D. & SEKALY, R. P. 2013. Barriers to a cure for HIV: new ways to target and eradicate HIV-1 reservoirs. *The Lancet,* 381**,** 2109-2117.

LIU, L., MAY, S., RICHMAN, D. D., HECHT, F. M., MARKOWITZ, M., DAAR, E. S., ROUTY, J.-P., MARGOLICK, J. B., COLLIER, A. C. & WOELK, C. H. 2008. Comparison of

algorithms that interpret genotypic HIV-1 drug resistance to determine the prevalence of transmitted drug resistance. *AIDS (London, England),* 22**,** 835.
LIU, T. F. & SHAFER, R. W. 2006. Web resources for HIV type 1 genotypic-resistance test interpretation. *Clinical infectious diseases,* 42**,** 1608-1618.
LODISH, H., BERK, A., ZIPURSKY, S. L., MATSUDAIRA, P., BALTIMORE, D. & DARNELL, J. 2000. Mutations: types and causes. *Molecular Cell Biology,* 4.
MARQUES, F. K., CAMPOS, F. M. F., SOUSA, L. P., TEIXEIRA-CARVALHO, A., DUSSE, L. M. S. & GOMES, K. B. 2013. Association of microparticles and preeclampsia. *Molecular Biology Reports,* 40**,** 4553-4559.
MASQUELIER, B., TAMALET, C., MONTÈS, B., DESCAMPS, D., PEYTAVIN, G., BOCKET, L., WIRDEN, M., IZOPET, J., SCHNEIDER, V. & FERRÉ, V. 2004. Genotypic determinants of the virological response to tenofovir disoproxil fumarate in nucleoside reverse transcriptase inhibitor-experienced patients. *Antiviral therapy,* 9**,** 315-324.
MESPLÈDE, T., QUASHIE, P. K., ZANICHELLI, V. & WAINBERG, M. A. 2014. Integrase strand transfer inhibitors in the management of HIV-positive individuals. *Annals of medicine,* 46**,** 123-129.
MEYNARD, J.-L., VRAY, M., MORAND-JOUBERT, L., RACE, E., DESCAMPS, D., PEYTAVIN, G., MATHERON, S., LAMOTTE, C., GUIRAMAND, S. & COSTAGLIOLA, D. 2002a. Phenotypic or genotypic resistance testing for choosing antiretroviral therapy after treatment failure: a randomized trial. *Aids,* 16**,** 727-736.

MITTON, J. 2000a. The sociological spread of HIV/AIDS in South Africa. *Journal of the association of Nurses in AIDS care,* 11**,** 17-26.
MOODLEY, P. 2006. HIV workshop. *Society of Medical Laboratory technologists of South Africa and Department of Virology Presentation*.
MOUTOUH, L., CORBEIL, J. & RICHMAN, D. D. 1996. Recombination leads to the rapid emergence of HIV-1 dually resistant mutants under selective drug pressure. *Proceedings of the National Academy of Sciences,* 93**,** 6106-6111.
PASOMSUB, E., SUKASEM, C., SUNGKANUPARH, S., KIJSIRIKUL, B. & CHANTRATIA, W. 2010. The application of artificial neural networks for phenotypic drug resistance prediction: evaluation and comparison with other interpretation systems. *Janapese Journal of infectious diseases,* 63**,** 87-94.
PERELSON, A. S., NEUMANN, A. U., MARKOWITZ, M., LEONARD, J. M. & HO, D. D. 1996. HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science,* 271**,** 1582-6.
PIERRET, J. 2007. An analysis over time of the experiences of living with HIV *Social Science Medicine,* 65**,** 1595-1605.
POONPIRIYA, V., SUNGKANUPARPH, S., LEECHANACHAI, P., PASOMSUB, E., WATITPUN, C., CHUNHAKAN, S. & CHANTRATITA, W. 2008. A study of seven rule-based algorithms for the interpretation of HIV-1 genotypic resistance data in Thailand. *Journal of virological methods,* 151**,** 79-86.
QUINN, T. C. 1998. Molecular variants of HIV-1 and their impact on vaccine development. *Int J STD AIDS,* 9 Suppl 1**,** 2.
RAJKUMAR, A. & REENA, G. S. 2010. Diagnosis of heart disease using datamining algorithm. *Global journal of computer science and technology,* 10**,** 38-43.
RAKOTOMALALA, R. & LALLICH, S. 2002. Construction d'arbres de décision par optimisation. *Revue d'intelligence artificielle,* 16**,** 685-703.
RICHMAN, D. & STASZEWSKI, S. 2000. *HIV drug resistance and its implications for antiretroviral treatment strategies,* Internat. Medical Press.

ROBERTSON, D., ANDERSON, J., BRADAC, J., CARR, J., FOLEY, B., FUNKHOUSER, R., GAO, F., HAHN, B., KALISH, M. & KUIKEN, C. 2000. HIV-1 nomenclature proposal. *Science,* 288**,** 55-55.

SIMBAYI, L., SHISANA, O., REHLE, T., ONOYA, D., JOOSTE, S., ZUNGU, N. & ZUMA, K. 2014. South African national HIV prevalence, incidence and behaviour survey, 2012. *Pretoria: Human Sciences Research Council*.

SINGH, Y. & MARS, M. 2012a. HIV Drug-Resistant Patient Information Management, Analysis, and Interpretation. *JMIR research protocols,* 1.

SINGH, Y. & MARS, M. 2012b. Predicting a single HIV drug resistance measure from three international interpretation gold standards. *Asian Pacific Journal of Tropical Medicine,* 5**,** 566-572.

SINGH, Y. & MARS, M. 2014. An investigation into the comparison of three human immunodeficiency virus (HIV) drug resistance interpretation algorithms. *African Journal of Microbiology Research,* 8**,** 3710-3715.

SNOECK, J., KANTOR, R., SHAFER, R. W., VAN LAETHEM, K., DEFORCHE, K., CARVALHO, A. P., WYNHOVEN, B., SOARES, M. A., CANE, P. & CLARKE, J. 2006b. Discordances between interpretation algorithms for genotypic resistance to protease and reverse transcriptase inhibitors of human immunodeficiency virus are subtype dependent. *Antimicrobial agents and chemotherapy,* 50**,** 694-701.

TANG, M. W., LIU, T. F. & SHAFER, R. W. 2012. The HIVdb system for HIV-1 genotypic resistance interpretation. *Intervirology,* 55**,** 98-101.

TANG, M. W. & SHAFER, R. W. 2012. HIV-1 Antiretroviral Resistance. *Drugs,* 72**,** e1-e25. THE EUROGUIDELINES GROUP FOR HIV RESISTANCE 2003. Clinical and laboratory guidelines for the use of HIV-1 drug resistance testing as part of treatment management: recommendations for the European setting. *AIDS,* 15**,** 309-320.

TOMIMATSU, T., MIMURA, K., ENDO, M., KUMASAWA, K. & KIMURA, T. 2016. Pathophysiology of preeclampsia: an angiogenic imbalance and long-lasting systemic vascular dysfunction. *Hypertension Research.*

TOOR, J. S., SHARMA, A., KUMAR, R., GUPTA, P., GARG, P. & ARORA, S. K. 2011. Prediction of drug-resistance in HIV-1 subtype C based on protease sequences from ART naive and first-line treatment failures in North India using genotypic and docking analysis. *Antiviral Research,* 92**,** 213-218.

TOWNSEND, C. L., CORTINA-BORJA, M., PECKHAM, C. S., DE RUITER, A., LYALL, H. & TOOKEY, P. A. 2008. Low rates of mother-to-child transmission of HIV following effective pregnancy interventions in the United Kingdom and Ireland, 2000–2006. *Aids,* 22**,** 973-981.

USACH, I., MELIS, V. & PERIS, J.-E. 2013. Non-nucleoside reverse transcriptase inhibitors: a review on pharmacokinetics, pharmacodynamics, safety and tolerability. *Journal of the International AIDS Society,* 16.

VERCAUTEREN, J. & VANDAMME, A. 2006. Algorithms for the interpretation of HIV-1 genotypic drug resistance information. *Antiviral Research,* 71**,** 335-342.

VERGNE, L., SNOECK, J., AGHOKENG, A. & OTHERS 2006. Genotypic drug resistance interpretation algorithms display high levels of discordance when applied to non-B strains from HIV-1 naive and treated patients. *FEMS Immunology and Medical Microbiology,* 46**,** 53-62.

VITINGHOFF, E., DOUGLAS, J., JUDON, F., MCKIMAN, D., MACQUEEN, K. & BUCHINDER, S. P. 1999. Per-contact risk of human immunodificiency virus tramnsmision between male sexual partners. *American journal of epidemiology,* 150**,** 306-311.

WILSON D, NAIDOO S, BEKKER L, COTTON M & G., M. 2002. *Handbook of Hiv Medicine,* Southern Africa, University Press

YEBRA, G., DE MULDER, M., DEL ROMERO, J., RODRÍGUEZ, C. & HOLGUIN, A. 2010. HIV-1 non-B subtypes: High transmitted NNRTI-resistance in Spain and impaired genotypic resistance interpretation due to variability. *Antiviral research,* 85**,** 409-417.