RHODES UNIVERSITY

Grahamstown • 6140 • South Africa

Influence of Non-Synonymous Sequence Mutations on the Architecture of HIV-1 Clade C Protease Receptor Site: Docking and Molecular Dynamics Studies

---

A mini-thesis submitted in partial fulfilment of the requirements for a degree of

MASTER OF SCIENCE OF RHODES UNIVERSITY

by

Coursework/Thesis

in

Bioinformatics and Computational Molecular Biology

In the Department of Biochemistry, Microbiology and Biotechnology

Faculty of Science

by

HARRIS ONYWERA

February 2013

DECLARATION

I, HARRIS ONYWERA, hereby declare that this thesis is a product of my original work and has not been submitted for a diploma or degree in any other university or college.

Signature: **Date:**

…… Wed., 6th February 2013

……

DEDICATION

A selfless reward, refining the zeal within…

To my finest friend, Bethwel Kiplagat Tanui, this is a selfless reward from the tune of that bitter past, the past that axed your life. Let this be a selfless reward refining the zeal within…

To my big family, it might not have been our year, but in spite of all the boundaries and bleakness, I often found solace in the river of your words, when I reflected upon them…

In my story of life, I've met distinct personalities, besides my kin, who sacrificed their time and resources to offer me both glee and encouragement, especially when sadness kept on staining my face. By chance, I might have treated you with floods of contempt, but you are still in my thoughts; Dr. Joseph Anejo-Okopi, and Messrs. Felix Lubwa and Harrison Fredrick. "Auntie" Martha wa Njoka, you've been a prophetess, and you prayed for my endeavours…

To all those who have suffered, are suffering and will suffer of HIV/AIDS, this one is for you again. To the infected and or affected, there is still a bunch of refined hope for us all…

Above all, I'm still in love with the Lamb of GOD, the King that was nailed to a tree. There is often a deep desire burning in my heart, which makes me run into eternity with insanity…

Just another reason, for me to give You praise…

ACKNOWLEDGEMENTS

Many thanks to the disciples whose songs were encouragements to me, especially when my hope of completion was ebbing away. Your words inspired me, even when Charmm, Perl, Python and Jython conspired to kill me despite the many tricks that I set upon them! Scripting kept me basking long in the weird and cold nights, writing command lines that at times didn't make sense to me, but I made sense to them!

Dr. Kevin A. Lobb, I know I might not have been the best student for the project, but even when my interests were docked to computational frustrations, your endless presence and motivation always freed and thrilled me, beyond appreciation. You are a mentor. Dr. Özlem Taştan Bishop, besides granting me the admission and endorsing me for the Henderson Bioinformatics award, I highlight my accolades for the workshops, putting me to participate in several journal clubs, and ensuring that challenges were always there, to keep me going. You were always behind me, ensuring that I toiled hard to complete my project, with a publication motive. I salute Prof. Perry T. Kaye, the quiet pharmaceutical surgeon, always operating in detail behind the scenes! Prof. Lynn Morris and team, from the National Institute of Communicable Diseases (NICD), my writings and smile today would have been inexistent were it not for the infant sequences from you. A standing ovation goes to other postgraduate lecturers: Prof. Philip Machanick, Prof. Nigel Bishop, Prof. Gunther Jaeger, Dr. Adrienne Edkins, Dr. Mike Ludewig, Mr. Jeremy Baxter, and Mr. Gustavo Adolfo. I am obliged to Mr. Alex Mathu, my academic predecessor, for the foundation. To all my referees (mostly Dr. Lillian Waiboçi-Muhia), discipleship family, friends, students, RUBi and Pharmaceutical Chemistry colleagues, thanks for your companies, trips, holistic uplifts and smiles. To the KEMRI/CDC HIVR-Lab folks, I echoed your training here, hereafter to lay my respect. To the "masked" external examiners, I owe you thanks for framing the ultimate tone my voice.

To my family in Kenya, you always believed in me. You never sided with me particularly when I aired-out my academic weaknesses to you (as an excuse to take a break). You always told me to press-on, I am a champion. A bunch of appreciation now that I'm happily done!

Today I look back, and see the purpose that You had for me, GOD. I must say, You're not yet done with me. Regardless of the many delays, let-downs and implausible blows in my life, You always saved me. You know my past and have my future. *Shalom Aleichem*… ♩ ♪♫♬

LIST OF ABBREVIATIONS

3P	3'-processing
3TC	Lamivudine
A	Adenine
ADT	Auto-Dock Tool
AIDS	Acquired Immunodeficiency Syndrome
Ala (A)	Alanine
ANOLEA	Atomic Non-Local Environment Assessment
ANRS	National Agency for AIDS Research
APV	Amprenavir
Arg (R)	Arginine
ARS	AntiRetroScan
ART	Antiretroviral Therapy
Asn (N)	Asparagine
Asp (D)	Aspartic Acid
ATV	Atazanavir
BLOSUM	BLOck of amino acids SUbstitution Matrix
C	Cytosine
CA	Capsid
CASTp	Computed Atlas Surface Topography of proteins
CD	Cluster of Differentiation
cDNA	Complementary Deoxyribonucleic Acid
CHARMM	Chemistry at HARvard Molecular Mechanics

CRF	Circulating Recombinant Forms
cRMSD	Cartesian/Coordinate Root Mean Square Deviation
CRS	Cis-acting Responsive Sequences
Cys (C)	Cysteine
d4T	Stavudine
DC-SIGN	Dendritic Cell-Specific Intercellular adhesion molecule-3-Grabbing Non-integrin
DLV	Delavirdine
DOPE	Discrete Optimized Protein Energy
DRM	Drug Resistance Mutation
DRV	Darunavir
EC	Enzyme Commission
EFV	Efavirenz
env	Envelope
FDA	Food and Drug Administration
FEZ- 1	Fasciculation and Elongation protein ζ1 (Zeta or Zygin I)
FPV/fAPV	Fosamprenavir
G	Guanine
gag	Group Specific Antigens
GDT-TS	Global Distance Test - Total Score
Gln (Q)	Glutamine
Glu (E)	Glutamic Acid
Gly (G)	Glycine
gp	Glycoprotein
HAART	Highly Active Antiretroviral Therapy
HIV	Human Immunodeficiency Virus

His (H)	Histidine
IAS	International AIDS Society
IDV	Indinavir
Ile (I)	Isoleucine
IN	Integrase
INH	Inhibitory/Instability RNA sequences
INI	Integrase Inhibitor
LANL	Los Alamos National Lab-HIV sequence database
Leu (L)	Leucine
LFA	Leukocyte Adhesion Receptor
LGA	Lamarckian Genetic Algorithm
LPV	Lopinavir
LTR	Long Terminal Repeat
Lys (K)	Lysine
MA	Matrix
Met (M)	Methionine
MIP	Macrophage Inflammatory Protein
MRP	Multidrug Resistance-associated Protein
ms	millisecond
MS	Molecular Surface
M-tropic	Macrophage-tropic
NC	NucleoCapsid
NFV	Nelfinavir
NF-Kb	NF kappa B
NIAID	National Institute of Allergy and Infectious Diseases
NICD	National Institute of Communicable Diseases

NNRTI	Nonnucleoside Reverse Transcriptase Inhibitor
NRTI	Nucleoside Reverse Transcriptase Inhibitor
ns	nanosecond
NSI	Non-Syncytia Inducing
PE	Psi Elements
P-gp	P-glycoprotein
Phe (F)	Phenylalanine
PI	Protease Inhibitor
pK_a	Acid Dissociation Constant
pM	Pico Molar
PR	Protease
Pro (P)	Proline
ProSA	Protein Structure Analysis
QMEAN	Quality Model Energy ANalysis
RANTES	Regulated on Activation, Normal T cell Expressed and Secreted
rev	Regulatory factor for HIV expression
RMSD	Root Mean Square Deviation
RNA	Ribonucleic Acid
RRE	rev Responsive Element
RT	Reverse Transcriptase
RTI	Reverse transcriptase inhibitor
RTV	Ritonavir
SDF-1	Stromal cell-Derived Factor-1
Ser (S)	Serine
SI	Syncytia Inducing
SIV	Simian Immunodeficiency Virus

SIVcpz	SIVs from wild chimpanzees
SIVgor	SIVs from wild gorillas
SIVmm	SIV from sooty mangabeys
SLIP	Slippery Site
SA	Solvent Accessible surface
SASA	Solvent Accessible Surface Area
sPI	Single PI
SQV	Saquinavir
Stats SA	Statistics South Africa
STD	Sexually Transmitted Disease
SU	Surface protein
T	Thymine
T-20	Enfuvirtide
T-Coffee	Tree-based Consistency Objective Function for Alignment Evaluation
TAR	Target sequence for viral transactivation
tat	Transactivator for HIV gene expression
Thr (T)	Threonine
TM	Transmembrane
TPV	Tipranavir
Trp (W)	Tryptophan
Try (Y)	Tyrosine
T-tropic	T cell tropic
UNAIDS	Joint United Nations Programme on HIV/AIDS
URF	Unique Recombinant Form
Val (V)	Valine
VGI	Visible Genetics Interpretation program

vif Viral Infectivity Factor

vpr Viral Protein R

WHO World Health Organization

LIST OF FIGURES

LIST OF TABLES

TABLE OF CONTENTS

ABSTRACT

Despite the current interventions to avert contagions and AIDS-related deaths, sub-Saharan Africa is still the region most severely affected by the HIV/AIDS pandemic, where clade C is the dominant circulating HIV-1 strain. The *pol*-encoded HIV-1 protease enzyme has been extensively exploited as a drug target. Protease inhibitors have been engineered within the framework of clade B, the commonest in America, Europe and Australia. Recent studies have attested the existence of sequence and catalytic disparities between clades B and C proteases that could upset drug susceptibilities. Emergence of drug-resistant associated mutations and combinatorial explosions due to recombination thwarts the attempt to stabilize the current highly active antiretroviral therapy (HAART) baseline. The project aimed at identifying the structural and molecular mechanisms hired by mutants to affect the efficacies of both FDA approved and Rhodes University (RU)-synthesized inhibitors, in order to define how current and or future drugs ought to be modified or synthesized with the intent of combating drug resistance. The rationale involved the generation of homology models of the HIV-1 sequences from the South African infants failing treatment with two protease inhibitors: lopinavir and ritonavir (as monitored by alterations in surrogate markers: CD4 cell count decline and viral load upsurge). Consistent with previous studies, we established nine polymorphisms: 12S, 15V, 19I, 36I, 41K, 63P, 69K, 89M, and 93L, linked to subtype C wild-type; some of which are associated with protease treatment in clade B. Even though we predicted two occurrence patterns of M46I, I54V and V82A mutations as V82A→I54V→M46I and I54V→V82A→M46V, other possibilities might exist. Mutations either caused a protracted or contracted active site cleft, which enforced differential drug responses. The *in silico* docking indicated susceptibility discordances between clades B and C in certain polymorphisms and non-polymorphisms. The RU-synthesized ligands displayed varied efficacies that were below those of the FDA approved protease inhibitors. The flaps underwent a wide range of structural motions to accommodate and stabilize the ligands. Computational analyses unravelled the need for these potential drugs to be restructured by (*de novo*) drug engineers to improve their binding fits, affinities, energies and interactions with multiple key protease residues in order to target resilient HIV-1 assemblages. Accumulating evidences on contrasting drug-choice interpretations from the Stanford HIVdb should act as an impetus for the customization of a HIVdb for the sub-Saharan subcontinent.

CHAPTER ONE

1. INTRODUCTION

1.1 Epidemiology and Classification of HIV

According to the UNAIDS 2011 World AIDS Day annual report, 34.0 million people were estimated to be living with the Human Immunodeficiency Virus (HIV) at the closure of 2010. There were 2.67 million new infections, 390,000 new infections in children and 1.76 million AIDS-related deaths that occurred globally in 2010. 2010 Regional statistics attest that the pandemic is high in sub-Sahara Africa with 22.9 million people living with HIV. 1.9 million new HIV infections, 1.2 million AIDS-related deaths and 5.0% adult prevalence (UNAIDS, 2011).

HIV infection is considered pandemic by the World Health Organization (WHO). The WHO Global HIV/AIDS Response, 2011 Progress Report, shows that 5.6 million South Africans were infected as of 2009, and this figure balances the infected population in Asia (WHO, UNAIDS & UNICEF, 2011). This contrasts against the total South African population (between mid 2009 and mid 2010) which estimated 49.32 - 49.99 million (Statistics South Africa, 2009, 2011).

HIV, a lentivirus in the Retroviridae family (Nielsen et al., 2005), is the causative agent for Acquired Immunodeficiency Syndrome (AIDS) (Douek et al., 2009; Weiss, 1993), a condition that destroys the immune system (Deeb & Jawabreh, 2012), thereby leading to manifestation of opportunistic infections (OIs) (Araya et al., 2011), coronary disease, metabolic anomalies and cancer (Boudová et al., 2012). Two HIV strains exist: HIV-1 and HIV-2 (Gilbert et al., 2003). HIV-2 exhibits 40-60% identity with HIV-1 and is thought to be less virulent and transmissible than HIV-1, hence not pandemic (Reeves & Doms, 2002). It displays a more gradual and milder infectivity rise during its immunodeficiency development. Compared with HIV-1, the duration of this increased infectivity is shorter (Gilbert et al., 2003).

HIV-1 is transmitted through sexual, percutaneous and perinatal routes. Transmission documentations state that 80% of adults get infected after (genital) mucosal sites have been

exposed to the virus, hence justifying that AIDS is primarily a sexually transmitted disease (STD) (Cohen et al., 2011; Hladik & McElrath, 2008).

HIV-1 is further classified into major group (Group M) and at least two minor groups. Group M exhibits considerable genetic diversity and is the most predominant in the global endemic (Zhu et al., 1998). Currently, it has 9 pure subtypes, known as clades (A, B, C, D, F, G, H, J and K) (Kantor & Katzenstein, 2004), at least 43 Circulating Recombinant Forms (CRF) and Unique Recombinant Forms (URF) as a result of recombination (Bulla et al., 2010; Quesnel-Vallières et al., 2011; Taylor et al., 2008). Two to three recombination events per genome per cycle occur in HIV-1 (Jetzt et al., 2000). These clades differ by up to 15% and 30% of amino acid in the *gag* and *env* genes respectively (Korber et al., 2001), therefore there is limited subtype cross-reactivity in terms of antibody titers (Mthunzi and Meyer, 2004).

Clades A and F are further split into different sub-lineages namely A1, A2, A3 and A4, and F1 and F2 respectively (Abecasis et al., 2007; Gao et al., 2001; Taylor et al., 2008; Triques et al., 1999). Group N ("non-M, non-O" or "Newer"), has only been reported in Cameroon since 1998 (Yamaguchi et al., 2006). Geographical distribution of HIV-1, reports that Group O ("Outlier") is also common in Cameroon (Peeters et al., 1997). A new putative group, Group P ("pending the identification of further human cases"), has been reported at least twice in Cameroon (Vallari et al., 2011). Its lineage has been more distinctively rooted to the Simian Immunodeficiency Virus that attacks wild gorilla, (SIVgor), than to SIVs from wild chimpanzees, (SIVcpz) (Van Heuverswyn et al., 2006; Plantier et al., 2009; Takehisa et al., 2009).

Vertical and horizontal transmission evidences have identified SIVmm from sooty mangabeys as the progenitor for HIV-2 due to its potentiality to cross the species genetic barrier (Hahn et al., 2000; Marx et. al., 2001; Santiago et al., 2005). West Africa has the highest prevalence of HIV-2 (Reeves & Doms, 2002) and as of 2010, eight clades of HIV-2 (A, B, C, D, E, F, G and H) and one CRF had been tabulated (Ibe et al., 2010).

Phylogenetic classification of HIV strains has effectuated the tracking of the diversity of the circulating strains, which account for different degrees of infections worldwide. For example, HIV-1 clade B viruses majorly occur in Europe, America and Australia, whilst clade C dominates in Southern Africa, India and Nepal (Kandathil et al., 2005; McCutchan, 2000).

1.2 HIV Structure and Landmarks

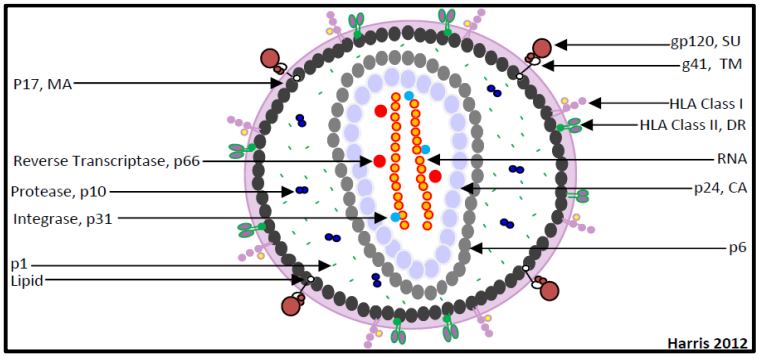


Figure 1-1: Schematic diagram of a mature HIV-1 structure, showing its key components (adapted and revised from http://www.microbiologybytes.com/virology).

HIV has a diameter of 120 nm (Kuznetsov et al., 2003; Song et al., 2009) (Figure 1-1). Its genome consists of two identical single-stranded sense RNA strands bound to NucleoCapsid (NC), p7, and enzymes. The genome is hemmed in an icosahedral capsid (CA), p24, that is enclosed in a Matrix (MA), p17 (Höglund et al., 2002). Anchored onto the surface are envelope proteins embedded in a phospholipid bilayer taken from host cells during budding. They appear as spikes formed by trimers each of noncovalently linked gp120 (SU) and gp41 (TM) and are vital for virus attachment and fusion during infectivity (Reeves & Doms, 2002).

The virus comprises at least nine genes (9.719 kb) (Nielsen et al., 2005; Song et al., 2009) namely *gag* (group specific antigens), *pol*, *env* (envelope), *tat* (transactivator for HIV gene expression), *rev* (regulatory factor for HIV expression), *vif* (viral infectivity factor), *vpr* (viral protein R, in HIV-1)/*vp*x (a duplicated *vpr* in HIV-2), *vpu* (viral protein U in HIV-1) and *nef* (negative factor) (Nielsen et al., 2005), and occasionally *tev* (a hybrid cardinally comprising of *tat* and *rev* and partly *env*) (Benko et al., 1990).

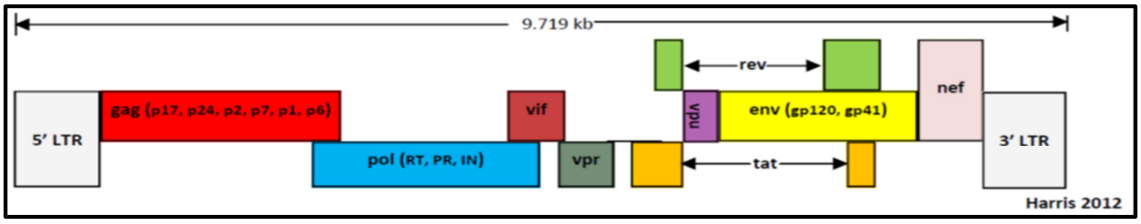


Figure 1-2: Landmarks of the HIV genome. The rectangles indicate the open reading frames (ORF) of its gene marks (revised from http://www.hiv.lanl.gov/content/sequence/HIV/MAP/landmark.html).

The Los Alamos National Lab-HIV sequence database identifies HIV as having seven genomic structural elements (Figure 1-2) namely LTR (long terminal repeat), TAR (target sequence for viral transactivation), RRE (*rev* responsive element), PE (Psi elements), SLIP (slippery site),

CRS (cis-acting responsive sequences) and INH (inhibitory/instability RNA sequences). The *gag*, *pol* and *env* code for structural proteins including the viral enzymes; *tat* and *rev* code for regulatory proteins for (post)transcriptional steps while *vif*, *vpr*/*vpx*, *vpu* and *nef* code for auxiliary proteins (Le Rouzic & Benichou 2005). The *pol* gene codes for reverse transcriptase (RT, p66: RT, p51 and RNase H, p15), protease (PR, p10) and integrase (IN, p31). These enzymes are initially synthesized as *gag-pol* precursor as a result of ribosome frameshifting near the 3' *gag* end (http://www.hiv.lanl.gov/content/sequence/HIV/MAP/landmark.html).

1.3 HIV Life Cycle and Treatment

HIV attacks macrophages, monocytes, myeloid dendritic cells, microglial cells, $CD4^+$ T cells (Regoes & Bonhoeffer, 2005) and at times spermatozoa due to its heparan sulphate, besides having CCR3 and CCR5 (Ceballos et al., 2009; Habasque et al., 2002). Neurons have the FEZ-1 (fasciculation and elongation protein ζ1) molecule that foils HIV infection. HIV exhibits viral tropism; there exists macrophage strains (M-tropic or non-syncytia inducing (NSI) or R5 strains) that attack via β-chemokine receptor CCR5 whose ligands are macrophage inflammatory protein (MIP-1α and MIP-1β) and RANTES (regulated on activation, normal T cell expressed and secreted) (Dragic et al., 1996; He et al., 1997; Lobritz et al., 2010; Wu et al., 1997); T-tropic (syncytia inducing (SI) or X4) infect $CD4^+$ T cells via the α-chemokine receptor whose ligand is SDF-1 (stromal cell-derived factor-1) (Maréchal et al., 1999).

During late AIDS stage, a co-receptor switch of R5→X4 usually occurs. Dual tropism is for viral adaptation and is due to a transitional switch of HIV-1 to utilize both receptors for infectivity (Regoes & Bonhoeffer, 2005). Such strains are referred as R5X4. R5 can also switch to R5X4 (Jones et al., 2010). Dendritic cells get infected either via the CD4-CCR5 route or mannose-specific C-type lectin receptors of which CD209 or DC-SIGN (dendritic cell-specific intercellular adhesion molecule-3-grabbing non-integrin) is an example (Cunningham et al., 2007). Intestinal dendritic cells are also targeted by HIV; once infected, they infect the T cells via the intestinal mucosa (Shen et al., 2010). Individuals with CCR5-Δ32 mutation are completely or nearly resistant to R5 depending on the form of mutational zygosity (Arts, 2010).

The concatenation biology of the HIV infectious cycle according to 2012 National Institute of Allergy and Infectious Diseases (NIAID), is as follows (Figure 1-3): 1) Attachment and fusion

of HIV to host receptors, 2) Discharge of RNA, enzymes and viral proteins into host cell, 3) Reverse transcription to form proviral DNA, 4) Importation of proviral DNA imported to nucleus, followed by its integration into host DNA, 5) Synthesized viral RNA used as genomic RNA and to translate viral proteins, 6) Migration of synthesized viral RNA and proteins to cell membrane to form immature HIV, and 7) Maturation and budding of the virions after proteolytic processing. This has also been reported in other studies (Nielsen et al., 2005; Buckheit et al., 2011; Abbas & Herbein, 2012).

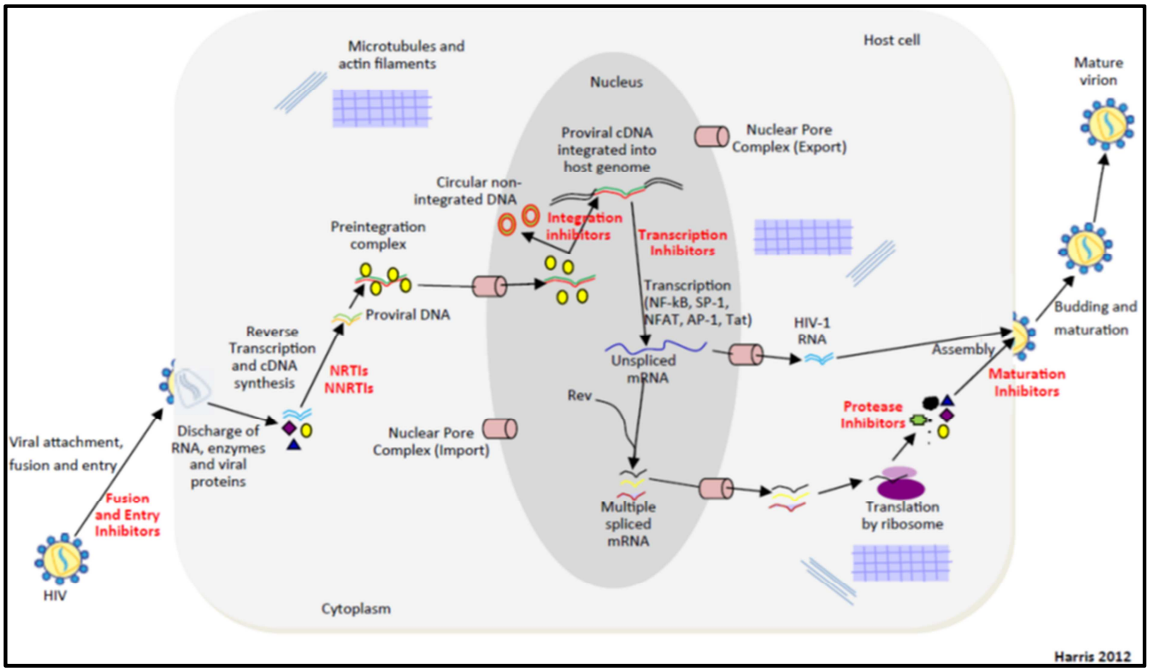


Figure 1-3: Diagrammatic representation of the HIV infectious cycle, indicating the drugs used to curtail infection (accessed and revised from Abbas & Herbein, 2012).

Cellular HIV infection commences with high affinity adsorption where gp120 recognizes and binds to $\alpha_4\beta_7$ thereby activating a central integrin called Leukocyte Adhesion Receptor (LFA-1) that institutes virological synapses (Hioe et al., 2011). The gp160 is made up of five variable domains (V1-V5) and five conserved domains (C1-C5). Its chemokine binding domains (neutralizing domains) in gp120 are exposed for attachment reinforcement once V3 undergoes conformation (Li & Pauza, 2011). Conformational change in gp120 also intensifies its binding to heparan sulphate (Vivès et al., 2005). It binds either to CCR5 and or CXCR4 co-receptors (Dragic et al., 1996; Pasquato et al., 2007; Song et al., 2009). The N-terminal of gp41 then permeates through the plasmalemma. Interaction of gp41 heptad repeat sequences, HR1 and HR2, collapses the extracellular potion of gp41 into a hairpin

loop consisting of coiled-coil helices that augments viral fusion, entry and release of the CA inclusions (Pomerantz & Horn, 2003), where sense cDNA is reversed transcribed from the viral RNA via an antisense DNA using RT which has three activities: reverse transcription, DNA-dependent DNA polymerase and ribonuclease activities. The RNA is destroyed by RNase H, a ribonuclease. Through microtubule- and dynein-based transport (Abbas & Herbein, 2012), the preintegration complex is imported into the nucleus. Integrase then executes ligation of the cDNA in a two-step reaction; 1) endonucleolytic 3'-processing (3P), and 2) strand transfer (ST) reaction (Wang et al., 2005). Virions are produced when NF-κB (NF kappa B), NFAT, AP-1 and SP1 are unregulated so that RNA polymerase II can bind to the TATA box to initiate transcription. The last phases involve exportation of unspliced mRNA to cytoplasm, transport of glycoproteins to the cell surface via endoplasmic reticulum and Golgi apparatus, viral assembly and release (Abbas & Herbein, 2012).

Due to the absence of an HIV vaccine, antiretroviral therapy (ART) is still in use. ART utilizes inhibiting strategies that target virus attachment, fusion, reverse transcription, integration and proteolysis (Nielsen et al., 2005; Pomerantz & Horn, 2003). HIV drugs have been designed to target specific points in the HIV life cycle (Arts & Hazuda, 2012), even though some can be combined to form multi-class combination products such as atripla (efavirenz, emtricitabine and tenofovir disoproxil fumarate). As of 2011, the US Department of Health and Human Services reported the four main classes as:

1) Reverse transcriptase inhibitors, (RTIs), (Nucleoside Reverse Transcriptase Inhibitors (NRTIs), e.g., lamivudine (3TC), stavudine (d4T), Nonnucleoside Reverse Transcriptase Inhibitors (NNRTIs), e.g., efavirenz (EFV), delavirdine (DLV)). NRTIs act as chain terminators once incorporated in the growing transcript while NNRTIs bind to and extinguish RT function (Arts & Hazuda, 2012).
2) Protease inhibitors (PIs), e.g., amprenavir (APV), tipranavir (TPV), indinavir (IDV), nelfinavir (NFV) which inhibit proteolytic processing the last stage of the cycle (Abbas & Herbein 2012; Arts & Hazuda, 2012).
3) Integrase strand transfer inhibitors, e.g., raltegravir, inhibit strand transfer reaction (Abbas & Herbein, 2012; Arts & Hazuda, 2012).
4) Fusion inhibitors, e.g., enfuvirtide (T-20) bind to gp41 to deny entry of HIV into host cells (Abbas & Herbein, 2012; Arts & Hazuda, 2012).

There are entry inhibitors, e.g., maraviroc, aplaviroc which are CCR5 co-receptor antagonists; and TNX-355 (Arts & Hazuda, 2012), KRH-1636 and AMD3100 which are CXCR4 antagonists (Briz et al., 2006; Lobritz et al., 2010). Generally, fusion inhibitors fall under entry inhibitors. In the pipeline, there are transcription inhibitors, e.g., RNAi, L50, etc., and HIV maturation inhibitors target the last phase of HIV-1 *Gag* processing (Abbas & Herbein, 2012), hence disrupting viral assembly and production, e.g., vivecon (Arts & Hazuda, 2012).

Combination therapy known as highly active ART (HAART) (Martin et al., 2005) that was devised in 1996 has transformed HIV/AIDS management (Chandwani & Shuter, 2008) and is nowadays employed to perturb HIV-1 pathogenesis equilibrium, hence reducing and suppressing viremia and delaying onset period of AIDS (Buckheit et al., 2011; Abbas & Herbein, 2012). For example two NRTIs may be combined with either one NNRTI or PI (Ortega et al., 2009). After much debate of clinical progression and drug-associated toxicities, the current therapeutic guidelines recommend initiation into the HAART to be when the baseline CD4 cell counts is <350 cells/µl. Before, initiation timing was when the count was <200 cells/µl, but due to rapid clinical progression, this was revised (Caroline & Andrew, 2009). The debate still continues. Beside the aforementioned baseline, upsurges in viral load are also considered (Arts & Hazuda, 2012).

In South Africa, first-line therapy in drug-naïve adults constitutes lamivudine, stavudine and efavirenz. In pregnant women, nevirapine replaces efavirenz. Children less than 3 years old capitalize on PIs as initial therapy. Second-line adult therapy commends zidovudine, didanosine and LPV/r (Bessong, 2008).

1.4 HIV Sequence Variation

The two most striking characteristics of HIV-1 include its high mutation turnover (ranging between 5×10^{-6} and 9×10^{-5} mutations per nucleotide per cycle of virus replication) (Smith et al., 2005), and recombination rate (42.4% per replication cycle, with markers 1 kb apart (Rhodes et al., 2003), leading to HIV-1 evolution (Abecasis et al., 2007). Genome plasticity of HIV-1 leads to its sustained biodiversification (Malim & Emerman, 2001).

Factors accountable for HIV genetic diversity include (i) the lack of proof reading capability by the HIV RT (Purohit et al., 2008) owing to the absence of 3'→5' exonuclease activity; (ii) the *in vivo* rate of viral turnover/replication; (iii) the accrual of proviral variants during the

contagion period; and (iv) recombination as a result of heterogeneity of infecting population or dual infection (Coffin, 1995; Shafer et al., 2000; Malim & Emerman, 2001; Nukoolkarn et al., 2004; Barbour & Grant, 2005).

1.5 HIV-1 Protease

1.5.1 HIV-1 Protease Structure

HIV-1 protease (EC 3.4.23.16 from 2012 IUBMB), an aspartyl protease (retropepsin, from sequence homology and inhibition by pepstatin (Brik and Wong, 2003)), is a homodimer composed of two non-covalently linked structurally identical monomers each having 99 amino acids (Kear et al., 2011; Shafer et al., 2001) (Figure 1-4). Crystallographic studies have revealed that its active site displays a perfect two-fold symmetry in the free form (Brik & Wong, 2003) and that this site is covered by two symmetry-related β hairpins, termed as flap connected by glycine rich loops (Baldwin et al., 1995; Brik & Wong, 2003; Hornak et al., 2006a). It resembles other aspartyl proteases due to its conserved triad, Asp-Thr-Gly (Asp25, Thr26 and Gly27) (Brik and Wong, 2003; Shafer et al., 2001).

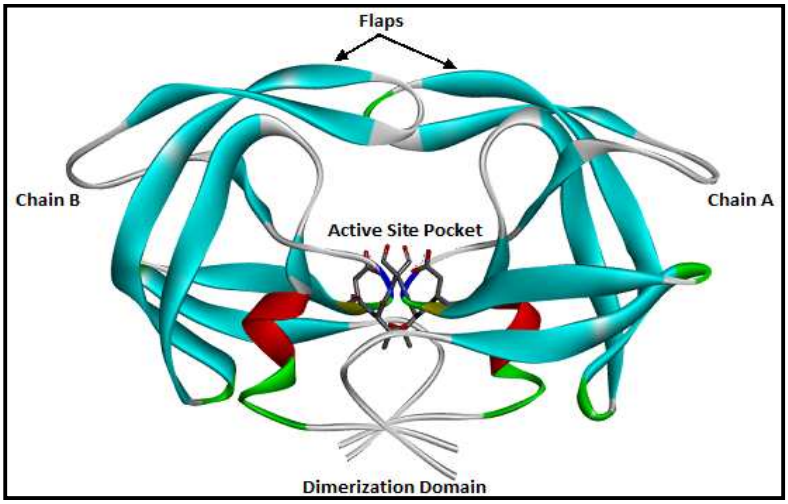


Figure 1-4: 3D structure of HIV-1 in closed conformation (PDB ID: 1HXB) as visualized by Pymol. Conserved catalytic triad indicated in stick representations.

Figure 1-5 shows that it is made up of two flanking cleavage sites (p6*PR: -5 to 5, **GREEN**) and PR/RT: 95 to 105, **GREEN**) and three functional regions, which include the active site (21 to 32, **RED**) lies between the identical subunits. The two Asp25 residues in the active region (one from each chain) act as the catalytic residues; if interchanged with Asn, Thr or Ala, the enzyme becomes inactive (Brik & Wong, 2003). The dimer has substrate binding clefts (78 to

88, BROWN), and two molecular "flaps" (37 to 61, BLUE) (Yu et al., 2011) that endow the backbone with a flexibility of up to 7 Å upon substrate binding (Miller et al., 1989; Nukoolkarn et al., 2004).

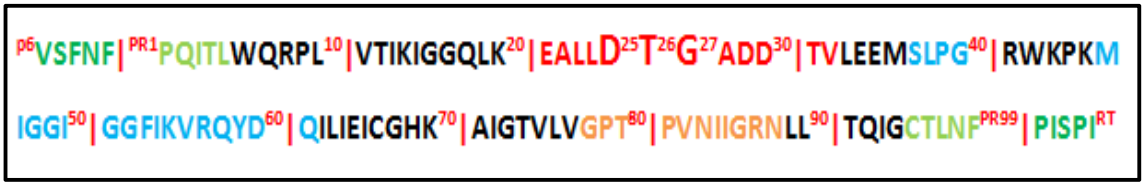


Figure 1-5: HXB2 isolate (HIV-1 wild-type) (the amino acid sequences obtained from HIV bioinformatics in Africa - http://bioafrica.mrc.ac.za/proteomics/HIV1-HXB2-PR.fasta). Green indicates the cleavage sites, red the active site, blue the flap regions while brown shows the substrate binding region. The superscripted numbers indicate the amino acid positions.

Hydrogen-bonding networks related to those in eukaryotic enzymes (Wlodawer & Vondrasek, 1998) occur between the active site residues and those in close proximity. This complex grid is known as the "fireman's grip", and it tasks to hold the loops of the active site together, rendering it rigid. This scaffold of hydrogen bonds stabilizes the dimer and arises when each Thr26 accepts a hydrogen bond from the opposing chain (Alcaro et al., 2009), then transfers it to the carbonyl oxygen of Leu24 (Ingr et al., 2003; Wlodawer & Vondrasek, 1998). Each Asp25 interacts with the amine group of the opposing Gly27 (Das et al., 2006).

The flaps regulate access to the active sites and these domains exist in three states: open, closed and semi-open structures (Freedberg et al., 2002); structures ranging from open and closed have been revealed in free protease. A ligand will best dock to the receptor site when the flaps are opened. Flaps closure then ensues (Chang et al., 2007). In the closed conformation, the flaps are in close proximity to the catalytic triad whereas the reverse is true in semi-open structures (Hornak et al., 2006a).

1.5.2 HIV-1 Protease Function

HIV-1 protease cleaves *gag* and *gag-pol* polypeptides into functional translates so that mature and infectious viral particles are spawned (Martin et al., 2005; Ode et al., 2007; Petrokova, 2006; Pomerantz & Horn, 2003). HIV-1 viruses encompassing inactive protease are debilitated to hydrolyze the polypeptide precursors during the maturation process and therefore cannot replicate and infect new cells (Wlodawer et al., 1989).

HIV-1 protease employs its substrate binding cleft for recognition and cleavage of at least nine different sequences of the polyprotein to yield the MA, CA, NC, and p6 proteins from

the *gag* translate, and the PR, RT and IN enzymes from *gag-pol* translate (Shafer et al., 2000, 2001). This is due to its high sequence selectivity and catalytic proficiency (Brik and Wong, 2003), besides its flap flexibility mechanism; where the unliganded protease populates the semi-open conformation while both closed and fully open structures comprises minor components of the overall ensemble (Freedberg et al., 2002; Hornak et al., 2006).

1.5.3 HIV-1 Protease Mechanism of Action

Synthesized polypeptide must bind to the active site for post-translational cleavage so that mature proteins can be produced for new virus assemblage (Panther & Libman, 2005). The binding pathway occurs when both the substrate and protease diffuse together and undergo orbital steering in terms of proximity and orientation in order to adopt specific conformation and contact. Figure 1-6 shows the nomenclature of the substrate and binding subsites. Binding occurs in two phases: the first phase is due to non-specific and long-range electrostatic forces, while the second one may be due to specific short-range interactions. A small-sized inhibitor such as cyclic urea can penetrate through the binding cleft even when the flaps are not completely open, but if the flaps are nearly closed, then the inhibitor must undergo conformational change to assist in orbital steering (Chang et al., 2007).

Mechanistically there are two protease classes: 1) those that require water molecule to hydrolyze the scissile bond (zinc metalloproteinases that use a zinc cation as the water activator, and aspartate proteases that use two aspartyl β-carboxy groups at the active site as water activators), and 2) those that use the nucleophilic atom of an hydroxyl or thiol moiety for hydrolysis. Various mechanisms of HIV-1 protease have been proposed based on various methods e.g., kinetics and structural studies; but they share mechanistic features with the concerted mechanism, a one-step process that was postulated by Jaskólski et al. (Figure 1-7). Protonation states of the two Asp25 are different; pK_a values 3.1 and 5.2. The acid-base chemical catalysis is the accepted mechanism. The monoprotonated state of the two catalytic Asp is a catalysis requisite (Brik & Wong, 2003).

Hydrolysis entails the general base (the Asp25 R group), which is ionized (COO^-), deprotonating the already nucleophilic water. The general acid, Asp25' R group, is protonated (COOH) and protonates the carbonyl oxygen atom of the scissile peptide bond (Yu et al., 2011). Suguna et al. indicated that there is an oxyanion tetrahedral intermediate

formed when the activated water attacks the carbonyl oxygen atom. It is this intermediate that is then protonated by Asp25' and hydrolyzed (Brik & Wong, 2003).

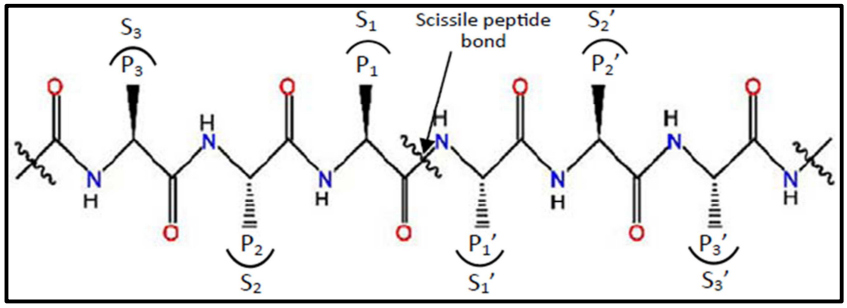


Figure 1-6: Standard nomenclature of peptide substrates (P_1…P_n, P_1'… P_n') and HIV-1 protease binding subsites (S_1…S_n, S_1'…S_n'). S_1 subsites are very hydrophobic. S_2 and S_3 are mostly hydrophobic except for Asp29 and Asp30 in S_2 (Figure accessed with permission from Brik and Wong, 2003).

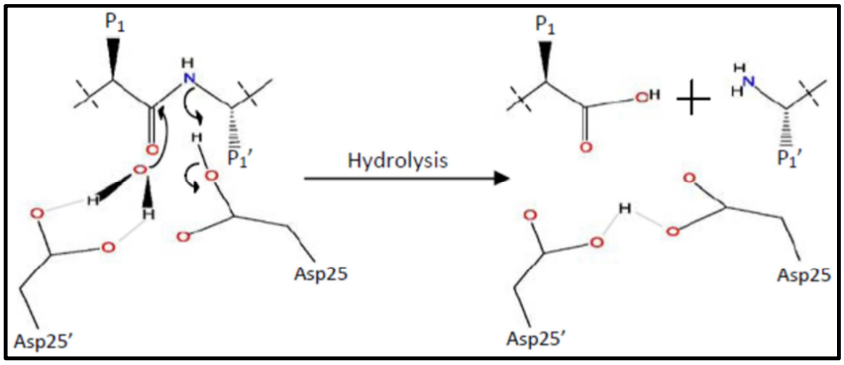


Figure 1-7: General acid-general base catalytic mechanism of HIV-1 protease. A concerted action involving Asp25 and Asp25' ensures that the scissile bond of the peptide is cleaved by the activated water molecule (Figure accessed with permission from Brik & Wong, 2003).

1.5.4 HIV-1 Protease as a Drug Target

Most of the FDA approved PIs exhibit poor pharmacokinetics in terms of 1) low aqueous solubility, 2) poor membrane permeability, 3) high binding plasma protein, P-glycoprotein (Pokorná et al., 2009), and multidrug resistance-associated protein (MRP1 and MRP2) efflux channel (Zeldin & Petruschke, 2004), and 4) insufficient metabolic stability (Wu et al., 2008).

Single PI-based regimens were used to significantly reduce AIDS-related deaths, but due to their poor pharmacokinetics, boosted PI-regimens that may include IDV, LPV, APV or SQV, are utilizing RTV nowadays to improve efficacy and tolerability through dosing and bioavailability (Kaplan & Hicks, 2005); thereby efficaciously suppressing viremia in first-line and salvage therapies in adults, adolescents and children (Alcaro et al., 2009; Chandwani & Shuter, 2008; Kaplan & Hicks, 2005; Zeldin & Petruschke, 2004). Kaletra (Lopinavir/Ritonavir, LPV/RTV or LPV/r) is also linked to high genetic barrier to resistance (Kaplan & Hicks, 2005).

Major side effects associated with Kaletra include gastrointestinal upsets, serum hyperlipidemia (hypercholesterolemia and hypertriglyceridemia) and lipodystrophy syndrome (Kaplan & Hicks, 2005; Pokorná et al., 2009).

HIV-1 protease inhibitors are divided into two major groups:

1) First generation protease inhibitors, e.g., saquinavir (SQV)/invirase, ritonavir (RTV)/norvir, indinavir (IDV)/crixivan, nelfinavir (NFV)/viracept, amprenavir (APV)/agenerase, and fosamprenavir/ (FPV, fAPV)/lexiva (Pokorná et al., 2009).
2) Second generation protease inhibitors, which were created to target strains resistant to the first generation PIs, and improve adherence due to minimized side effects and improved dosing (once-daily prescription), e.g., lopinavir (LPV)/Kaletra/aluvia (Chandwani & Shuter, 2008), atazanavir(ATV)/reyataz, and tipranavir (TPV)/aptivus (Pokorná et al., 2009).

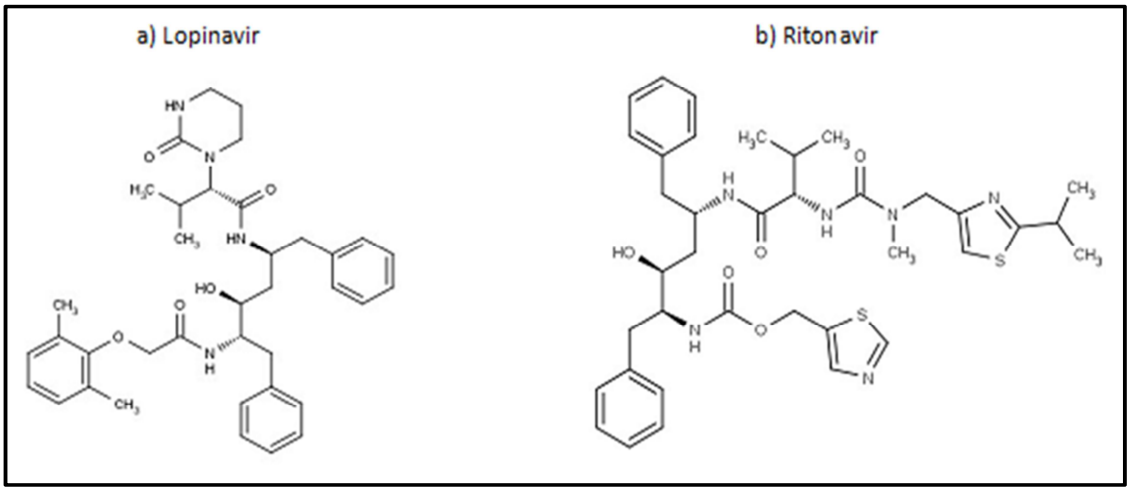


Figure 1-8: Chemical structures of two HIV-1 PIs; Lopinavir (LPV) and Ritonavir (RTV), adapted from the drug bank (http://www.drugbank.ca/).

Two important PIs that are currently in use in Southern Africa are RTV (Zeldin & Petruschke 2004) and LPV (Sham et al., 1998) (Figure 1-8). Kaletra is cited as the consensus first-line PI in the current ART guidelines (Kaplan & Hicks, 2005; Pokorná et al., 2009). It is more potent than NNTRIs and is associated with lower viral loads. First-line therapy in children <3 years old maximized two NTRIs with one PI (Zyl et al., 2011). Its antiviral activity is chiefly due to LPV, the most widely used PI in drug-naïve patients. The inclusion of low pharmacokinetic levels of RTV boosts LPV levels in the blood (Pokorná et al., 2009). RTV enhances bioavailability of other PIs by inhibiting cytochrome P-450 CYP3A4 enzyme thereby reducing

their catabolism; altering the area under the curve (AUC), maximum concentration (Cmax), minimum concentration (Cmin) and half-life ($t^1/_2$). It also inhibits P-glycoprotein and MRP channels thus permitting PIs to transverse cellular boundaries (Zeldin & Petruschke, 2004).

RTV full or single PI (RTV sPI) dose which were initially used in infants at most six months old or in co-treatment of tuberculosis using rifampicin (a strong CYP3A and P-glycoprotein inducer (Frohoff et al., 2011)) or in children awaiting for formulation of therapeutic guidelines prior to 2007, have been substituted with Kaletra, which is the common PI in the HAART in South Africa; and it is the recommended regimen in HIV-infected subjects less than three years. WHO recommends that infants failing therapy-containing nevirapine should be switched to therapy containing PI (Frohoff et al., 2011). Most PIs are active against HIV-2 with the exception of LPV (Pokorná et al., 2009).

1.5.5 Drug Resistance Mutations in HIV-1 Protease

Since the approval of PIs, the global therapeutic response snapshots have been remarkable as evinced by noteworthy decline in deaths (Pokorná et al., 2009). Despite intra-host variation, HAART remains effective because selection for multiple drug-associated mutations is difficult unless it emerges sequentially (Korber et al., 2001). Drug resistance and cross-resistance are as a result of the dynamics in genetic evolution (Barbour & Grant, 2005; Brik & Wong, 2003; Kantor & Katzenstein, 2004). Mutations are polymorphisms that can either occur naturally or caused by drug-selective pressures. There can either be major or minor mutations (Ohtaka & Freire, 2005), even in protease (Tang et al, 2012). Both major and accessory mutations occur in HIV-1 protease (Shafer et al., 2001), with major mutations being conservative (Ohtaka & Freire, 2005) and emerge first due to the presence of a particular drug and reduce susceptibility to that drug. Minor mutations are selected later after occurrences of major mutations and possess infinitesimal effects on the virus phenotype and are enforcing viral fitness, i.e. refining the replicative capability (Pokorná et al., 2009). They define the onset of resistance (Ohtaka & Freire, 2005). Both host selection pressure (adaptive immunity) and viral factors (drug selection pressures) have impact on the viral fitness (Barbour & Grant, 2005; Nicastri et al., 2003). Mutations alter viral fitness to different extents and only strains with high-level resistance and functional protease will be populated (Nicastri et al., 2003). For instance I47A assigns resistance to both HIV-1 and HIV-

2 to LPV, but at the expense of its fitness, therefore this mutation is quite uncommon (Pokorná et al., 2009). As viral fitness continues to set in, disease progression increases (Arnott et al., 2010). Type and location of mutations are constrained to maintain viral fitness (Ohtaka & Freire, 2005).

The International AIDS Society (IAS) - USA's 2011 update of the HIV-1 drug resistance mutations lists V32I, I47V/A, L76V and V82A/F/T/S as the major mutations associated with LPV/RTV resistance; minor mutations include L10F/I/R/V, K20M/R, L24I, L33F, M46IL, I50V, F53L, I54V/L/A/M/T/S, L63P, A71V/T, G73S, I84V and L90M (Johnson et al., 2011). Other PI mutations are retrievable from the same source. High-level resistance to PIs is slower since it requires the overall participation of other mutations, unlike in some NRTIs and all NNRTIs where a single mutation is enough to cause a high-level resistance in a predetermined fashion (Hirsch et al., 2003). Resistance to Kaletra is due to cumulative increment of nine to eleven positional mutations ("LPV mutation score") (Pokorná et al., 2009). I47A (and possibly 147I) and V32I confer high-level resistance, and inclusion of L76V to 3 PI resistance-associated mutations significantly increases resistance to LPV/r. In HIV-1 primary transmission networks, polymorphisms conferring therapeutic failure and are known to exist in the wild-type are usually ignored (Johnson et al., 2011). Frequency of cross-resistance of LPV/r and other PIs is low (Chandwani & Shuter, 2008). Existences of L24I, I50L/V, F53Y/L/W, I54L, and L76V have been linked with improved virologic response to TPV. V82A and I84V confer resistance to RTV (Pokorná et al., 2009).

In the HIV-1 protease alone, over 87 mutations in at least 49 codon positions have been documented and at times linked to multiclass drug-resistance (MDR) (Ohtaka & Freire, 2005). Even though the development of primary resistance is hard due to PIs, greater than 20 substitution signatures have been found to confer resistance (Arts & Hazuda, 2012). PI resistance has been associated with virologic failure during HAART. Non-synonymous mutations affect substrate binding architecture, binding affinity and replicative capacity. To compensate for the altered dynamics and retain optimal viral functionality, compensatory mutations occur in the protease itself or the C terminal of the *Gag* polyprotein; NC/p1 and p1/p6 cleavage sites (van Maarseveen et al., 2012). Polymorphisms and drug associated mutations in HIV-1 protease are always compensated for in the protease cleavage sites (Nukoolkarn et al., 2004).

Structurally, mutations can either be active site or non-active site mutations. Most major mutations are the former and often distort the binding pocket, whereas the latter alter binding affinity and have compensatory roles (Ohtaka & Freire, 2005). Compensatory mutations affect the enzyme activity through conformational flexibility-assisted molecular mechanisms which maintain the electrostatic characteristics of HIV-1 protease (Piana et al., 2002). M46I variant protease is the commonest compensatory mutation (Wideburg et al., 1994) and unlike the wild-type protease, it stabilizes the closed flaps conformation (Collins et al., 1995). It exhibits minimal alteration in the binding affinity (Pazhanisamy et al., 1996), thereby suggesting the possibilities of other dynamics participating in this mutation. The high level of flap mobility permits attainability of an immense number of conformations on a nanosecond (ns) to millisecond (ms) timescale. This explains the relevance of flap mutations in binding kinetics (Piana et al., 2002).

The V82F/I84V mutation leads to loss of binding affinity for most PIs (Perryman & Lin 2004). LPV/r is an efficient drug since no clinical failure against it has been accumulated against it over long-term usage (Chandwani & Shuter, 2008). Variations in the HIV-1 protease active site selectively perturbs the binding energy of the inhibitor to a greater extent than of that of the substrate (Pazhanisamy et al., 1996). In drug-resistant HIV-1 variants, the S_3 binding region is condensed, thus the hydrophobic binding site is impaired from interacting with P_3/P_3' and P_1/P_1' groups of the inhibitors. Inhibitors with small or lacking P_3 group e.g., TL-3, have been shown to be effective against both wild-type and mutant strains and are associated with a significant delay in the emergence of resistance (Brik & Wong, 2003).

M36I is a common polymorphism in non-clade B that reduces the binding site cleft through positional shifts, i.e., inward displacements of Leu33/Leu33' and Val77/Val77' followed by active site conformational changes in Thr31/Thr31' and Pro79/Pro79'. Positional shifts are conformational changes measured as either contractions (negative) or elongation (positive) distances of residues (in mutants) from the binding cavity center. M36I is associated with slight increase in affinity for NFV. M36I increases the emergence of the N88S mutation that confers resistance to NFV. V82F/I84V lowers the binding affinity and dissociation kinetics of the currently available PIs by distorting the equilibrium between closed and semi-open flap conformations, thus semi-open conformation is populated (Perryman & Lin, 2004).

D30N is a rare non-clade B polymorphism because it lowers the rate of replication, as also reported in L89M polymorphism. NFV is still potent in D30N non-subtype B mutants. In clade B, D30N alone confers NFV resistance by inhibiting hydrogen bonding between of NFV with N30 (Ode et al., 2007). D30N mostly coexists with N88D. Selection of L90M can also confer NFV resistance (Pokorná et al., 2009). Again, M36I/D30N has been shown to impose bonding between N30 and NFV, a network that is not apparent in D30N mutants. Apart from diminished binding affinity, M36I/D30N confers NFV resistance due to a two-fold outward distortion of Asp29, greater than what is seen from either M36I alone or D30N alone. This distortion can be relieved by other polymorphisms with the exception of M36I. Compared to M36I, M36V reduces the volume of the binding pocket to a lesser extent (Ode et al., 2007).

X-ray crystallography has revealed that LPV in Kaletra inhibits mutants selected for by RTV, by avoiding the hydrophobic interaction between RTV and the isopropyl side chain of valine at position 82 (that switches to alanine, threonine or phenylalanine) in HIV-1 protease (Kaplan & Hicks, 2005; Pokorná et al., 2009; Sham et al., 1998).

Emergence of drug-resistant associated mutations and combinatorial explosions due to recombination have antithetical effects on the active site pocket, dynamics and *Gag* polyprotein cleavage sites and have consequently thwarted the attempt to stabilize the current treatment baseline. In spite all these, viral fitness is still maintained (Ali et al., 2010).

1.5.6 HIV-1 Protease in Drug Design and Development

The potent functional capability of the HIV-1 protease is being exploited in substrate-based inhibitor design via substitution of scissile bond for a non-cleavable isostere (Petrokova, 2006). The Lock and Key hypothesis with inhibitor conformational constraints (Ohtaka & Freire, 2005), permitted structure-based drug design of the FDA approved PIs (Ali et al., 2010). Today, ten FDA approved HIV-1 PIs exist (Alcaro et al., 2009; Arts & Hazuda, 2012). RTV and LPV/r for example, were approved on 1st March 1996 and 15th September 2000, respectively (Pomerantz & Horn, 2003). Atomistic simulations of ligand binding processes is useful for drug discovery since it identifies optimal association pathways and design ligands possessing good binding kinetics (Chang et al., 2007), thus the structure-based drug design concept can be used to demystify drug-resistance mechanisms (Ali et al., 2010).

Drug development continues due to emergence of numerous mutants and rapid replication and viral transcription errors (Wu et al., 2008). HIV-1 PIs, first launched in triple combination therapy in 1995 through advanced drug discovery processes (Pomerantz & Horn, 2003), have the effect of rigidifying HIV-1 protease flaps (Heal et al., 2011) which are mobile in the native state but rendered rigid in the presence of an active-site inhibitor (Hornak et al, 2006a). Thus polyproteins are not cleaved for maturation and infectivity to occur (Kandathil et al., 2009).

Most inhibitors have a non-hydrolyzable hydroxyethylene or hydroxyethylamine moiety that mimics the tetrahedral transition state of proteolytic reaction. Some inhibitors, e.g., two-carbon-elongated inhibitor, associate with the active site by direct hydrogen bonds and indirect hydrogen bonds via two water molecules, and unlike shorter inhibitors they only accept hydrogen bonds from one of the Asp25 (Wu et al, 2008).

LPV was first developed based on RTV architecture where P_3 isopropylthiazolyl group of RTV that associated with the wild-type V82 residue was first removed. The thiazolylmethoxycarbonyl group in the P_2' was then replaced with dimethylphenoxyacetyl group, yielding LPV whose P_1-P_1' positions were occupied by the same hydroxyethylene peptidomimetic as in RTV (Pokorná et al., 2009).

1.6 Problem Statement and Justification

Subtype C is the most predominant in South Africa (Bessong, 2008; Papathanasopoulos et al., 2002), and epidemiological trends had hypothesized this clade, which is widespread in Africa, would dominate the HIV pandemic in the future (Kandathil et al., 2005). It now accounts for a 50% estimate of the global infection (Dalai et al., 2009) and its global prevalence is rapidly increasing (Archary et al., 2010). According to the WHO 2011 Progress Report, the infected population in South Africa stands at 5.6 million, and represents one of the highest global endemic. Its 1.5% incidence is still high despite the decline from 2.4%. Nonetheless, the country represents one of the few countries where both child and maternal mortalities have increased in the 21st century, but now there exists national policies to address the scourge in a country where the 2011 mid-year population estimates were 50.59 million (Statistics South Africa, 2012). According to 2012 National Strategic Campaign, the South African Department of Health hopes to significantly reduce HIV/AIDS

incidence, prevalence and mortalities in adults and children by 2015 through accurate disease management (CARMMA, 2012).

HIV-1 clade B commonest in Western Europe and most drugs in the market were designed to target clade B, yet non-subtype B is pandemic and present residue variability, for example in the protease translate. Prevalence of subtype B is subsiding whereas that of subtype C is rapidly increasing hence unmasking the need to engineer drugs based on C consensus, but the role of other non-synonymous mutations in clade C remains an enigma (Bessong, 2008). There is poor characterization of drug-resistance mutational effects in non-clade B (Ohtaka & Freire, 2005), for instance, the data in hand for clade C are mostly conflicting (Bessong, 2008). Much biological and therapeutic information of non-clade B, subtype C in particular, remains unexploited and unavailable (Archary et al., 2010; Ode et al., 2007). At the global consensus subtype C differs from subtype B in eight positions: T12S, I15V, L19I, M36I, R41K, H69K, L89M and I93L. Unlike other clades, clade C presents a disproportionate increase due to transmissibility ease and viral fitness (Bessong, 2008). There are reported differences in both viral fitness and pathogenesis rates between sub-type B and C (Jakobsen et al., 2010). Current HIV-1 drug resistance and susceptibility indicates that in absence of drugs, non-subtype B strains still have high frequency rates to develop non-synonymous mutations that are associated with drug resistance in subtype B; again non-subtype B have a different mechanism for developing drug-associated resistance (Nukoolkarn et al., 2004). Clade genomic variations can be up to 30% (Ohtaka & Freire, 2005).

HIV therapy started to improve when protease inhibitors were introduced (Olsen et al., 1999). Emergence of drug-resistant HIV-1 protease variants thwarts a serious blow to the available inhibition therapies (Piana et al., 2002) even with the availability of FDA approved PIs designed to curb the infection (Arts & Hazuda, 2012; Brik & Wong, 2003). The disease continues to globally spread to catastrophic dimensions (Petrokova, 2006; Sham et al., 1998), claiming millions of life every year (Wu et al., 2008; Deeb & Jawabreh, 2012). Due to the constant resistance of the virus to protease inhibitors, there is the need to design arsenal of new PIs. In South Africa RTIs are associated with resistance within the first two years of initiation (Bessong, 2008). Poor pharmacokinetic properties of majority of the current FDA approved HIV-1 protease inhibitors also emphasizes the urgent need for effective therapies (Kaplan & Hicks, 2005; Wu et al., 2008).

The Lock and Key model based drug construction only informs the development of PIs with conformational constraints that become ineffective upon selection of certain mutations. Therefore adaptive drugs with structural optimization endowed with high affinity, specificity and response need to be designed. Optimization is done by enforcing flexible asymmetric functionalities in the inhibitor moieties oriented towards the mutated receptor sites. This can only be effectuated through gathering of information pertaining to critical interactions involved in thermodynamics especially in HIV-1 mutants (Ohtaka & Freire, 2005). Inhibition of HIV-1 protease renders the virions non-infective (van der Kuyl, 2012). Binding dynamics of ligands is now the subject of computational drug research design (Mao, 2011).

The pandemic is great in developing countries, yet HAART therapy is very expensive (Brik & Wong, 2003) even as intervention urgency rests high (Bessong, 2008). There is no conventional approach to curtail drug resistances (Brik & Wong, 2003) as evidenced by clinical, immunological and virologic failure despite both optimal adherence and therapeutic drug range (Zyl et al., 2011), therefore the sustained need (Martin et al., 2005) to develop potent PIs (Li et al., 2011) with lower tendencies of leading to drug-resistances (Ali et al., 2010). Limited information subsists for the drug resistance mechanisms in non-clade B strains, especially for the PIs (Martinez-Cajas et al., 2012), but so far the protease represents the most attractive target for drug discovery (Deeb & Jawabreh 2012; Ode et al., 2007).

1.7 Aim and Objectives

Sequences were obtained from 29 infants infected with HIV-1. These were HIV-1 *pol* sequences prior to and after drug exposure to combination regimen containing either RTV or boosted LPV. Homology modelling was performed to generate the 3-dimensional (3D) protease structures that were used to interpret the impact of non-synonymous mutations on both FDA approved and RU-synthesized protease inhibitors utilization using both docking and molecular dynamic simulations. Apart from the molecular aspects, the structural mechanisms leading to drug resistance were also studied in order to guide drug engineering.

1.8.1 Goal

To identify both structural and molecular mechanisms by which non-synonymous mutations affect the docking and efficacies of protease inhibitors in order to determine how current and or future drugs ought to be modified or developed to combat drug resistance.

1.8.2 Objectives

1) To build homology models of the HIV-1 C proteases from existing crystal structures.
2) To identify amino acid signatures at baseline and under drug selection pressure that confer resistance in clade C infected infants
3) To predict the pattern of occurrence of mutations linked to drug resistance.
4) To identify drug resistance mechanisms using architectural studies and docking studies complemented by and molecular dynamics calculations.
5) To use the ligand-receptor atomic and surface interactions to determine changes in drug engineering to target HIV-1 drug mutants.

1.8 Hypotheses

1) The Rhodes University engineered drugs are potent in HIV-1 protease antagonists.
2) Some HIV-1 clade C protease polymorphisms at baseline are linked to resistance and can be used as signatures for therapy selection focussed to abate virological failure.
3) HIV-1 clade C still has more masked molecular mechanisms leading to PIs drug resistance and cross-resistance.
4) Drug-resistant mutations have a pattern of occurrence that could be predicted only if such mutations reveal synergism.

1.9 Limitation of this Study

1) The sample size of 29 sequences from infants is not an ideal representation of the South African population currently infected.
2) There is restriction to the number of the heterocyclic analogues of RTV that the RU Organic research group can synthesize from the Baylis-Hillman reaction.
3) The *in silico* docking is meant to identify the mechanisms by which non-synonymous mutations affect the docking and efficacies of protease inhibitors. This project assumes that the mechanisms leading to drug resistance can be attested from the determination of docking energies, inhibition constants and atomic interactions, yet the interaction between the receptor and ligand comprises multifaceted steps: approach, ligand and binding site desolvation, penetration of ligand into binding cleft, orbital steering, conformational adoption, and interaction through hydrogen,

electrostatic, van der Waals and hydrophobic forces. This *in vitro* course may not be captured absolutely but the *in silico* docking will approximate the binding energy via:

$$\Delta \boldsymbol{G_{bind}} = \Delta \boldsymbol{G_{sol}} + \Delta \boldsymbol{G_{tor}} + \boldsymbol{G_{conform}} + \Delta \boldsymbol{G_{hbond}} + \Delta \boldsymbol{G_{vdw}} + \Delta \boldsymbol{G_{elec}}$$

where = ΔG_{bind} = Free binding energy, ΔG_{sol} = Energy due to desolvation effects, ΔG_{tor} = Energy due to internal ligand torsions, $\Delta G_{conform}$ = Energy due to deviation from the covalent geometry = ΔG_{hbond} = Energy due to hydrogen bonding, ΔG_{wdw} = Energy due to dispersion/repulsion, ΔG_{elec} = Energy due to electrostatic forces (Toor et al., 2011).

4) Competition of PI with substrate is not accounted for.
5) Time-scale is also a drawback. Atomistic simulations are limited by computational time to shorter timescales.

CHAPTER TWO

2. INTRODUCTION

2.1 Protease Sequence Analysis and Homology Modelling Scope

This section attempts to characterize the mutations that are commonly, uncommonly and rarely associated with HIV-1 protease C, particularly in drug-naïve patients. The grade of polymorphisms of these mutations and the consequences of all mutations in drug susceptibilities are evaluated. The sequences of occurrence of mutations linked to drug-resistance are also predicted. 3D structures of the target sequences are also generated using Modeller v9.10 and validated before being docked to the PIs.

2.2 HIV-1 Protease Molecular Characterization

Protease is one of the key elements in the life cycle of HIV. It facilitates production of a functional viral assemblage (Petrokova, 2006). HIV is a dynamic virus that maximizes on viral fitness through constant evolution (Ali et al., 2010; Barbour & Grant 2005). Molecular and epidemiological studies have indicated that there are different HIV subtypes in circulation (Bessong, 2008; Kandathil et al., 2005; McCutchan, 2000). Mutations occur either naturally or in presence of drugs (Ohtaka & Freire, 2005; Toor et al., 2011) in the protease gene and can be used to subtype the virus (Kear et al., 2011; Shafer et al., 2001). Certain natural polymorphisms are more greatly associated with a specific subtype than in others. For instance, 8 natural genetic variations (already cited in Problem Statement and Justification subsection) have been observed to have high frequency of occurrences in clade C in South Africa, as when compared against B (Bessong, 2008). The Indian subcontinent where clade C is also endemic, 9 codon differences have been reported in clade C wild-type. They include 12S, 15V, 19I, 36I, 41K, 63P, 69K, 89M and 90L (Toor et al., 2011). Existences of natural mutations in non-subtype B have the potential to confer drug resistance (Kantor et al., 2002; Nukoolkarn et al., 2004) that can be transmitted between individuals (Shafer et al., 2001).

Drug resistance researches have been based on subtype B (Nukoolkarn et al., 2004). It can be carried out using phenotypic, genotypic and or docking studies (Toor et al., 2011). Normally, mutations can usually be assessed using algorithms modelled to suit subtype B. These include the Stanford HIVdb (the gold standard), National Agency for AIDS Research

(ANRS), Sierra Web service (Tang et al., 2012), the Visible Genetics/Bayer Diagnostics Guidelines 6.0 (VGI) and AntiRetroScan (ARS) (Zazzi et al., 2004). The Stanford HIVdb computation entails an alignment between the wild-type B *pol* protein and the query protein (that has been translated from nucleotides using an inbuilt conversion algorithm) (Tang et al., 2012). It gives information on sequence quality, flags and defines mutations, and computes mutational scores (approximations of diminished efficacies) derived from summation of mutation penalty scores of each significant mutation on a given drug. The mutation score has five levels: susceptible (net score of 0-9), potential low-level resistance (net score of 10-14), low-level resistance (net score of 15-30), inter-mediate resistance (net score of 30-59), high-level resistance (net score of ≥ 60) (Tang et al., 2012; Zazzi et al., 2004).

RIP, a Los Alamos program, uses similarity search and bootscanning similarity while REGA uses bootscanning and phylogenetic signals during subtyping. REGA subtyping is defined after tight clustering has been executed, with boostrap replicates of ≥ 70% (de Oliveira et al., 2005).

2.3 Homology Modelling Scope

Homology modelling, also known as computational or comparative modelling (Sali & Blundell, 1993), is used to predict reliable and accurate the 3D structure of unknown proteins (Guex et al., 2009) based on high sequence identity between the unknown and the high-resolution evolutionary structure proteins that act as templates. This technique replaces the slow experimental approaches (Nuclear Magnetic Resonance and X-ray crystallography techniques) that have not been able to generate protein structures at the same rate as the genes generated from sequencing. It comprises of 4 key steps (Figure 2-1):

1) template identification using database search programs ,e.g. ScanP, PsiBLAST,
2) sequence and structure alignment using Clustal, 3DCoffee, FUGUE, etc.
3) model creation and refinement using Modeller or other programs, and,
4) validation of the model using Procheck, ANOLEA (Taştan Bishop et al., 2008), ProSA (di Luccio & Koehl, 2011) among others. Validation flags regions of conformational anomalies both at the streochemical and energy levels (Guex et al., 2009).

Various approaches exist for generating models. The Modeller program, for instance, builds the models from restraints satisfaction extracted from alignment (Sali & Blundell, 1993).

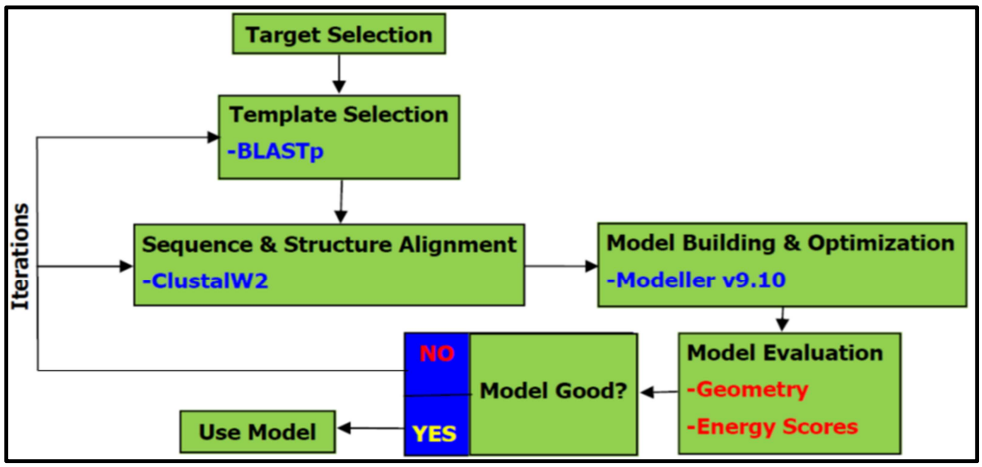


Figure 2-1: Overview of the homology modelling steps.

Every model has errors but these are of less concern if they are localized and do not occur in the region of interest (Krieger et al., 2003). There are programs that provide geometrical statistics and or energy scores of the residues to assess the protein conformation, as supported by overall stereochemical quality of the residues falling within the experimental range (Sánchez & Sali, 1997). For example, in ProSA, a protein is affirmed to be stable if the energy distribution/fluctuations of the amino acids are below the zero flagging value.

Homology modelling can be used to build membrane structures, design drugs (Taştan Bishop et al., 2008). Other uses include directed mutagenesis, analysing mutational effects and SNPs (Guex et al., 2009), and prediction of protein-protein interactions among others (Dibrov et al., 2009). Despite its success, it is still subject to criticisms, thus has not been approved as an alternative to experimental structures (di Luccio & Koehl, 2011).

2.4 METHODOLOGY

2.4.1 Quality Assessment and Subtype Characterization

58 *pol* sequences from 29 infants before and after treatment with either RTV or LPV/r in addition to NRTIs were obtained from the National Institute of Communicable Diseases (NICD) (Table 2-1). The infants were first treated with RTV for the first 6 months before their therapy was switched to LPV/r. Both clinical and virological failures were used as therapeutic monitoring modules. Therapy-failing infants had >1,000 viral copies/ml despite the implementation of both optimal therapeutic dosage and adherences. The nucleotide sequences were trimmed to 297 bases long, only to house the protease region. The RT frame was eliminated. A script was built, PRnucl-Truncator.py (Supplementary

Data/AllScripts/SampleProcessing), which was used to carry out this. The script was run in Python v2.7.3. The nucleotides were converted into amino acids using the web-based ExPasy Translate Tool (Artimo et al., 2012). Subtype assignment was performed using the Stanford HIVdb v6.2.0 (that also accesses the sequence quality), REGA HIV-1 subtyping tool v2.0 and RIP (Recombinant Identification Program) v3.0 (de Oliveira et al., 2005) (Table 2-2).

Table 2-1: Details of the infant cohort indicating their sample identities, duration of sample recollection after PI therapy initiation. The protease inhibitor(s) utilized in the therapy is/are also shown (Mathu, 2012).

Patient ID	Sample Recollection After Therapy	Protease Inhibitor(s) used
3018	12 months	RTV and LPV/r
3021	12 months	RTV and LPV/r
3043	Pre-random	-
3051	12 months, 52 weeks	RTV and LPV/r
3059	9 months	RTV and LPV/r
5014	24 weeks	-
5032	52 weeks	RTV and LPV/r
5045	9 months	RTV and LPV/r
5046	12 months	RTV and LPV/r
5074	36 weeks	RTV and LPV/r
5079	6 months	RTV
5080	12 months	RTV and LPV/r
5086	12 months	RTV and LPV/r
5089	9 months	RTV and LPV/r
5094	24 weeks	RTV
5114	12 months	RTV and LPV/r
5117	12 months	RTV and LPV/r
5144	12 months	RTV and LPV/r
5169	9 months	RTV and LPV/r
5175	*	RTV
5178	24 weeks	RTV
5198	9 months	RTV and LPV/r
5207	6 months	RTV
5211	12 months	RTV and LPV/r
5228	12 weeks	RTV
5242	3 months	RTV
5245	4 weeks	RTV
5261	12 weeks	RTV

Sample 3043 had blood recollected during an unscheduled date. * indicates that the patient did not turn up for follow-up. In cases where PI therapy was not used, "-" is used as an indicator.

Table 2-2: The online programs used for translation, subtyping and mutation assessment.

Name of the online tool	URL
ExPasy Translate Tool	http://web.expasy.org/translate/
Stanford HIVdb v6.2.0	http://hivdb.Stanford.edu/
REGA HIV-1 subtyping tool v2.0	http://dbpartners.stanford.edu/RegaSubtyping/
RIP v3.0	http://www.hiv.lanl.gov/content/sequence/RIP/RIP.html/

2.4.2 Assignment, Frequencies and Pattern Determination of Non-Synonymous Mutations

Both the nucleotide and protein sequences were assessed for any character specific change. Any unusual base or deletion apart from the 'CATG' code required detection and counting. To achieve this, the MutationScanFlagger.py script was written. However, at the protein level, the consensus B in the Stanford HIVdb was used. Consensus B nucleotide sequence from RIP was translated by ExPasy Translate Tool, and aligned by Clustalw2 (Kandathil et al., 2009) with consensus C protein generated from the alignment of the drug-naïve sequences as viewed in Jalview v2.7. Alignment between the two clade consensuses was also visualized in Jalview to access conservation and polymorphisms in HIV-1 clade C. The frequencies of each missense mutation at baseline and after treatment initiation were calculated based on consensus C protease sequence to evaluate their degrees of occurrence/polymorphism and therapeutic criticality, i.e., mutations linked to drug resistances and susceptibilities of the different PIs. Mutations were interpreted using the Stanford HIVdb where drug resistance signatures were also identified for each sequence, at baseline and after therapy initiation.

2.4.3 Template Search and Selection

Possible templates, both open and closed conformations, were searched from the PDB based on resolution of $\leq$ 2.00 Å. This was followed by an alignment between target and the templates in order to select the template of choice based on percentage identity score and the expectation value (E-value). BLASTp (protein-protein Basic Local Alignment Search Tool) from the National Center for Biotechnology Information (NCBI) was used. Parameterization entailed use of 3 amino acid window size, a threshold of 10, BLOSUM62 pairwise alignment scoring matrix and gap costs of 11 and 1 for existence and extension, respectively for gaps. Selected templates were evaluated using the following web-based programs (Table 2-3).

Table 2-3: The online programs used for template search and selection, target-template alignment, and model validation.

Name of the online tool	URL
PDB	http://www.rcsb.org/pdb/home/home.do
BLASTp	http://blast.ncbi.nlm.nih.gov/
Verify3D	http://nihserver.mbi.ucla.edu/Verify_3D/
ANOLEA	http://swissmodel.expasy.org/workspace/index.php?func=tools_structureassessment1
Procheck	http://swissmodel.expasy.org/workspace/index.php?func=tools_structureassessment1
ProSA	http://www.came.sbg.ac.at/prosa.php
QMEAN6	http://swissmodel.expasy.org/workspace/index.php?func=tools_structureassessment1

2.4.4 Validation of the Homology Modelling Scripts

Two modelling scripts, homodimer.py and model_m2.py, from the Modeller package 9.10 (http://salilab.org/modeller/manual/) were revised and validated with respect to which best predicted a reliable model. DOPE score computations were implemented into each of the scripts to develop the homodimer+dope.py and model_m2+dope.py (Supplementary Data/AllScripts/HomologyModelling) in order to automate rankings of the generated models and aid towards selection of the most reliable model. Unlike the latter script, the former script builds multi-chain models through the introduction of extra restraints and satisfaction of symmetry restraints, ensuring that the two chains are rendered identical, with symmetry violations being reported. Models were generated using the "very slow" refine mode. Out of the 100 models built for each target sequence, the model with the least DOPE and Z-DOPE scores (computed using the adapted and re-edited getdopezdope_scores.py (Supplementary Data/AllScripts/HomologyModelling) were evaluated for satisfaction of restraints, e.g., stereochemical restraints, involved in building. Energy scores were also considered (Sánchez and Sali 1997). Web-based programs were used for validation.

2.4.5 Generation and Evaluation of the 3D Structures of the HIV-1 C Protease

Modeller v9.10 was used for homology modelling, which is a multistep process that encompasses manual template identification, sequence and structural alignment, model building and refinement, and model validation (Sánchez & Sali, 1997). Possible templates were searched for and manually selected from PDB. These sequences were used to construct an in-house database. Target sequences were then searched against the in-house database using BLASTp and appropriate templates selected based on identity scores and E-values.

Using existing crystal structure data as templates for a series of monomers, homology modelling and protein prediction of the dimeric HIV-1 protease were performed. Target-template alignment and conversion of the fasta to PIR file were carried out using Clustalw2 (http://www.ebi.ac.uk/Tools/msa/clustalw2/). Parametization was set as follows: the alignment type was slow, BLOSUM as the weighted matrix and gap opening and extension set at 10 and 0.1 respectively. Most of the parameters were left as default.

2.5 RESULTS AND DISCUSSION

2.5.1 Quality Assessment and Subtype Characterization

By comparison of the three subtyping algorithms, all the sequences were confirmed to be of subtype C except for sample 5252 (both before and after drug exposure). It did not meet the minimum criteria to be included in the study. This sequence was of subtype A according to the Stanford HIVdb. Similarity search and bootscanning similarity according to RIP were non-significant. REGA, that clusters query with the reference (de Oliveira et al., 2005), defined it as subtype C. An agreement of at least two subtyping tools was used to define the subtype.

2.5.2 Assignment, Frequencies and Pattern Determination of Non-Synonymous Mutations

The mutations were analyzed both before and after drug exposure. The script MutationScanFlagger.py was developed and used (Supplementary Data/Chapter II/AllScripts/SampleProcessing), to show that at the nucleotide level, both synonymous and non-synonymous mutations do occur. The result, N2baseline failing-TRIMMED-SCANNED is included in the supplementary information (Supplementary Data/Chapter II/HARRiS_MutationFlagger_v1.0). Synonymous mutations occur in the third position of the codon. Due to the degeneracy feature of the genetic code, they manifest as silent mutations at the amino acid level. Non-synonymous mutations occur in any position of the codon.

In drug-inexperienced infants, 9 mutational positions were found to occur in high frequencies (Table 2-4). They included 12S (78.6%), 15V (92.9%), 19I (96.4%), 36I (92.9%), 41K (89.3%), 63P (78.6%), 69K (100.0%), 89M (92.9%), and 93L (96.4%). Our data were in agreement with findings from South Africa and India by Bessong, 2008 and Toor et al., 2011, respectively. 63P was absent in works by the former. These are the nine consensus positions in which clade C that differ from clade B. Figure 2-2 indicates the 9 differences between consensuses B and C. Consensus C for the study was generated from the multiple sequence alignment of the drug-naïve sequences. The identity score for the two sequences was 90.91%. Common mutations e.g., 20R (28.6%) and 35E (39.3%), were also observed in drug-naïve infants. Some mutations such as 20M (3.6%), 23I (3.6%), 57K (3.6%) and 70R (3.6%) were classified as rare mutations due to their low frequencies of occurrences. Some amino acid positions had the propensity of at least two amino acids either occurring (in those

positions). Examples of this are 19I/T/V, 36I/L, 37E/K/S, 41I/N/K, and 63L/T/V. Deletions were observed to occur rarely but one of the samples (5079) among the drug-naïve had a deletion at position 10 in the *pol* region.

Table 2-4: Type and prevalence of natural and drug-induced mutations in infant cohort. Consensuses B and C have been used as references.

Reference: Consensus B Amino Acid	Mutations Before Treatment			Reference: Consensus C Amino Acid	Mutations After Treatment (Besides Naturally Occurring Polymorphisms)		
	Mutation(s)	Frequency (%)			Mutation(s)	Frequency (%)	
		As per each signature	Total			As per each signature	Total
L10	-	3.6	3.6	L10	F/I/M	7.1/3.6/3.6	14.3
V11	D	3.6	3.6	S12	K/S/T	3.6/3.6/3.6	10.7
T12	A/S	3.6/75.0	78.6	K14	K	3.6	3.6
I13	V	7.1	7.1	I19	I	3.6	3.6
K14	N/R	3.6/3.6	7.1	K20	R	3.6	3.6
I15	V	92.9	92.9	L23	I	3.6	3.6
G16	E	7.1	7.1	E35	D	7.1	7.1
L19	I/T/V	67.9/21.4/7.1	96.4	N37	S	3.6	3.6
K20	M/R	3.6/28.6	32.1	K45	R	3.6	3.6
E35	D	39.3	39.3	M46	I	10.7	10.7
M36	I/L	82.1/10.7	92.9	I54	V	25.0	25.0
N37	E/K/S	3.6/10.7/7.1	21.4	D60	E	3.6	3.6
R41	I/N/K	3.6/3.6/82.1	89.3	Q61	H	3.6	3.6
R57	K	3.6	3.6	L63	P/T	10.7/3.6	14.3
D60	E	14.3	14.3	K70	R	3.6	3.6
Q61	D/E	3.6/3.6	7.1	T74	T	3.6	3.6
I62	V	3.6	3.6	G78	R	3.6	3.6
P63	L/T/V	53.6/3.6/21.4	78.6	V82	A/I	28.6/3.6	32.1
I64	M	3.6	3.6	M89	I	3.6	3.6
C67	Y	7.1	7.1	L90	M	3.6	3.6
H69	K	100.0	100.0	**KEY**			
K70	R	3.6	3.6	Small nonpolar	G, A, S, T		Orange
A71	T	3.6	3.6	Hydrophobic	C, V, I, L, P, F, Y, M, W		Green
T74	S	17.9	17.9	Polar	N, Q, H		Magenta
V77	I	7.1	7.1	Negatively charged	D, E		Red
V82	I	10.7	10.7	Positively charged	K, R		Blue
L89	M	92.9	92.9	Deletion	-		Black
I93	L	96.4	96.4				

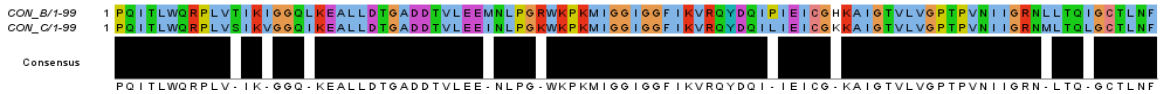


Figure 2-2: Pairwise sequence alignment between HIV-1 consensuses B and C (which was obtained from our sequences). The white regions in the black block indicate the nine key sequence differences between consensus B and consensus C.

Most of these mutations are non-active site mutations. Figure 2-3 and Figure 2-4 indicate that even though mutations e.g., I13V, R41K, D60E, etc., occur in HIV-1 C protease, there is still a significant aspect of conservation in terms of physicochemical properties in order to retain the enzyme functionality. The evident structural change is in terms the residue size.

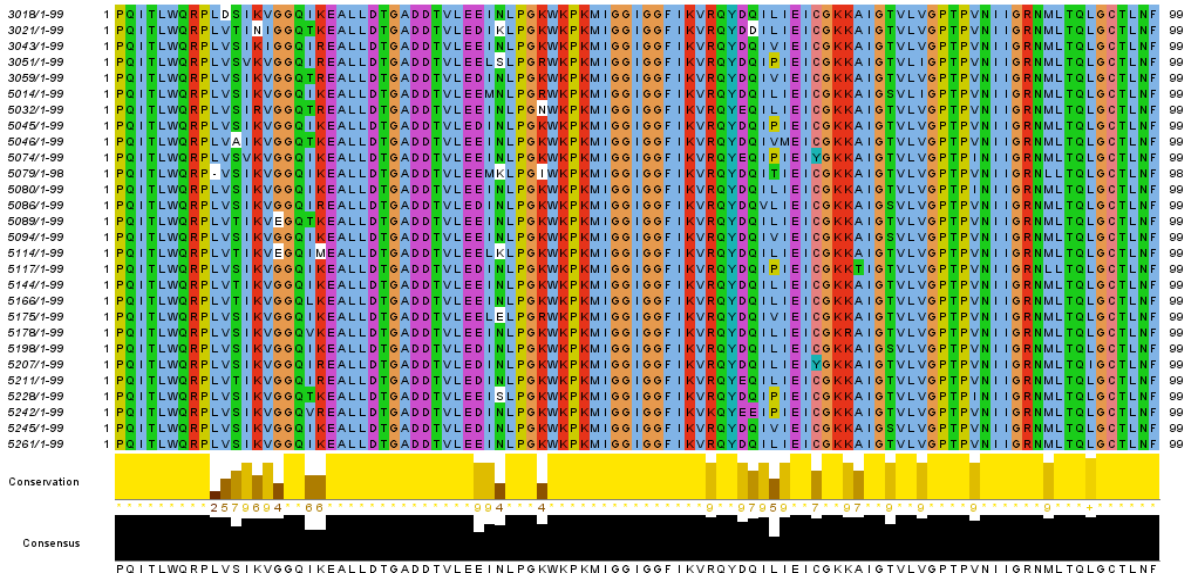


Figure 2-3: Multiple sequence alignment of sequences from drug-naïve infants, indicating consensus C and conservation according to physico-chemical features. White spots in the alignment columns indicate amino acids with different features.

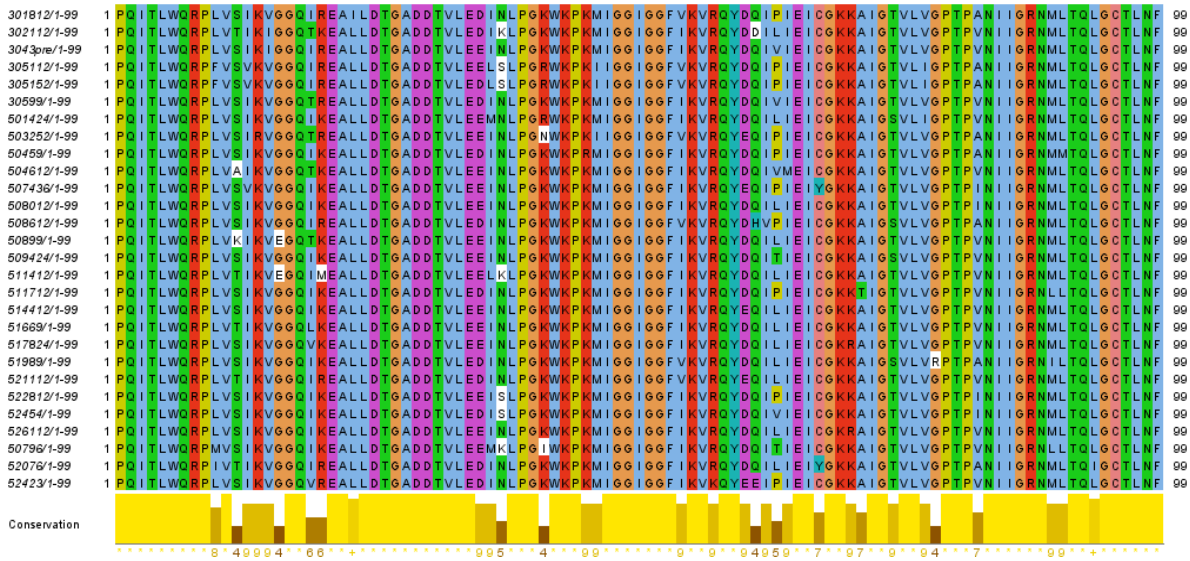


Figure 2-4: Multiple sequence alignment of sequences from drug-failing infants, indicating conservation according to physico-chemical features. White spots in the alignment columns indicate amino acids with different features.

In drug-failing infants, drug-associated mutations occurred, besides natural polymorphisms. Out of the 28 patients (one eliminated after subtyping), only 16 (57.1%) manifested changes after drug exposure as shown the pairwise alignments (Figure 2-5). The samples include 3018, 3021, 3051, 5032, 5045, 5079, 5086, 5089, 5094, 5144, 5198, 5207, 5211, 5228, 5245, and 5261. This was consistent with the preceding study (Mathu, 2012).

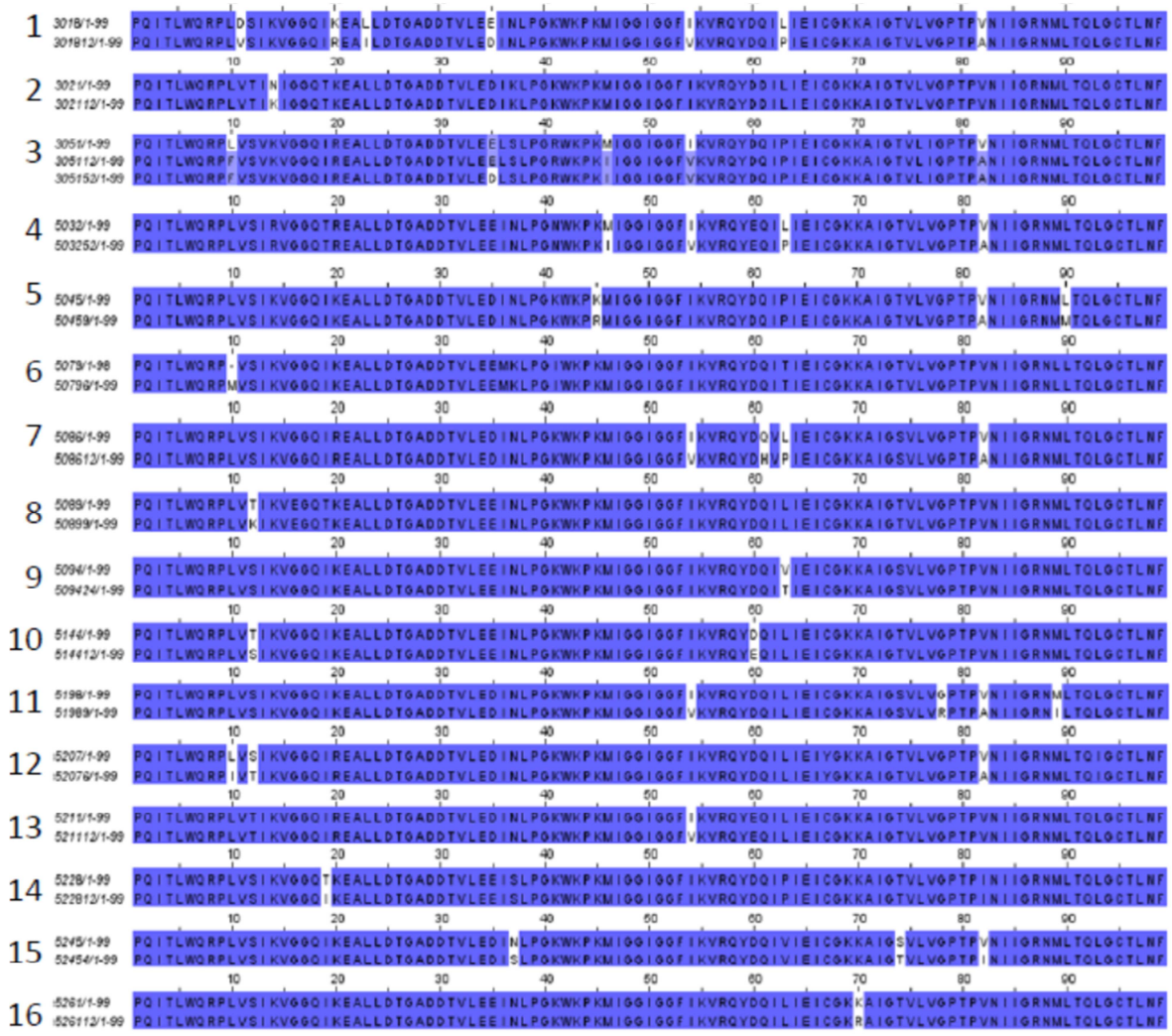


Figure 2-5: Pairwise sequence alignment between drug-naïve and drug-exposed infant sequences. White region(s) in the alignment column indicate(s) amino acid differences between two alignments, before and after therapy (Mathu, 2012).

Mutations arose naturally or were caused by drug and or host selection pressures. 9 infants acquired minor and or major mutations (Table 2-5); out of which 8 (88.9%) had both major and minor mutations. 19 (67.9 %) of the infants lacked drug-associated mutations. Infants on RTV did not show drug-linked mutations (according to Stanford HIVdb) probably due to the short timeline between drug initiation and sample collection. For those on LPV/r therapy, the reason is the association of this drug to low mutation incidences (Kaplan & Hicks, 2005).

Table 2-5: Stanford HIVdb drug resistance reports of patient samples with either major or minor or both of mutations.

Patient ID	Regimen used	Major Mutation	Minor Mutation	Other Mutations	Drugs to which Resistance is Conferred (Low, Intermediate, High)
301812	RTV, LPV/r	I54V, V82A	L23I	T12S, I15V, L19I, K20R, E35D, M36I, R41K, L63P, H69K, L89M, I93L	ATV/r, fAPV/r, IDV/r, LPV/r, NFV, SQV/r, TPV/r
305112	LPV/r	M46I, I54V, V82A	L10F	T12S, I13V, I15V, L19I, K20R, M36L, N37S, L63P, H69K, V77I, L89M, I93L	ATV/r, fAPV/r, IDV/r, LPV/r, NFV, SQV/r, TPV/r
305152	LPV/r	M46I, I54V, V82A	L10F	T12S, I13V, I15V, L19I, K20R, E35X, M36L, N37S, L63P, H69K, V77I, L89IM, I93L	ATV/r, fAPV/r, IDV/r, LPV/r, NFV, SQV/r, TPV/r
503252	LPV/r	M46I, I54V, V82A	F53FL	T12S, K14R, I15V, L19T, K20R, E35DE, M36I, R41N, D60E, L63P, H69K, L89M, I93L	ATV/r, fAPV/r, IDV/r, LPV/r, NFV, SQV/r, TPV/r
50459	LPV/r	V82A, L90M	NONE	T12S, I15V, L19I, K20KR, E35D, M36I, R41K, K45R, L63P, H69K, L89M, I93L	ATV/r, fAPV/r, IDV/r, LPV/r, NFV, SQV/r
508612	LPV/r	M46IM, I54V, V82A	L10FIL	T12S, K14R, I15V, L19I, K20R, E35D, M36I, R41K, Q61H, I62V, L63P, H69K, T74S, L89M, I93L	ATV/r, fAPV/r, IDV/r, LPV/r, NFV, SQV/r, TPV/r
511712	LPV/r	NONE	A71T	T12S, I15V, L19I, E35D, M36I, R41K, L63P, H69K, I93L	NONE
51989	LPV/r	I54V, V82A	L33FL	T12S, I15V, L19I, K20KR, E35D, M36I, R41K, H69K, T74S, L89I, I93L	ATV/r, fAPV/r, IDV/r, LPV/r, NFV, SQV/r, TPV/r
52076	RTV	V82A	L10I	I15V, L19I, K20R, E35D, M36I, R41K, K45KR, C67Y, H69K, L89M	ATV/r, fAPV/r*, IDV/r, LPV/r, NFV, SQV/r*
521112	RTV, LPV/r	I54V	NONE	I15V, L19I, K20R, E35D, M36I, R41K, D60E, H69K, L89M, I93L	ATV/r, fAPV/r*, IDV/r, LPV/r*, NFV, SQV/r, TPV/r

*Potential-low level resistance.

PI resistance reports offer avenues for predicting patterns of mutations (Mao, 2011). The sequence of occurrence of V82A, I54V and M46I major mutations in the infant cohort was predicted based on the proportion of infant samples harbouring these mutations. However, this was limited to the fraction of samples having mutations. V82A was due to RTV selection pressure. 1 infant on RTV and 7 infants on LPV/r dose developed major drug resistance. 7 (87.5%) had V82A, which could be due to the previous RTV therapy before switching to LPV/r. Out of these 8 infants, 2 (25.0%) had both V82A and 154V after drug exposure. 1 (12.5%) had I54V being selected on its own, 2 (25.02%) and 3 (37.5%) dual and triple synergism respectively. It appears that the occurrence and drug response impact of M46I could be dependent on the presence of V82A/I54V. One sample had V82A/L90M dual major mutation. Occurrence pattern of these two mutations could not be defined due to sample and mutation constraints. Studies by Dirauf et al., 2010 revealed that both V82A single and I54V/V82A dual mutations highly occur with L90M of the dimer interface. We encountered a 14.2% frequency. Subtype differences could account for this. 3 (42.9%) infants acquired the triple major mutation of V82A/I54V/M46I, a signal proposing that M46I could be selected for by the mutants harbouring both V82A and I54V. Based on the results, the predicted patterns of occurrence are V82A→I54V→M46I or I54V→V82A→M46I, but further analysis in a (retrospective) longitudinal study is needed to confirm this since there could be a possibility

of M46I being selected for before I54V, but the chance of occurrence of M46I→I54V could be much less as compared to I54V→M46I. I54V is likely to be selected for after V82A. Gradual accumulation of these mutations might also be reinforcing viral fitness besides causing resistance. The degree of cross-resistance increased as the level of the aforementioned major mutations accumulated. The triplet mutation, V82A/I54V/M46V, is linked to high-grade cross-resistance according to the Stanford HIVdb report, and this is due to the extra involvement of M46I, since V82A/I54V dual mutation mostly correlates with intermediate cross-resistance. I54V in the aforesaid dual mutation causes high-level NFV resistance, since V82A alone does confer intermediate NFV resistance.

Drug resistance is discussed in detail in chapter III which also compares our docking against those from Stanford HIVdb. According to the Stanford HIVdb, in subtype B, both M36I and I93L are accessory mutations that weakly confer resistance, but the former requires participation of other mutations are in background. M36L is uncommon and its role remains as an enigma. K20R, L63P and V77I are common polymorphisms selected for by PIs. Among other amino acids such as Met, Ile, Thr and Val that can occupy codon position 20, it is only Arg that has the effect on PI resistance. V77I is known to be selected by NFV. T74S is also alters NFV susceptibility and has a 5% incidence in subtype C drug-naïve subjects. A71T/V emerges in 2-3% of untreated individuals and usually its incidence concomitantly increases when therapy is initiated. V82I in substrate cleft is also a common polymorphism in some non-B subtypes and has infinitesimal effect on drug response. V77I polymorphism is common and is linked to drug-resistance in phenotypic studies (Shafer et al., 2001).

The database further reports I54V, a flap mutation (Shafer et al., 2001), conferring susceptibility only to DRV. It works with V82A/S/T to confer PI resistance. V82A lowers response to boosted IDV and LPV. It imparts resistance to NFV, boosted ATV, SQV and fAPV. Coexistence of V82A and I54V is highly associated with high-grade NFV resistance. L23I is a rare mutation positioned in the substrate cleft and enforces a low-grade NFV resistance. M46I/L is a flap mutation (Johnston et al., 2004; Shafer et al., 2001) that reduces response to IDV/r, fAPV, LPV/r, ATV/r and NFV in presence of other mutations. L10I/V/F/R/Y is an accessory mutation (Shafer et al., 2001) that confers resistance to most when other mutations are present. The last two amino acids substitutions at codon position 10 are poorly studied. L10I/V has 5-10% incidence in drug-naïve population. L10F is linked to all PI

therapy interference with the exception of ATV/r, SQV/r and TPV/r. L10M is an unusual mutation. L90M can alter susceptibility if it exists as either singly or with other mutations. On its own, it affects NFV, SQV/R, IDV/r and ATV/r therapy. LPV/r and fAPV are affected when other mutations are present. F53L is a flap mutation that reduces susceptibility to SQV/r and ATV/r. L33F is due to DRV/r, LPV/R, ATV/r, TPV/r, and fAPV/r drug pressures.

2.5.3 Template Search and Selection

Possible templates were searched for and extracted from the PDB with particular attention to low-resolution structures. There is a direct correlation between resolution and structures. Low-resolution values correspond to high-resolution structures while high-resolutions relate to low-resolution structures. Accurate models are built when high-resolution crystallized structures are used as templates (Taştan Bishop et al., 2008). The Supplementary Data/Chapter II/Validation/DataofPossible_HIV-1C-ProteaseTemplate has more details.

Templates were selected from the list of possible templates based on the sequence identity scores and E-values that were generated from the BLASTp search. BLASTp is a similarity search tool that uses a protein sequence as a query to search for its related sequences in the database containing protein sequences. It is a heuristic algorithm that searches for high-scoring segment pairs that have a maximal score above a defined threshold.

Scoring alignments are used to compute similarities between the query and the available templates in the database. For the BLASTp, BLOSUM was used since it yields better searches than matrices constructed on evolutionary rates. It gives improved alignments and searches in closely related sequences (Henikoff & Henikoff, 1992). Table 2-6 and Table 2-7 show part of the data from BLASTp search.

Sequences identity scores ranged 84-97% and 78-90% for the closed and open conformation, respectively. Reliable models are often built if the target-template similarity is >75% (Rodriguez et al., 1998), else the models are likely to be compromised mainly if the score is <30% due to increase in alignment errors (Taştan Bishop et al., 2008). >40% predicts good models (di Luccio & Koehl, 2011). Sequence coverage was 100% and the E-values ranged 1e-69 - 5e-58 and 2e-67 - 1e-57, in the same order of conformation. E-value is a statistical measure depicting the probability that the target-template alignments are by chance.

Table 2-6: Selected templates (closed conformation) for modelling after search from blast (based on E-value and coverage).

No.	SEQUENCE ID	IDENTITY (%)	E-VALUE	COVERAGE (%)	TEMPLATE(S) CLOSED	RESOLUTION (Å)
1.	3018	89	6e-66	100	1HXB	2.30
2.	301812	86	1e-64	100	1RL8	2.00
3.	3021	89	6e-65	100	1HXB	2.30
4.	302112	90	1e-65	100	1HXB	2.30
5.	3043	89	3e-66	100	1HXB	2.30
6.	3043pre	89	3e-66	100	1HXB	2.30
7.	3051	88	3e-66	100	1HXB	2.30
8.	305112	85	3e-59	100	1RL8	2.00
9.	305152	84	1e-58	100	1RL8	2.00
10.	3059	87	3e-65	100	1HXB	2.30
11.	30599	87	3e-65	100	1HXB	2.30
12.	5014	90	1e-67	100	1HXB	2.30
13.	501424	90	1e-67	100	1HXB	2.30
14.	5032	87	8e-65	100	1HXB	2.30
15.	503252	85	5e-58	100	1RL8	2.00
16.	5045	88	3e-65	100	1HXB	2.30
17.	50459	87	9e-65	100	1RL8	2.00
18.	5046	88	2e-65	100	1HXB	2.30
19.	504612	88	2e-65	100	1HXB	2.30
20.	5074	85	2e-63	100	1HXB	2.30
21.	507436	85	2e-63	100	1HXB	2.30
22.	5079	90	8e-64	100	1HXB	2.30
23.	50796	90	4e-66	100	1RL8	2.00
24.	5080	89	1e-66	100	1HXB	2.30
25.	508012	89	1e-66	100	1HXB	2.30
26.	5086	86	2e-65	100	1HXB	2.30
27.	508612	84	9e-64	100	1RL8	2.00
28.	5089	90	6e-66	100	1HXB	2.30
29.	50899	89	3e-65	100	1HXB	2.30
30.	5094	88	3e-66	100	1HXB	2.30
31.	509424	88	7e-66	100	1HXB	2.30
32.	5114	89	2e-65	100	1HXB	2.30
33.	511412	89	2e-65	100	1HXB	2.30
34.	5117	88	8e-65	100	1HXB	2.30
35.	511712	88	8e-65	100	1HXB	2.30
36.	5144	91	3e-67	100	1HXB	2.30
37.	514412	89	3e-66	100	1HXB	2.30
38.	5166	92	1e-67	100	1HXB	2.30
39.	51669	92	1e-67	100	1HXB	2.30
40.	5175	90	5e-67	100	1HXB	2.30
41.	5178	89	2e-66	100	1HXB	2.30
42.	517824	89	2e-66	100	1HXB	2.30
43.	5198	88	5e-66	100	1HXB	2.30
44.	51989	87	8e-65	100	1RL8	2.00
45.	5207	88	5e-65	100	1HXB	2.30
46.	52076	87	2e-64	100	1HXB	2.30
47.	5211	88	1e-65	100	1HXB	2.30
48.	521112	87	2e-65	100	1HXB	2.30
49.	5228	89	7e-66	100	1HXB	2.30
50.	522812	89	3e-66	100	1HXB	2.30
51.	5242	84	3e-63	100	1HXB	2.30
52.	52423	84	1e-63	100	1HXB	2.30
53.	5245	87	1e-65	100	1HXB	2.30
54.	52454	88	2e-66	100	1HXB	2.30
55.	5261	90	6e-67	100	1HXB	2.30
56.	526112	89	2e-66	100	1HXB	2.30
57.	CON_B	97	1e-69	100	1HXB	2.30
58.	CON_C	90	6e-67	100	1HXB	2.30

Table 2-7: Selected templates (open conformation) for modelling after search from blast (based on E-value and coverage).

No.	SEQUENCE ID	IDENTITY (%)	E-VALUE	COVERAGE (%)	TEMPLATE(S) OPEN	RESOLUTION (Å)
1.	3018	81	3e-62	100	1TW7	1.30
2.	301812	82	3e-64	100	1TW7	1.30
3.	3021	80	2e-60	100	1TW7	1.30
4.	302112	81	3e-61	100	1TW7	1.30
5.	3043	82	5e-63	100	1TW7	1.30
6.	3043pre	82	5e-63	100	1TW7	1.30
7.	3051	80	2e-63	100	1TW7	1.30
8.	305112	82	1e-58	100	1TW7	1.30
9.	305152	81	4e-58	100	1TW7	1.30
10.	3059	80	6e-62	100	1TW7	1.30
11.	30599	80	6e-62	100	1TW7	1.30
12.	5014	81	5e-63	100	1TW7	1.30
13.	501424	81	5e-63	100	1TW7	1.30
14.	5032	79	6e-61	100	1TW7	1.30
15.	503252	82	1e-57	100	1TW7	1.30
16.	5045	82	6e-64	100	1TW7	1.30
17.	50459	83	5e-65	100	1TW7	1.30
18.	5046	81	3e-62	100	1TW7	1.30
19.	504612	81	3e-62	100	1TW7	1.30
20.	5074	80	4e-62	100	1TW7	1.30
21.	507436	80	4e-62	100	1TW7	1.30
22.	5079	82	9e-60	100	1TW7	1.30
23.	50796	82	7e-62	100	1TW7	1.30
24.	5080	82	5e-63	100	1TW7	1.30
25.	508012	82	5e-63	100	1TW7	1.30
26.	5086	80	3e-62	100	1TW7	1.30
27.	508612	82	7e-64	100	1TW7	1.30
28.	5089	82	4e-62	100	1TW7	1.30
29.	50899	81	2e-61	100	1TW7	1.30
30.	5094	81	7e-63	100	1TW7	1.30
31.	509424	81	4e-63	100	1TW7	1.30
32.	5114	80	1e-60	100	1TW7	1.30
33.	511412	80	1e-60	100	1TW7	1.30
34.	5117	83	5e-64	100	1TW7	1.30
35.	511712	83	5e-64	100	1TW7	1.30
36.	5144	83	1e-63	100	1TW7	1.30
37.	514412	81	2e-62	100	1TW7	1.30
38.	5166	84	7e-64	100	1TW7	1.30
39.	51669	84	7e-64	100	1TW7	1.30
40.	5175	82	1e-62	100	1TW7	1.30
41.	5178	81	1e-62	100	1TW7	1.30
42.	517824	81	1e-62	100	1TW7	1.30
43.	5198	80	2e-62	100	1TW7	1.30
44.	51989	81	4e-62	100	1TW7	1.30
45.	5207	80	3e-61	100	1TW7	1.30
46.	52076	83	1e-62	100	1TW7	1.30
47.	5211	80	5e-62	100	1TW7	1.30
48.	521112	81	3e-62	100	1TW7	1.30
49.	5228	82	2e-63	100	1TW7	1.30
50.	522812	82	1e-63	100	1TW7	1.30
51.	5242	78	7e-62	100	1TW7	1.30
52.	52423	78	7e-62	100	1TW7	1.30
53.	5245	80	2e-62	100	1TW7	1.30
54.	52454	80	6e-62	100	1TW7	1.30
55.	5261	82	4e-63	100	1TW7	1.30
56.	526112	81	9e-63	100	1TW7	1.30
57.	CON_B	90	2e-67	100	1TW7	1.30
58.	CON_C	82	4e-63	100	1TW7	1.30

The E-values were significant, indicative of how far the alignment was by chance. Both the identity score (based on percentage sequence coverage) and E-values were manually used to select the template since initial alignment error is often minimized if templates are manually selected (Sánchez & Sali, 1997). Template selection is therefore a crucial step and depends on the alignment accuracy. For example, a one amino acid shift will result in a 3.8 Å distortion in the predicted structure (di Luccio & Koehl, 2011). Modelling depends on the type of amino acid, amino acid equivalence in correlated proteins and protein similarity (Sali & Blundell, 1993). 1HXB and 1RL8, of closed conformation were used as templates. For the open conformation, 1TW7 was used.

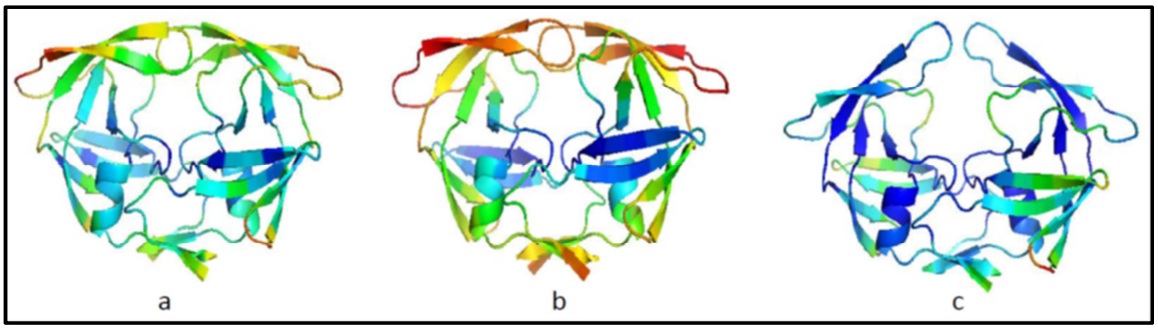


Figure 2-6: 3D structures of 1HXB (a), 1RL8 (b) and 1TW7 (c) as visualized by Pymol. Regions highlighted according to the B factor profiles.

Only 1HXB and 1RL8 had SQV and RTV, respectively, bound to their active sites. Their quality assessment results are shown in Figure 2-6. Due to time restriction, not all data could be written. For 1TW7, only statistics have been used, and the associated figures are in the Supplementary Data/Chapter II/Validation/Summary-TemplateSelection&Validation.

Procheck evaluates stereochemical quality. From its geometrical report, the templates were of reliable quality and could be employed in homology modelling even though a few atomic bond lengths and angles were distorted. A Ramachandran plot is a map of the torsions phi (ϕ) and psi (ψ) that show the conformation distributions of amino acids in making the backbone. It has 4 quadrants: quadrant 1 (left-handed helix), quadrant 2 (beta sheet), quadrant 3 (right-handed helix) and quadrant 4. Figure 2-7A indicates the plots of the two templates. Evaluation of experimental results dictates that a good quality model should have at least 90% amino acids (spotted using black markers) in the most favourable core (region 3: red). Templates 1HXB, 1RL8, and 1TW7 had 95.6%, 94.9% and 94.2% amino acids in within region 3. Within the additionally allowed region (region 2: yellow), 4.4%, 5.1% and

5.8% amino acids were marked. Red markers are used to map residues in unfavourable regions. None of the residues were in the generously allowed region (region 1: cream) and the disallowed region (region 0). Stereochemical evaluation alone, for example, from the Ramachandran plot, is not satisfactory hence the need to use other verification programs (di Luccio & Koehl, 2011).

According to the Verify3D (Figure 2-7B), the quality folding of each residue was good. The threshold values for 1HXB and 1RL8 were 0.32 and 0.3 respectively. The scoring profiles for the three templates were above their set thresholds, suggesting that both their secondary structures and solvent exposures ranks were good, i.e., the quality of the residues in the proteins were good since most of the residues occurred in the favourable environment.

The ANOLEA accesses the non-local atomic interactions energy. Figure 2-7C shows that part of the flap regions were of low quality since there interaction energies were above the zero threshold value. Generally the templates were good since most of the amino acid residue positions were defined as favourable regions. The same scenario was apparent for 1TW7.

ProSA works by computing energies in single residues and residues in pairs (Sánchez & Sali, 1997). It evaluates protein solvent interaction and pairwise potential using the Cβ atoms (di Luccio & Koehl, 2011). From the Z scores -4.59 and -4.41, the overall model qualities were within the experimentally acceptable limit, as shown in Figure 2-7D. 1TW7 had a score of -5.15. These scores were compared against those from X-ray crystallography and NMR. The overall quality model was within the normal range as the energy distribution/fluctuations of the amino acids were below the zero flagging value, hence protein stable. From the Jmol (the interactive mol viewer) the 3D structure of the protein had some erroneous parts as translated by the energy distribution. This information is in the Supplementary Data. The ProSA plot of residues scores which graphs energy as a function of amino acid position confirmed the good local quality of the two templates. The energy values of the residues were < zero-threshold. Positive values indicate erroneous regions. None of these regions were observed in the optimized window size of 40 residues, which systematically computes the energy terms. In the background, the window size is 10 amino acid residues, but with large energy fluctuations. A plot of energy as a function of a single residue is greatly associated with energy fluctuations, and this explains why a large window size was used.

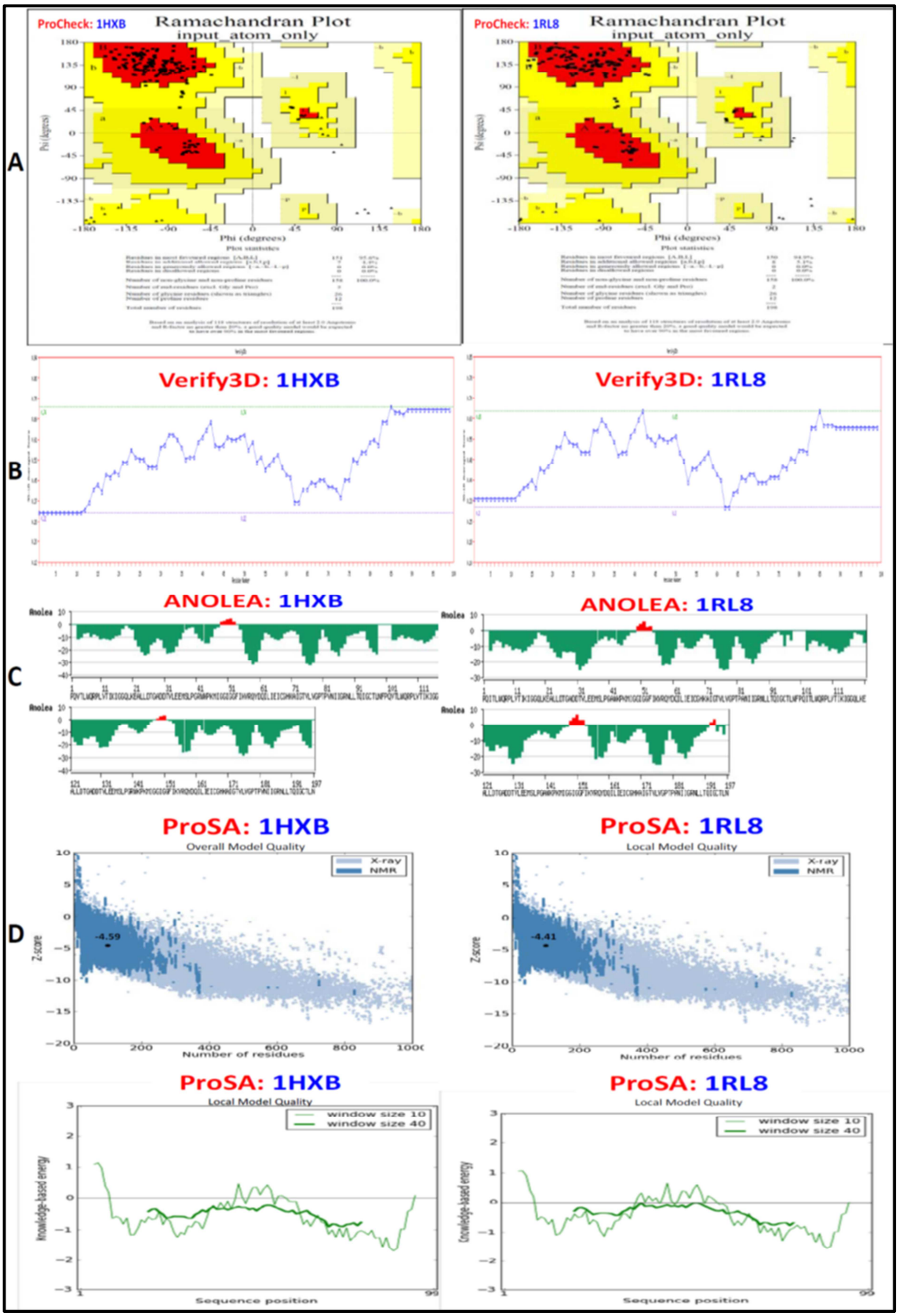


Figure 2-7: 1HXB and 1RL8 validation profiles from four web-based programs. **A** - Ramachandran plots from Procheck, **B** - Verify3D, **C** - ANOLEA models for non-local atomic interaction energy, and **D** - Overall and local qualities from ProSA.

QMEAN (Quality Model Energy Analysis) is also another web-based embedded program that was used for model evaluation and is as effective as ProSA and DOPE. It is an "absolute indicator" similar to the R-factor of X-ray crystallography that evaluates the geometrical aspects of a model (di Luccio & Koehl, 2011). A reliable model is one with a score of 0-1 (Benkert et al., 2009). Templates 1HXB and 1RL8 had QMEANscores of 0.613 and 0.624, respectively as seen in Figure 2-8A, while that of 1TW7 was 0.585.

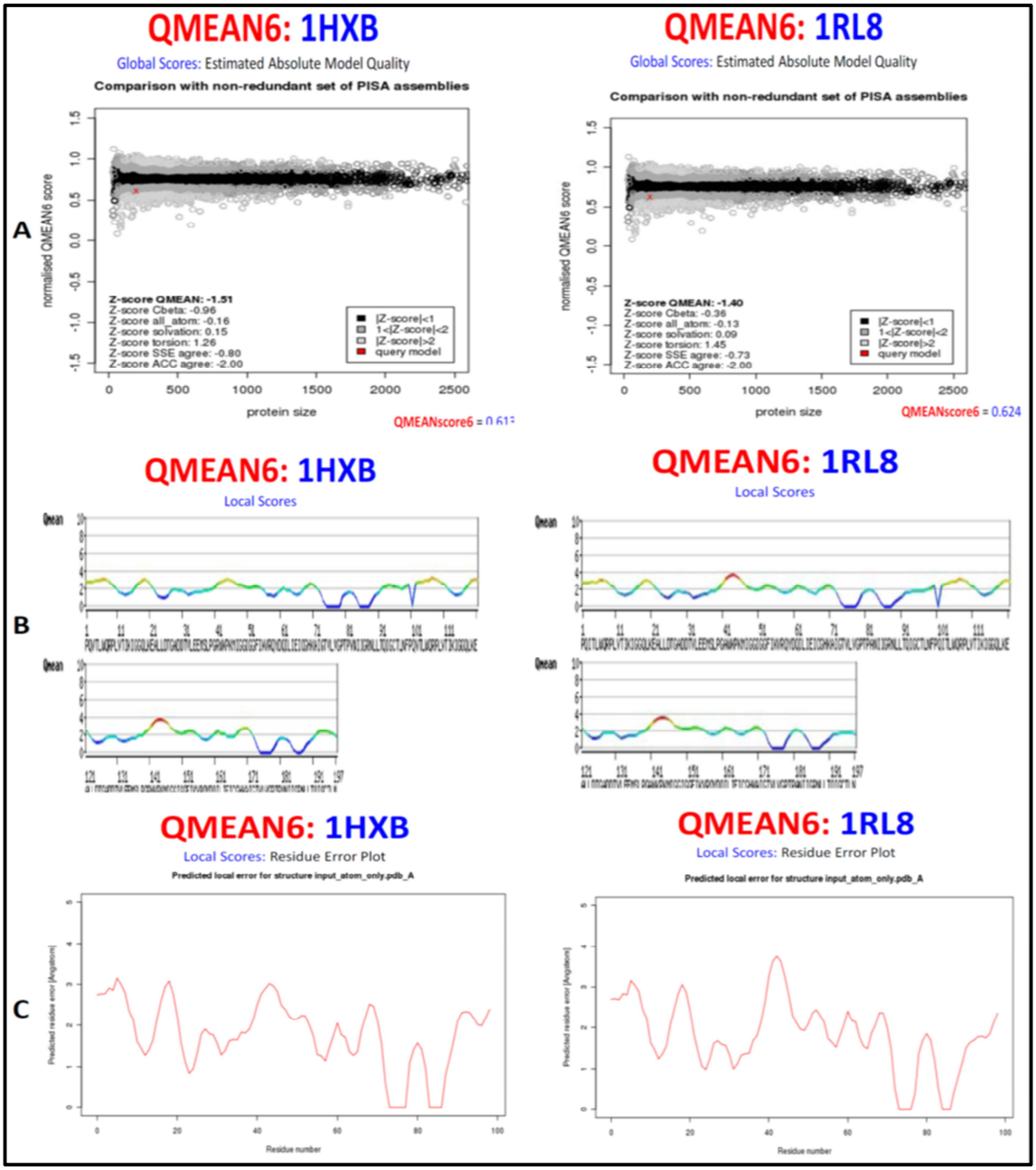


Figure 2-8: QMEAN results of template verification. **A** - Global evaluation based on 6 parameters for estimation of absolute model quality. **B** - Local quality assessment indication favourable and unfavourable residues, **C** -Predicted local error for the structure (one chain).

QMEAN scoring function has two evaluation options: composite scoring profile (QMEAN) that evaluates a model based on its geometry and evolution based on known structures. The consensus or clustering-based scoring profile (QMEANclust) does both local and global evaluation against a number of diverse models that the user feeds into the server. It computes the Z-score which is a measure of the statistical potential of mean force. The Z-score Cβ and all_atoms are distance-dependent interaction potentials that analyze long-range interactions. Solvation score evaluates hydrophobicity rank of the residues while torsion potential describes the local geometry over three amino acids window size. “SSE agree” and “ACC agree” designate the match between the predicted and computed secondary structure and solvent accessibility. Of these scores, only the torsion and solvation potentials compared to those of non-redundant set of PISA (Protein Interfaces, Surfaces and Assemblies). Figure 2-8B displays the quality of the residues. The local scores for the residues were good except for those in the loop regions. 1HXB has one of its chains having a problematic flap. For 1RL8, both the flaps had been flagged as erroneous regions. 1TW7 reported the same as 1RL8. Predicted local error for the structure with respect to a combination of eight parameters, uses a nine-amino acid window size. Figure 2-8C, displays the residue error due to inappropriate conformation as a result of solvent exposure. For all the templates, the error was quite great in regions 40 - 60, where flaps rest. Jmol is used to visualize problematic areas (Benkert et al., 2009). It colours residues by error, with blue being associated with the least energy whereas red is the highest. More QMEAN data in the Supplementary Data proves that the two templates were of good quality. Based on comparison of these validation suites, all the templates were of good quality.

2.5.4 Validation of the Homology Modelling Scripts

Clustalw2 was used for global target-template pairwise alignment and to generate the PIR format file. The PIR format is the alignment format that Modeller (one of the modelling programs (di Luccio & Koehl, 2011) recognizes. The homodimer+dope.py and model_m2.py were validated by rebuilding the three templates (1HXB, 1RL8 and 1TW7) and building and evaluation of a sample randomly selected (3018). Based on the consensuses of evaluation of the modelled structures, both the scripts constructed equally good models, as compared from the available crystal structures initially validates. Only the flap regions were reported to possess unfavourable regions, as also seen in the crystal structures evaluation reports.

Results from Procheck (Figure 2-9 and Figure 2-10) and Verify3D in Supplementary Data/Chapter II Validation/Summary – ModelScript_Validation&Selection show no notable differences in the models from the two scripts, for the three templates.

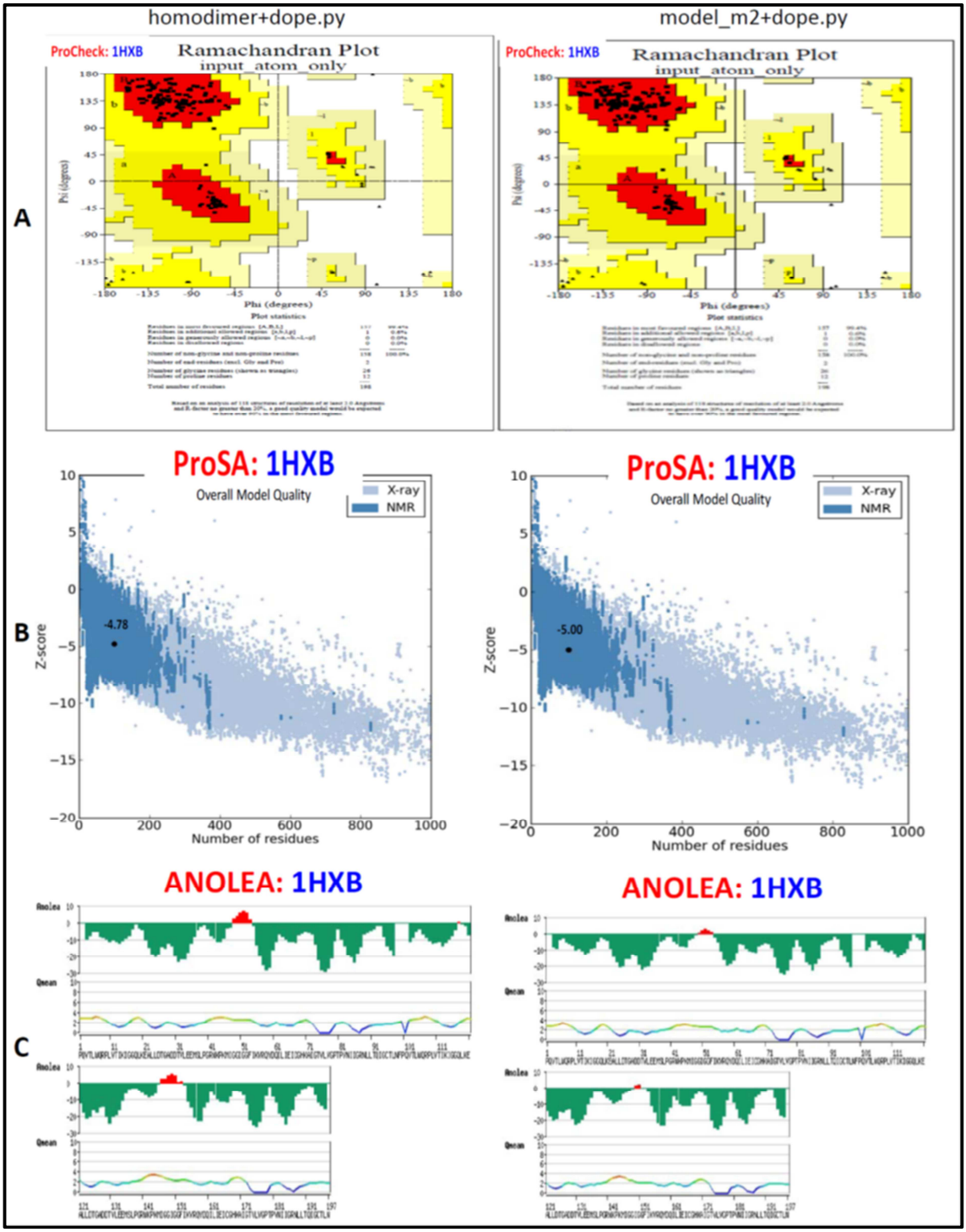


Figure 2-9: Quality assessment of the 1HXB template from the two modelling scripts (Diagrams on right and left are from homodimer+dope.py and model_m2+dope.py, respectively). **A** - Ramachandran plots from Procheck, **B** - Overall quality from ProSA and **C** – ANOLEA for non-local atomic interaction energy.

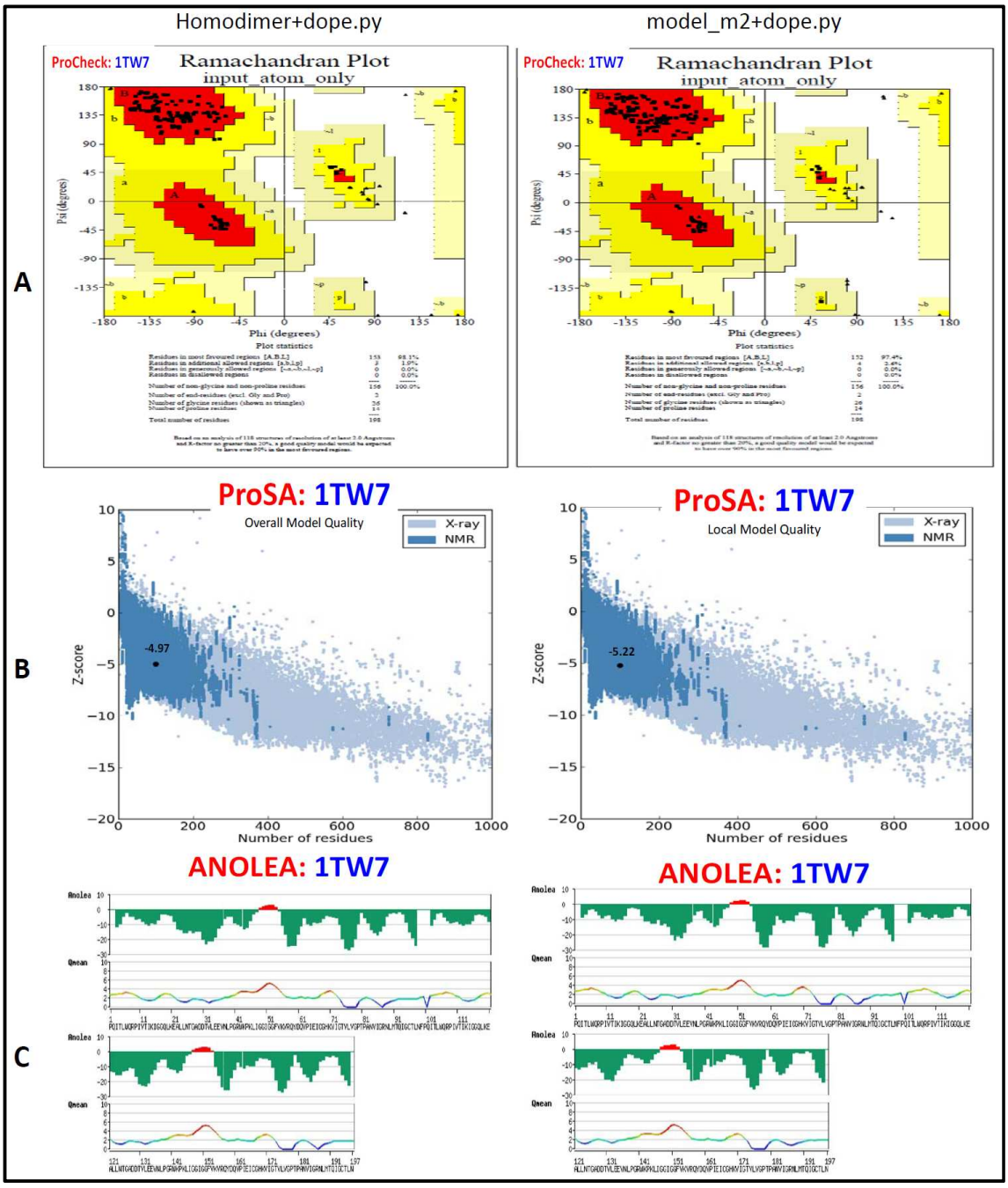


Figure 2-10: Quality assessment of the 1TW7 template from the two modelling scripts (Diagrams on right and left are from homodimer+dope.py and model_m2+dope.py, respectively). **A** - Ramachandran plots from Procheck, **B** - Overall quality from ProSA and **C** - ANOLEA for non-local atomic interaction energy.

Ramachandran statistical data for the rebuilt 1HXB indicated that the best model built from the homodimer+dope.py had 99.4% and 0.6% of its residues in the most favourable and additionally allowed regions respectively. Similarly, the best model from model_m2+dope.py had 99.4% and 0.6% in that order. For 1TW7, the scores from the two scripts were 98.1% and 1.9, and 97.4% and 2.6%, correspondingly.

ProSA Z scores for homodimer+dope.py and model_m2+dope.py were -4.78 and -5.00 for 1HXB whereas those of 1TW7 were -4.97 and -5.22, respectively. ANOLEA results depicted good models for both but tended to favour model_m2+dope.py in 1HXB reconstruction. For 1TW7, the ANOLEA showed that both the scripts are equally good. The reconstruction of the crystallized structure by homology modelling was used as a positive control, of which it passed, as analyzed by the validation results.

Similarly, the constructed models of 3018 from the two scripts were of good quality and were comparable to the template in closed conformation that was in alignment. Some verification programs tended to show that model_m2+dope.py was more reliable than homodimer+dope.py, while others were either in favour of homodimer+dope.py or both. For example, ANOLEA showed that the model generated by model_m2+dope.py was better than that by homodimer+dope.py, when 1HXB was utilized as a template. Modelling of 1TW7 showed no preference to a particular script according to ANOLEA. Based on the Verify3D Ramachandran plot, both the scripts generate models whose stereochemical qualities are similar. ProSA statistics depicted favourability in the model structures, but homodimer+dope.py constructed a model with a Z score of -4.58 that was close to that of the template (-4.59); model_m2+dope.py Z score was -4.56 for reconstructed 1HXB. In 1TW7 reconstruction, the scores were -4.67 and -4.55 for homodimer+dope.py and model_m2+dope.py, respectively, while that of the template was -5.15. Even though all the ProSA scores were within the range that is usually generated from native proteins of similar sizes; hence showing that the models still good, ProSA favoured homodimer+dope.py.

DOPE (Discrete Optimized Protein Energy) and DOPE Z scores were also used in evaluation. DOPE in Modeller, a statistical potential that evaluates the native structure of a protein, identifies the best models that have large negative values. The best model for the 3018 target built from script that introduces extra restraints had a DOPE score of 21,452.18 whilst that from the other script had 21,382.46 for the closed conformation. For the open conformation, the scores were -20,786.14 and -20,743.00, respectively. DOPE Z is a normalized DOPE that is dependent on atomic distance and compares proteins of different sizes. Reliable models have negative scores (di Luccio & Koehl, 2011). The scores for the 3018 models were -1.15 and -1.13, and -0.91 and -0.90, for each script and each model

respectively. Both of these energy scores showed that homodimer+dope.py is more superior to model_m2+dope.py in homology modelling.

Consensuses results of all these validation programs indicate that the two scripts build equally good models. Evaluation of models is crucial since model accuracy determines how fit the model is for a defined use. A poor model will often be flagged as containing problematic zones (Benkert et al., 2009). The models were good; only few errors particularly in the flap regions were translated from the template. Aligned structure of model against template had the Cα atoms in traceable positions. This was measured by RMSD (Root Mean Square Deviation) which is a score of the distance of two Cα atoms in alignment that fall within the allowed region based on cut-offs (Sánchez & Sali, 1997).

2.5.5 Generation and Evaluation of the 3D Structures of the HIV-1 C Protease

There are various approaches for homology modelling. A few examples include distance geometric approach, energy based approach and Modeller, whose models possess good stereochemistry. Modeller builds models by automatic extraction and satisfaction of spatial restraints (expressed as probability density function, pdfs) from alignments (Sali & Blundell, 1993). There are several alignment algorithms and alignment scoring approaches. Clustalw2 was used in our study. Models construction is dependent on the capability to find a related structure and the quality of building and refinement (di Luccio & Koehl, 2011). Initial alignment error is often minimized if templates are manually selected. Key control points with the exception of manual template selection and editing that could be used to improve on the model quality include iteration and model building coupled to validation (Sánchez & Sali, 1997) of the final optimized model (di Luccio & Koehl, 2011).

Template selection was based on the E-value and identity scores. The lowest identity for target-template alignment was 78%. The E-values were all significant, showing that the alignments (based on sequence coverage) were not by chance. The whole process of model building and evaluation was automated due to the large number of target sequences whose structures were to be predicted. Ultimately 100 models were to be generated for each target sequence and each had to be validated. Therefore the whole process was automated to eliminate the cumbersomeness associated with subjection of each of the models to the suites of web-based embedded verification programs. 100 models were built for each target

sequence in order to increase the statistical chance of building and selecting the best model from the constructed and available model library. Figure 2-11 shows examples of some of the generated models for both open and closed conformations.

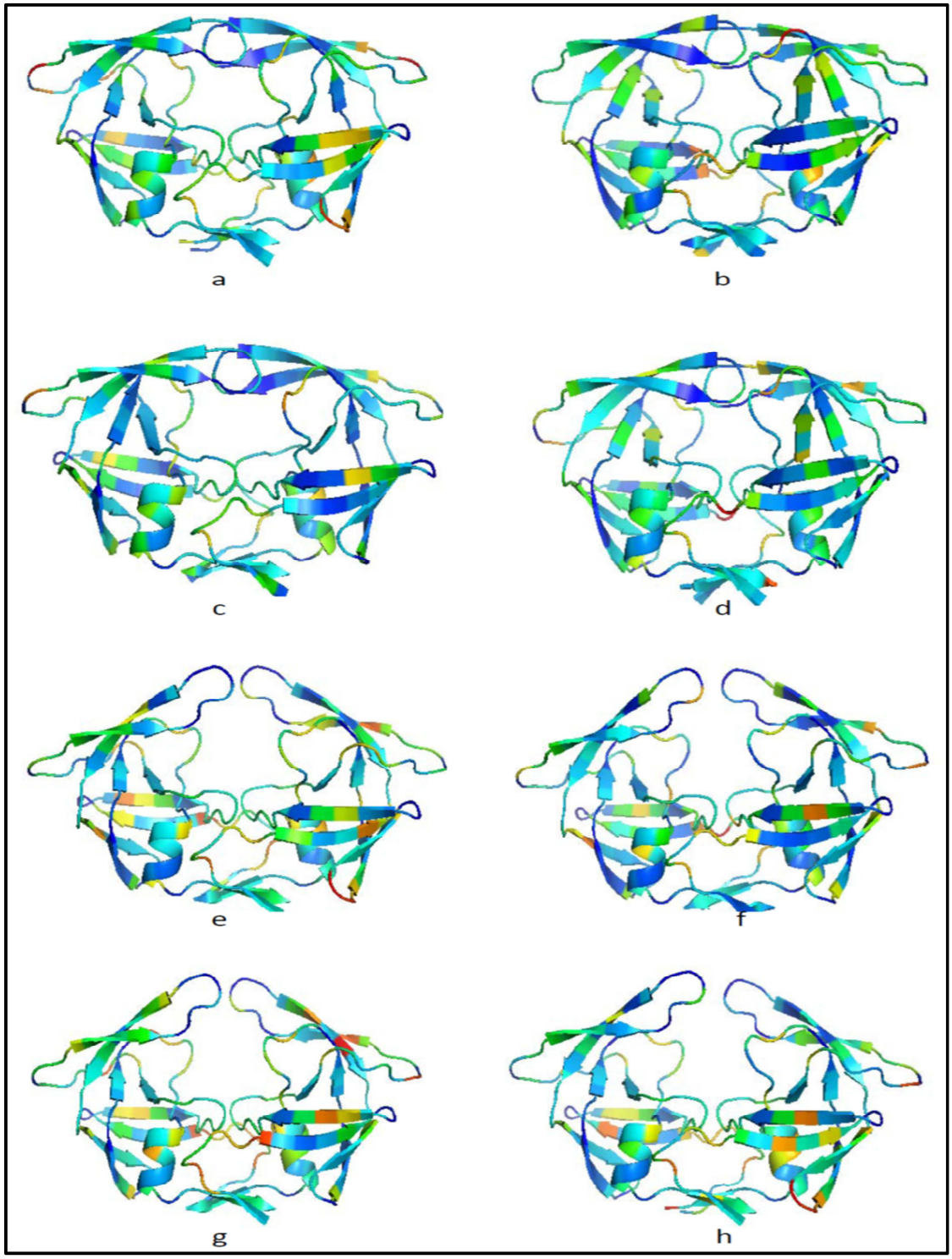


Figure 2-11: Diagrams of selected generated closed models: 3051 (a), 305112 (b), 5086 (c), and 508612 (d); open models: 3018 (e), 301812 (f), 3021 (g) and 302112 (h) as visualized by Pymol. Regions highlighted according to the B factor profiles.

In comparative modelling, some of the stereochemical restraints are bond distances, bond angles, bond lengths, chiralities of Cα and side chains, torsions and peptide bond and side chain planarity. The molecular pdfs restraints are optimized to avoid violations. Non hydrogen atom loci are modelled using conjugate gradient algorithm (Sali & Blundell 1993).

Homodimer+dope.py was imported from a separate script, modelling.py in order to automate the process. All the models built were evaluated in order to select reliable models to be used in protease architectural studies, and in docking and molecular dynamics studies.

GDT-TS (Global Distance Test - Total Score), cRMSD, GA341, (Cartesian/Coordinate Root Mean Square Deviation) DOPE and DOPE Z scores were integrated into the modelling processing to automatically access the models and select the best one(s). DOPE and DOPE Z scores selected the best model while the other computations were used to evaluate model quality. Models selected by DOPE and normalized DOPE scores were highly negative and compared to those of the templates. Good models are the ones that have high negative scores (di Luccio & Koehl, 2011). The scores were in agreement in all cases for the good models as seen in Table 2-8 and Table 2-9. For instance, the best model search by DOPE and DOPE Z scores selected identical models as the best. The DOPE Z score for the best model selected by DOPE was calculated and compared to DOPE Z score of the best model selected by the DOPE Z approach. In all the cases, the two DOPE Z scores matched, thus no deviation. The Supplementary Data (Chapter II/Validation/AutoValidation_allscores) contains these scores. The zero deviations of the DOPE Z scores provided confidence in the models consensus that were finally evaluated and used.

GDT-TS uses Cα to gauge similarity between the model and the experimental structures that are superpositioned (Benkert et al., 2009). It does this by counting the number of equivalent Cα atom pairs that are within a predefined distance 1, 2, 4 and 8 Å (Li et al., 2011). When these cut-offs were used, the scores were always 100%. Therefore to make it more sensitive, new cut-offs were set to be 0.1875, 0.375, 0.75 and 1.5 Å. The generated models for both closed and open conformations had high GDT scores with some (3059, 5032, 5045, 5117, 30599, 301812, 503252, 511712 and CON_B sequences all of closed conformation) being as high as 100%, hence were of good quality based on backbone. GDT score does not consider the side chains (Benkert et al., 2009).

Table 2-8: Scores from the statistical evaluation approaches that guided the search process for the best "closed" model.

No.	Sequence ID.	DOPE Score	DOPE Z Score	% GDT Score	GA341 Score	cRMSD	Model used
1.	3018	-21452.17 (33)	-1.15378 (33)	95.58081	1	0.14129	33
2.	301812	-22061.67 (82)	-1.49437 (82)	100.00000	1	0.00000	82
3.	3021	-22154.62 (11)	-1.40607 (11)	95.32828	1	0.14625	11
4.	302112	-22180.60 (58)	-1.38509 (58)	95.58081	1	0.14496	58
5.	3043	-22440.05 (48)	-1.38401 (48)	95.58081	1	0.14421	48
6.	3043pre	-22440.05 (48)	-1.38401 (48)	95.58081	1	0.14421	48
7.	3051	-21947.11 (87)	-1.34721 (87)	94.69697	1	0.15586	87
8.	305112	-21605.05 (93)	-1.22670 (93)	98.48485	1	0.10785	93
9.	305152	-21670.16 (74)	-1.25942 (74)	98.48485	1	0.10927	74
10.	3059	-22215.59 (04)	-1.4821 (15)	100.00000	1	0.00000	15
11.	30599	-22215.59 (04)	-1.48021 (04)	100.00000	1	0.00000	04
12.	5014	-22050.29 (82)	-1.23899 (82)	95.45455	1	0.15087	82
13.	501424	-22050.29 (82)	-1.23899 (82)	95.45455	1	0.15087	82
14.	5032	-22221.63 (79)	-1.44616 (79)	100.00000	1	0.00000	79
15.	503252	-21502.32 (96)	-1.40625 (96)	100.00000	1	0.00000	96
16.	5045	-22070.67 (38)	-1.40820 (38)	100.00000	1	0.00000	38
17.	50459	-21461.36(49)	-1.27036 (49)	98.35859	1	0.10930	49
18.	5046	-21843.79 (81)	-1.31059 (81)	95.32828	1	0.15023	81
19.	504612	-21843.79 (81)	-1.31059 (81)	95.32828	1	0.15023	81
20.	5074	-21910.22 (40)	-1.30446 (40)	92.92929	1	0.16996	40
21.	507436	-21910.22 (40)	-1.30446(40)	92.92929	1	0.16996	40
22.	5079	-21478.29 (63)	-1.19190 (63)	-	1	-	63
23.	50796	-21966.07 (20)	-1.21712 (20)	97.22222	1	0.12839	20
24.	5080	-22200.82 (100)	-1.25869 (100)	95.83333	1	0.13990	100
25.	508012	-22200.82 (100)	-1.25869 (100)	95.83333	1	0.13990	100
26.	5086	-22071.74 (94)	-1.31194 (94)	92.80303	1	0.17332	94
27.	508612	-21619.06 (10)	-1.34900 (10)	96.96970	1	0.13133	10
28.	5089	-22073.08 (43)	-1.37493 (43)	96.33838	1	0.14130	43
29.	50899	-22042.09 (20)	-1.36568 (20)	95.70707	1	0.14478	20
30.	5094	-21920.28 (42)	-1.25364 (42)	96.21212	1	0.13741	42
31.	509424	-21696.70 (27)	-1.26568 (27)	95.95960	1	0.14317	27
32.	5114	-22535.20 (43)	-1.26791 (43)	96.21212	1	0.14079	43
33.	511412	-22535.20 (43)	-1.26791 (43)	96.21212	1	0.14079	43
34.	5117	-22077.63 (03)	-1.46902 (03)	100.00000	1	0.00000	03
35.	511712	-22077.63 (03)	-1.46902 (03)	100.00000	1	0.00000	03
36.	5144	-22150.86 (38)	-1.27551 (38)	95.95960	1	0.14812	38
37.	514412	-22137.89 (29)	-1.26995 (29)	95.58081	1	0.14257	29
38.	5166	-22124.16 (04)	-1.26523 (04)	93.18182	1	0.17003	04
39.	51669	-22124.16 (04)	-1.26523 (04)	93.18182	1	0.17003	04
40.	5175	-22109.76 (73)	-1.29417 (73)	95.95960	1	0.14203	73
41.	5178	-22157.32 (54)	-1.32459 (54)	95.07576	1	0.15350	54
42.	517824	-22157.32 (54)	-1.32459 (54)	95.07576	1	0.15350	54
43.	5198	-22044.18 (79)	-1.26348 (79)	95.45455	1	0.15361	79
44.	51989	-22030.78 (82)	-1.27092 (82)	97.09596	1	0.12462	82
45.	5207	-22591.88 (66)	-1.39391 (66)	95.20202	1	0.14939	66
46.	52076	-22481.58 (29)	-1.37471 (29)	96.33838	1	0.13682	29
47.	5211	-22482.96 (50)	-1.38980 (50)	95.95960	1	0.13941	50
48.	521112	-22289.72 (29)	-1.36371 (29)	95.83333	1	0.14941	29
49.	5228	-21763.26 (92)	-1.38280 (92)	96.33838	1	0.13941	92
50.	522812	-21866.96 (03)	-1.28493 (03)	96.33838	1	0.14008	03
51.	5242	-21746.62 (41)	-1.37045 (41)	95.20202	1	0.14919	41
52.	52423	-21746.62 (41)	-1.37045 (41)	95.20202	1	0.14919	41
53.	5245	-22046.65 (03)	1.30838 (03)	95.83333	1	0.14636	03
54.	52454	-22276.50 (28)	-1.31541 (28)	95.83333	1	0.14200	28
55.	5261	-22093.05 (08)	-1.26312 (08)	96.59091	1	0.14349	08
56.	526112	-22275.74 (67)	-1.32402 (67)	95.83333	1	0.14089	67
57.	CON_B	-22481.49 (82)	-1.46142 (82)	100.00000	1	0.00000	82
58.	CON_C	-22114.72 (21)	-1.27097 (21)	98.61111	1	0.11377	21

Table 2-9: Scores from the statistical evaluation approaches that guided the search process for the best "open" model.

No.	Sequence ID.	DOPE Score	DOPE Z Score	% GDT Score	GA341 Score	cRMSD	Model used
1.	3018	-20786.14 (17)	-0.91547 (17)	97.09596	1	0.13336	17
2.	301812	-21359.77 (95)	-1.24687 (95)	96.71717	1	0.13586	95
3.	3021	-21466.12 (37)	-1.15965 (37)	93.05556	1	0.16977	37
4.	302112	-21561.46 (37)	-1.16371 (37)	96.08586	1	0.14360	37
5.	3043	-21807.08 (96)	-1.15750 (96)	97.47475	1	0.11866	96
6.	3043pre	-21807.08 (96)	-1.15750 (96)	97.47475	1	0.11866	96
7.	3051	-21139.62 (61)	-1.05805 (61)	97.47475	1	0.12192	61
8.	305112	-20883.13 (78)	-0.96835 (78)	93.68687	1	0.16916	78
9.	305152	-21051.13 (73)	-1.03785 (73)	96.21212	1	0.14608	73
10.	3059	-21550.90 (59)	-1.25272 (59)	94.06566	1	0.16483	59
11.	30599	-21550.90 (59)	-1.25272 (59)	94.06566	1	0.16483	59
12.	5014	-21203.27 (63)	-0.93603 (63)	97.72727	1	0.12624	63
13.	501424	-21203.27 (63)	-0.93603 (63)	97.72727	1	0.12624	63
14.	5032	-21489.23 (08)	-1.19365 (08)	96.46465	1	0.14251	08
15.	503252	-20858.98 (63)	-1.17833 (63)	95.83333	1	0.14889	63
16.	5045	-21331.57 (79)	-1.16930 (79)	92.92929	1	0.19338	79
17.	50459	-20806.69 (42)	-1.03608 (42)	96.84343	1	0.13314	42
18.	5046	-21183.33 (58)	-1.07450 (58)	95.70707	1	0.14347	58
19.	504612	-21183.33 (58)	-1.07450 (58)	95.70707	1	0.14347	58
20.	5074	-21280.25 (23)	-1.07868 (23)	97.72727	1	0.11642	23
21.	507436	-21280.25 (23)	-1.07868 (23)	97.72727	1	0.11642	23
22.	5079	-20674.91 (08)	-0.90148 (08)	-	1	-	08
23.	50796	-21130.89 (29)	-0.91787 (29)	96.59091	1	0.14022	29
24.	5080	-21545.54 (88)	-1.02443 (88)	94.31818	1	0.16555	88
25.	508012	-21545.54 (88)	-1.02443 (88)	94.31818	1	0.16555	88
26.	5086	-21455.21 (65)	-1.09132 (65)	95.95960	1	0.14195	65
27.	508612	-21017.86 (82)	-1.13423 (82)	96.46465	1	0.13533	82
28.	5089	-21355.82 (49)	-1.11861 (49)	96.71717	1	0.13554	49
29.	50899	-21380.75 (76)	-1.12925 (76)	96.84343	1	0.11875	76
30.	5094	-21327.12 (45)	-1.04148 (45)	97.47475	1	0.12002	45
31.	509424	-21036.10 (34)	-1.02915 (34)	97.09596	1	0.12833	34
32.	5114	-21653.28 (11)	-0.95243 (11)	96.96970	1	0.12695	11
33.	511412	-21653.28 (11)	-0.95243 (11)	96.96970	1	0.12695	11
34.	5117	-21430.55 (18)	-1.23798 (18)	96.71717	1	0.12987	18
35.	511712	-21430.55 (18)	-1.23798 (18)	96.71717	1	0.12987	18
36.	5144	-21519.60 (85)	-1.04971 (85)	96.84343	1	0.13228	85
37.	514412	-21458.37 (44)	-1.02694 (44)	97.47475	1	0.11916	44
38.	5166	-21474.47 (71)	-1.03278 (71)	94.31818	1	0.16768	71
39.	51669	-21474.47 (71)	-1.03278 (71)	94.31818	1	0.16768	71
40.	5175	-21368.34 (98)	-1.02905 (98)	96.21212	1	0.13240	98
41.	5178	-21558.02 (30)	-1.11032 (30)	95.95960	1	0.14879	30
42.	517824	-21558.02 (30)	-1.11032 (30)	95.95960	1	0.14879	30
43.	5198	-21438.29 (83)	-1.04681 (83)	94.31818	1	0.17245	83
44.	51989	-21387.08 (86)	-1.04070 (86)	96.46465	1	0.13016	86
45.	5207	-21858.46 (22)	-1.13155 (22)	93.05556	1	0.17518	22
46.	52076	-21987.08 (65)	-1.19814 (65)	95.20202	1	0.14796	65
47.	5211	-21771.06 (70)	-1.13529 (70)	96.96970	1	0.12539	70
48.	521112	-21707.21 (62)	-1.15547 (62)	95.58081	1	0.14972	62
49.	5228	-21104.94 (78)	-1.14723 (78)	95.83333	1	0.15547	78
50.	522812	-21193.69 (38)	-1.04423 (38)	96.84343	1	0.14025	38
51.	5242	-21113.56 (50)	-1.14397 (50)	97.09596	1	0.14713	50
52.	52423	-21113.56 (50)	-1.14397 (50)	97.09596	1	0.14713	50
53.	5245	-21477.19 (71)	-1.10455 (71)	96.96970	1	0.12454	71
54.	52454	-21695.81 (88)	-1.10778 (88)	95.20202	1	0.15025	88
55.	5261	-21412.51 (03)	-1.01979 (03)	95.83333	1	0.14438	03
56.	526112	-21563.94 (02)	-1.06960 (02)	96.46465	1	0.13282	02
57.	CON_B	-21685.44 (53)	-1.18670 (53)	96.59091	1	0.13810	53
58.	CON_C	-21412.51(03)	-1.01979 (03)	95.83333	1	0.14438	03

Only models, both closed and open, of the target sequence 5079 did not have a GDT scores since the number of Cα atoms in the model did not match that of the template. The amino acid sequence of this sample had a deletion at position 10 as reported in the early section of this chapter (see subsection: 2.4.2 Assignment, Frequencies and Pattern Determination of Non-Synonymous Mutations), hence loss and absence of 1 Cα. GDT only considers pairs of the Cα atoms (Benkert et al., 2009; Li et al., 2011).

The cRMSD is a coordinate-based similarity evaluation that is dependent on the pairwise superposition of the 3D structure of both the template and model. The score is an average of all the RMSD from the alignment of the model and the template. The models had low cRMSD values that were close to zero, hence were of good quality. The cRMSD is directly proportional to RMSD. RMSD inversely proportional to alignment accuracy (Guex et al., 2009). Good models should have low RMSD which means low traceable backbone error (Sánchez & Sali, 1997) and that the model is as good as the experimental structures(s) (Guex et al., 2009). “Closed” models of 3059, 5032, 5045, 5117, 30599, 301812, 503252, 511712 and CON_B sequences had backbones that matched those of their templates, but this does not portray that they are perfect models, since their side chains were not considered by these scores.

The GA341 was implemented in the building script. It assesses the compactness of a model. GA341 score ranges between 0-1, with values towards 1 being good models. Each of the selected models had a GA341 score 1. Even among the 100 models generated for each target sequence, all of their scores were 1. Therefore despite the utility of GA341 in model quality evaluation, it appears that it is not a good criterion for discriminating good models from bad models.

Generally, selection of the best models was complemented by GDT scores. Target sequences without change after drug exposure always had the same reliable models with identical verifications scores being built and predicted (before and after drug exposure). This can be seen in some of the samples, e.g., 5014 (drug-naïve) and 501424 (drug-experienced). All the selected models for each of the target sequences could not be viewed to assess whether or not they had significant problematic regions, so that refinement could be carried out. In this case, we had two assumptions based on confirmed findings: 1) that based on the

template evaluation and high target-template sequence identity, the models constructed were likely to have good qualities, and 2) that the three statistical evaluation approaches (GDT, DOPE and normalized DOPE) selected reliable models from the library of generated models. Hence refinement was omitted.

There is the possibility of erroneous regions (especially in the loops) in the models, and as such energy minimizations with for the whole protease were carried out prior to molecular dynamics experiments. Similarly, docking results especially those of ligand-protein complexes have been shown to have better predicted energy of bound conformation when Arg8 is modified to torsions of freedom (Morris et al., 2009), so the introduction of this freedom eases some of the strain on the model.

2.6 CONCLUSION

At the sequence level, there are nine major amino acid positions differences between subtypes B and C drug-naïve consensuses. These positions have a high propensity for the existence of other amino acids that never occur in the subtype B consensus. Common, rare and uncommon polymorphisms likewise occur in subtype C. Some of the polymorphisms in clade C are associated with PI therapy hence there is need to set effective therapeutic platforms when treating HIV-1 C infected individuals. Intra-variation within clade C in the drug-naïve sequences was minimal. Both major and minor mutations rarely occurred in the drug-naïve patients. Despite the fact that some naturally occurring mutations are associated with PI therapy and have low thresholds, their roles need to be further characterized in such cases. Comparison of part of our work against those of subtype B, for example, from Dirauf and colleague, reveal that the association of drug-linked mutations in HIV-1 clade C drug-exposed subjects is dissimilar from those reported in subtype B. Thus this represents different mechanisms employed by subtype C to confer drug-resistance, hence the need for continued characterization.

Patterns of occurrence of drug-resistance mutations can be better characterized with sophisticated algorithms that still involve frequencies of these mutations. The sequences in this particular study should be scaled-up, and be from a longitudinal study, i.e., samples collected at a significant number of timelines, e.g., at least six spread-out time-points after drug exposure in order to follow the evolution of the virus. Again, these predictions need to

be validated. Knowledge of mutation prediction will enhance optimization of therapy with a view to minimize chances of developing (cross-) resistance. Novel vaccines and drugs for better therapeutic outcome can only be effectuated if drug-associated mutations are well studied (De Oliveira et al., 2005). Several mutations (which usually occur cumulatively) increase susceptibility of one or more ARVs. Even though V82A, I54V, M46I, L33F and F53L mutations emerged due to selection pressure from RTV and LPV/r therapy, these mutations exhibit cross-resistance to other PIs.

For the homology modelling, several programs were used in assessing model quality, and the best model in each case was taken forward to the next stage of the study.

CHAPTER THREE

3. INTRODUCTION

3.1 Scope of HIV-1 Protease Structure, Docking and Molecular Dynamics

Despite the vast information in various databases such as the Stanford HIVdb concerning drug resistances, the mechanisms driving these resistances are still understudied (Dirauf et al., 2010). Few studies have been directed towards finding the impact of HIV clade variation in drug-linked mutations (Ali et al., 2010). This section describes the use of atomic simulations in the investigation of drug susceptibilities in the infant cohort. Structural changes described here include fluctuations of active site surface area and volume. Molecular dynamics and docking studies are significantly not well exploited (Luo & Chan, 2012; Hetényi & Spoel, 2002) but they have been used define changes in drug susceptibilities by implementing thermodynamic computation of inhibition constants and binding energies. Fingerprinting profiles are used to analyze docking results. The utility of molecular dynamics simulation is in the observation of conformational transformations of the flaps both in the absence and presence of drug-resistant mutations. It is important to note that part of the results in Chapter II will be referred to where relevant.

3.2 HIV-1 Protease Structure

HIV protease appears like the head of a cat or bulldog (Perryman & Lin, 2004). It is a symmetrical (Li et al., 2011; Brik & Wong, 2003) homodimer of 99 amino acids per chain (Mao, 2011; Ode et al., 2007) that has a catalytic triad called the DGT motif (Brik & Wong, 2003; Kear et al., 2011) and two molecular flaps (Hornak et at., 2006a; Yu et al. 2011) that guide entry of substrate, or of ligand to the active site (Freedberg et al., 2002) to form a ligand-receptor complex (Luo & Chan, 2012). The chains are identified as chain A and chain B (Heal et al., 2012). The symmetry is lost upon ligand binding. There are various conventions describing the HIV protease topology (Perryman & Lin, 2004). According to the Cantilever-Fulcrum model based on flap opening-closure mechanism, the HIV protease can be divided into four regions namely: the fulcrum (10 - 23), the flap elbow (35 - 42), the flap tip (49 - 52), and the cantilever (59 - 78) (Kandathil et al., 2009). Based on substrate binding, protease has been found to have substrate binding subsites (Prabu-Jeyabalan et al., 2002).

The other convention is independent of the flap flexibility. The protease has whiskers (1 - 5 and 95 - 99, which form the dimerization interface), nose (6 - 10), cheek turn (11 - 22), eye (23 - 30), ear flap (35 - 42), flap (43 - 58), cheek sheet (59 - 75), wall turn (79 - 84, which is part of the active site's wall), and helix (86 - 90) (Perryman & Lin, 2004).

Mutations in HIV protease are known to affect the architecture of the active site, binding kinetics and replication (van Maarseveen et al., 2012). The architecture of receptor site is useful in the binding of ligands (Dirauf et al., 2010). There are many algorithms for prediction/detection of binding sites in proteins, but no accurate approach exists. Programs for binding site detection include Q-SiteFinder, Pocket Finder, CASTp, eF-site (Laurie & Jackson, 2006) and Pocket Program (Ode et al., 2007) among the many examples.

CASTp (Computed Atlas of Surface Topography of proteins) is an online program (Dundas et al., 2006) that is used to search for and estimate the surface area and volume of the pockets (largest binding sites) in proteins. It measures the Richard's surface (SA, solvent accessible surface) and Connolly's surface (MS, molecular surface) (Purohit et al., 2008). MS is made up of two surfaces; van der Waals surface (contact surface) and re-entrant surface (interior surface) (Bhat & Purisima, 2006). Likewise, CASTp offers information regarding the void regions (solvent inaccessible regions that result from the removal of HET groups; important residues that define the functional sites and size of the cavity opening (Binkowski et al., 2003) that allow molecules use to access the binding site (Purohit et al., 2008). These indicators are vital in structure-based drug design, characterization of MDR, and docking studies as they minimize the search space in the advent of binding (Laurie & Jackson, 2006). Even with these accounts, there is inadequate data about the structure of the binding sites (Logsdon et al., 2004).

3.3 *In Silico* Molecular Docking and Fingerprinting

Docking has been used to find the best binding mode between a ligand and a receptor (Hetényi & Spoel, 2002). When a ligand complexes with a HIV protease receptor, a conformational change occurs in the receptor (Luo & Chan, 2012). Docking experiments are exploited in structural biology and pharmaceutics (Huey et al., 2007). Computational time and accuracy have been effectuated via better molecular representation, energy calculation and conformational search strategies (Morris et al., 2009).

AutoDock computes simulated docking (Goodsell et al., 1996) that estimates the binding energy using the Lamarckian Genetic Algorithm (LGA) (Umamaheswari et al., 2012), which is a search algorithm that recognizes a good solution from the vast possibilities of protein conformations due to virtually incalculable protein flexibility and it is based on energy function (Morris et al., 1998). It searches for the best conformers of a given ligand (Gowthaman et al., 2008), i.e., both the correct position/orientation and conformation (Hetényi & Spoel, 2002). AutoDock uses AutoTors to set the ligand torsions, and the energy cost for placing a probe atom at a particular x, y and z coordinate of a grid centered on the active site are calculated by AutoGrid (Goodsell et al., 1996). This defines the search space in the receptor and allows for rapid energy evaluations of the bound conformation (Morris et al., 2009) upon submission to the AutoDock (Goodsell et al., 1996).

Ligand-receptor association involves interactions (Chang et al., 2007; Toor et al., 2011) that can be visualized via fingerprints (2D maps depicting the surface interactions between the ligand atoms and the receptor residues). Fingerprinting profiles between the ligand substructures and protein domains can be used to unravel uncharacterized interdependency between the two molecules and can expound more on the topic of specificity by proteins. These can be piped to drug design and development (Luo & Chan, 2012; Kandathil et al., 2009).

3.4 Molecular Dynamics Simulations

Molecular dynamics is likely to attract a greater application in pharmaceutical engineering and in determining receptor binding sites. This is due to its ability to capture and understand “lightning-speed” occurring events that cannot be captured experimentally (Durrant & McCammon, 2011). It has enabled the upgrading of kinetic processes, ligand-receptor associations and dissociations that have been employed to design both FDA approved and non-FDA approved regimens (Li et al., 2011). Unlike docking studies, molecular dynamics is also able to account for conformational changes in bioprocesses (Durrant & McCammon, 2011). It has revealed the relevance of flexibility (Yu et al., 2011) in ligand binding, unlike in docking where only the ligand locus and energy associated to it are described (Durrant & McCammon, 2011).

The “pebble game” algorithm uses a set of rules to match the torsions and constraints as a result of bonding (Heal et al., 2011, 2012) to determine rigid and flexible domains in a protein (Heal et al., 2011). It provides a profile of protein containing underconstrained, isostatic, and overconstrained regions (Heal et al., 2012). The flap tips exhibit flexibility in unbound form (Heal et al., 2012; Li et al., 2011). Glycine residues are the ones endowing it with flexibility. Molecular dynamics studies on flap tip “curling” based on the clustering of hydrophobic amino acid residues have specified the key role flap conformation in the protease function (Li et al., 2011) where it regulates binding. This concurs with the NMR findings (Perryman & Lin, 2004). Flaps allow the protease to acquire different conformations (Li et al., 2011): open, semi-closed and closed. Ligands have been shown to bind best to the open conformation (Chang et al., 2007). Once binding occurs, the flaps collapse close to the catalytic triad (Hornak et al., 2006a). Inhibitors work by hindering this flexibility. Inhibitors such as TPV render the active site stiff and therefore the flaps cannot close (Heal et al., 2012). HIV protease activity is as a result of high selectivity, efficient catalytic (Brik & Wong, 2003) and flap dynamics (Freedberg et al., 2002; Hornak et al., 2006a).

Hydrogen bonds at the flaps-inhibitor interface aid in protease catalysis. This network of bonds is due to w301 (Yu et al., 2011) that stabilizes inhibitors (Li et al., 2011) and substrates by ensuring that they attain an apt conformation (Yu et al., 2011). Local dynamics have revealed that deviation of ligand from the active site emerges due to mutations in the protease that cause simultaneous or stepwise loss of the hydrogen network at the flap tips and active site. This deviation might lead to the reduction of loading force to zero, hence escape of the ligand from the active site (Li et al., 2011).

3.5 METHODOLOGY

3.5.1 Architecture Variation: Calculation of Volume and Surface Area of Binding Cavity

CASTp suite (Table 3-1) that is constructed under the Delaunay Triangulation (Laurie & Jackson, 2006) and Alpha Shape theory (Purohit et al., 2008) was used to assess deviations in the volume and surface area of binding cleft of the proteases. The best models were uploaded in PDB format and submitted (Binkowski et al., 2003). The 1.4 Å default value was used as the radius of the water probe (Bhat & Purisima, 2006). The pocket architecture was

visualized the Mage (11) Java applet, which permits manipulation (e.g., rotate, translate, or zoom) the 3D structure of the protein. The results were e-mailed as attachments (Binkowski et al., 2003). The HET groups are not included in the calculation (Dundas et al., 2006).

Table 3-1: The online program used to evaluate protease internal architecture.

Name of the online tool	URL
CASTp	http://cast.engr.uic.edu

3.5.2 Construction of a Series of HIV-1 Protease Inhibitors

RU-synthesized ligands were constructed and edited by addition of hydrogen atoms at the selected relevant positions using the Accelrys Discovery Studio Visualizer 3.1. Using a fast, Dreiding-like force field, the geometries of these models were cleaned for optimization using the same program. FDA approved ligands were searched for and downloaded from the ZINC database (http://zinc.docking.org/) as mol2 files (Irwin & Shoichet, 2006) and converted to pdb using OpenBabel.

3.5.3 *In Silico* Molecular Docking

The AD4.2 Release v4.2.3 (http://autodock.scripps.edu) was exploited. The ADT (http://mgltools.scripps.edu/downloads) was used for the stages in the docking process starting from coordinate preparation right through to the analysis of results through clustering performance (Morris et al., 2009). Docking using 1HXB and SQV was carried out to validate the docking prior to the docking the two sets of ligands: 9 FDA approved and 14 RU-synthesized protease inhibitors, to the proteases both in the closed and open conformations. Reproducibility was evaluated using the Accelrys Discovery Studio Visualizer 3.1.

Preparation of the protein and ligand coordinate files entailed the creation of the PDBQT files with special atom type (Umamaheswari et al., 2012) and Gasteigner charges. Arg8 and Asp25 residues in each chain were set flexible, while the rest of the protein residues were left rigid (Morris et al., 2009; Purohit et al., 2008). Both the grid parameter and docking parameter files were created in that order, followed by analysis of the results (Figure 3-1).

Docking parameters were set: the population size to 150, followed by assignment of 0.02 mutation rate, 0.8 crossover and 100 LGA runs per 150 population size, 4500000 and

450000 energy evaluations (Magalhães et al., 2004) for FDA approved and RU-synthesized inhibitors respectively, and 27000 generations before LGA run termination (Archana et al., 2010). Other parameters were left as default (Hetényi & Spoel, 2002). Our docking was automated by use of custom scripts in a way similar to batch docking (Morris et al., 2009).

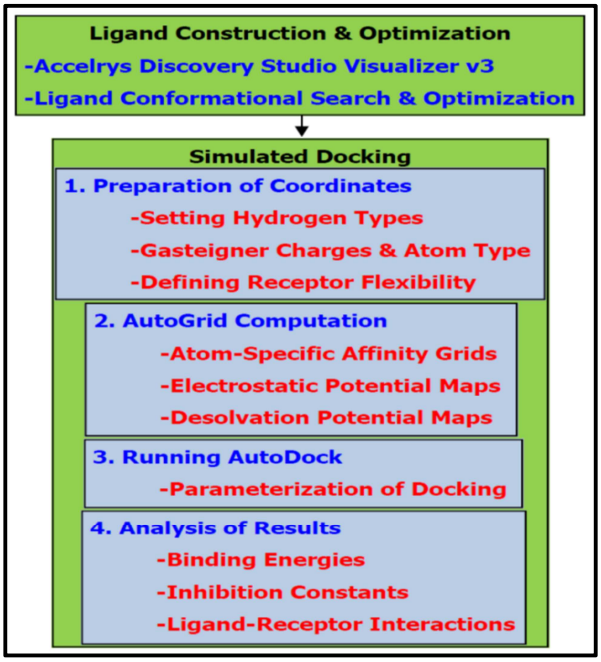


Figure 3-1: Summary of the ligand construction and optimization process and automated docking process.

3.5.4 Molecular Dynamics Simulations

Molecular dynamics v35b2 that exploits both the CHARMM22 force field and TIP3 water model were used (Zhu et al., 2003). HIS residues were changed appropriately to the applicable protonation state, the CHARMM22 force field was used in the treatment of the protein. The ligand was treated using the CGenff force field. Standard scripts (http://www.charmmtutorial.org/index.php/Full_example; www.charmming.org) were used with little modification to set up the system for dynamics. To mimic the *in vivo* aqueous environment, firstly periodic boundary conditions were set (Dirauf et al., 2010) at 60 x 60 x 60 $Å^3$ for the complexed protease based on the shape of ligand-receptor complex from gas-phase. Water molecules were added, while those within 1.8 Å of the protein atoms were deleted, leaving each system fully solvated. Chloride ions (or potassium ions) were used as counter ions (Perryman & Lin, 2004) to neutralize any charges (Li et al., 2011) but at a 0.15M concentration in the protein-ligand complex (Piana et al., 2002). To compute the long-range Coulombic (electrostatic) interactions, the Particle Mesh Ewald (PME) summation algorithm

(Piana et al., 2002) was used. The steepest descent algorithm was used to minimize the structure (Perryman & Lin, 2004; Dirauf et al., 2010; Li et al., 2011) throughout the 10000 steps. 1 fs integration time step of simulation, a 14.0 Å cut-off for nonbonded contacts that were updated in a cycle of 10 steps were set (Li et al., 2011). The temperature of each system was steadily heated to 310 K over 100000 steps and checked periodically for deviations. In case of deviations beyond $\pm$ 5 K then the velocities were scaled up using Berendsen algorithm (Piana et al., 2002; Perryman & Lin, 2004; Li et al., 2011) (and randomly selected from a Gaussian distribution). The SHAKE algorithm was used to fix and constrain any bond where hydrogen-heavy atoms participated (Perryman & Lin, 2004; Dirauf et al., 2010), by setting them to their default values so that the high-frequency simulation of hydrogen-heavy atom bond would take reasonable time. Equilibration was done for 200000 steps, 0.001ps time step. The production run was simulated at 400000 steps by the constant NVE (particle, volume and energy) model (http://www.ch.embnet.org/MD_tutorial/pages/CHARMM.Exercise2.html) also known as isotropic NPT model (Zhu et al., 2003) (isothermal-isobaric conditions) (Perryman & Lin, 2004). The VMD (Visual Molecular Dynamics) v1.9.1 and CHARMM Scripts (Supplementary Data/Chapter III/AllScripts/MolecularDynamics) were used for imaging, structural and energy analyses of trajectory data.

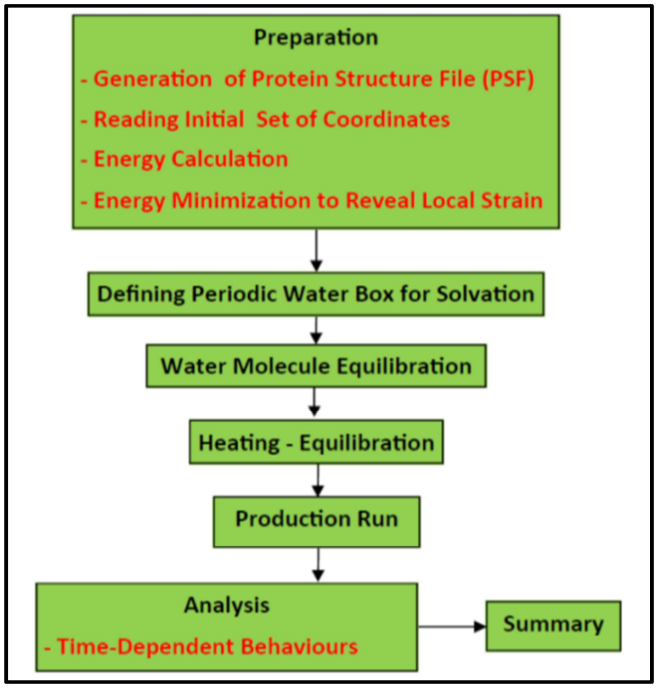


Figure 3-2: Outline of the molecular dynamics procedure.

3.6 RESULTS AND DISCUSSION

3.6.1 Architecture Variation: Calculation of Volume and Surface Area of the Binding Cavity

Geometry is crucial in both structure and function of proteins (Bhat & Purisima, 2006). The study aimed at evaluating the topological changes exerted on the HIV-1 subtype C protease enzyme by mutations. The short-term objective was to relate the evolution of geometry to the viral fitness using the molecular docking and dynamics studies. The long-term objective considers the implications of the structural dynamics of the binding cavity of the mutants to drug engineering in order to improve and or build inhibitors, probably third-generation, with high genetic barrier to resistance that are able to treat patients failing medications.

Data from the active site structure (Dirauf et al., 2010) has enabled most FDA approved PIs to be engineered based on the geometric complementarity (Ohtaka & Freire, 2005) with a view to dock to the closed conformation of the protease enzyme (Martin et al., 2005). Despite the FDA approved PIs in circulation, continued emergence of mutations in the protease, especially MDR-mutations, have been shown to cause structural changes in protease (van Maarseveen et al., 2012; Martin et al., 2005) thereby disrupting the lock and key binding phenomenon. As a result, the rigid drug becomes incapable of acquiring new conformation that fits into the transformed cavity of the enzyme (Ohtaka & Freire, 2005).

CASTp server was used for structural analysis since it has a graphical user interface, flexible interactive visualizations and is up-to-date (Binkowski et al., 2003). Its basis is on the weighted Delaunay Triangulation and Alpha Shape theory. Using a virtual spherical probe, it scans and quantifies the size of the pocket. The solvent accessible surface area (SASA), computed by the “rolling ball algorithm” coined by Shrake and Rupley, indicates the solvent accessibility of the residues. SASA represents the area from the center of a spherical probe as it rolls over the van de Waals surface of the protein (Purohit et al., 2008).

Table 3-2 shows a portion of the molecular surface area and volume geometric scores from the CASTp server. Data relating to the solvent accessible area and residues found in the mouth of the cavity are not included here. Both subtypes B and C were included as references in order to better characterize the contribution of the mutations in HIV-1 C protease receptor from both drug-naïve and drug-failing infants.

Table 3-2: CASTp results for identification and measurement surface area and volume in the HIV-1 protease cavity.

Sequence ID.	Surface Area ($Å^2$)	Volume ($Å^3$)	Net Change in Volume ($Å^3$) and Effect on Cavity	Predicted Notable Changes in the Pocket of the Protease after Drug-Exposure
3018	746.3	1185.3	+41 Expansion	Displacement of L10 and gain of P9. Continuous extension of the active site floor: 23-32.
301812	774.9	1226.4		
3021	693.1	1107.4	+195.9 Expansion	Gain of F53 and R87 and displacement of T80 (in chain A). Generally these changes were of non-active nature.
302112	739.5	1303.3		
3043	650.0	1036.8	0 None	None.
3043pre	650.0	1036.8		
3051	630.9	1001.5	+33.7 +212.8 Expansion	Displacement of R87. L10F and N83 becomes part of the pocket. Extension of the binding cavity.
305112	699.9	1035.2		
305152	748.0	1214.3		
3059	610.5	991.8	0 None	None.
30599	610.5	991.8		
5014	710.9	1158.2	0 None	None.
501424	710.9	1158.2		
5032	772.4	1195.0	-300.6 Contraction	L10 absent and the binding cavity extends in size.
503252	573.7	894.4		
5045	675.5	1086.7	+103.5 Expansion	Gain of L10 and extension of the active site floor from 23-32 to 22-32.
50459	724.6	1190.2		
5046	706.5	1106.4	0 None	None.
504612	706.5	1106.4		
5074	588.6	912.6	0 None	None.
507436	588.6	912.6		
5079	698.9	1145.4	-18.5 Contraction	Active site floor reduces.
50796	691.2	1126.9		
5080	627.8	1037.8	0 None	Displacement of L10 from the binding cavity
508012	627.8	1037.8		
5086	768.9	1181.5	-94.4 Contraction	Displacement of T26 and R87. Overall loss in the number of residues in the pocket.
508612	642.1	1087.1		
5089	647.8	1037.4	+39 Expansion	Displacement of L10.
50899	684.2	1076.4		
5094	665.4	1086.0	-38.7 Contraction	Displacement of T26 in chain A and shifting of R87.
509424	621.3	1047.3		
5114	626.0	1021.8	0 None	None.
511412	626.0	1021.8		
5117	622.5	1021.0	0	None.

511712	622.5	1021.0	None	
5144	739.8	1141.2	+60.3	Displacement of T26 and 87R, and gain of L10 in chain A.
514412	760.7	1201.5	Expansion	
5166	547.2	867.4	0	None.
51669	547.2	867.4	None	
5175	595.5	949.0	*	*
5178	758.0	1189.4	0	None.
517824	758.0	1189.4	None	
5198	720.6	1109.0	+43.2	Gain of L10 and loss of T26 and R87 in chain B.
51989	701.4	1152.2	Expansion	
5207	644.6	1060.7	+339.3	Gain of t26, F53 and R87. An extension of the floor of the active site cavity.
52076	793.1	1400.0	Expansion	
5211	616.8	982.6	-29.1	Displacement of R8 in chain A and gain of L10 and T80 in chain B.
521112	610.5	953.5	Contraction	
5228	628.4	1038.1	+99.5	Gain of L10 and displacement of T80 in chain B.
522812	677.3	1137.6	Expansion	
5242	681.7	1092.1	0	None.
52423	681.7	1092.1	None	
5245	645.8	1013.9	-47.9	Displacement of R8 of chain B as a binding residue.
52454	593.9	966.0	Contraction	
5261	708.7	1154.0	-240.3	Displacements of R8 and L10, and T24 and R87, of chains A and B respectively.
526112	550.3	913.7	Contraction	
CON_B	676.1	1088.8	*	*
CON_C	707.6	1065.3	*	*

*****Indicates HIV-1 protease structures whose architectural variation could not be estimated due to lack of corresponding structure at a different timeline.

Subtype B was used as a reference for consensus C that was subsequently used as a reference for the other subtype C. Sequences before drug exposure were used as references for the HIV variants. Protease structures that lacked sequence changes after drug-exposure were compared against subtype C structure in order to isolate those individual or group of mutations associated with changes in the protease pocket.

The above data from the *in silico* architectural studies showed that both drug-linked and non-drug-linked mutations affect the structure of the HIV-1 protease enzyme. The effect can either be an expansion or a contraction, unless no mutations occur after drug exposure or by chance (of which we propose that this is highly unlikely) and that the compensatory mutations lead to a net change of zero in the active site structure.

In this context, the term “displacement” has been used to refer to the scenario where a residue or a group of residues is/are dislodged and no longer become part of the cavity. The term “gain” designates when a residue or a group of residues change(s) its/their structural orientation and become constituents of the binding cavity.

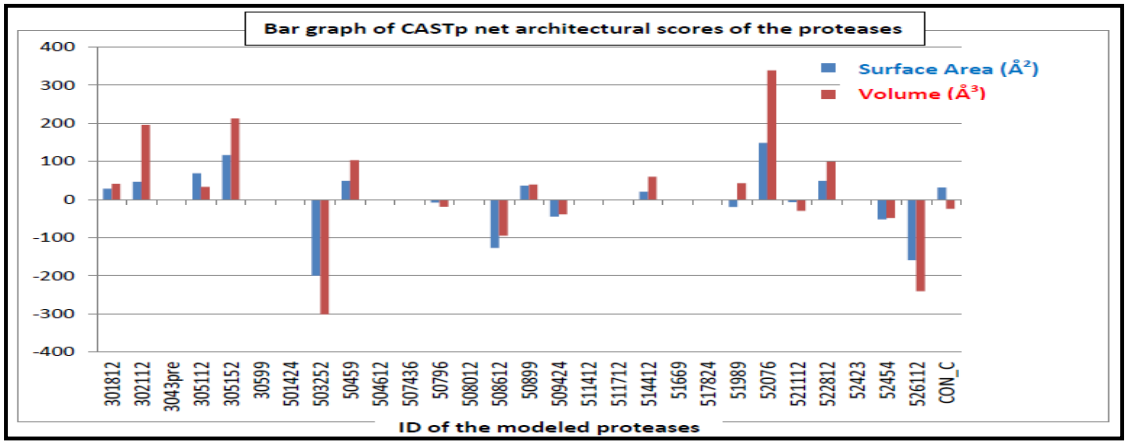


Figure 3-3: CASTp bar graph of the architectural variations of the HIV-1 protease active site volume and molecular surface area. Each point on the plot represents the change in binding cavity size going from drug-naïve to drug experienced models. Positive and negative values indicate active site expansion and contraction respectively after drug-exposure.

The plot in Figure 3-3 summarizes the data in Table 3-2 except that the residue changes are not captured here. The y-axis represents both the statistics for the volume and surface area (either constant, increase or decrease) while the x-axis denotes the protease identification name, each after drug-exposure (CON_C was compared against CON_B). Compared to subtype B, HIV-1 clade C protease has a slightly contracted binding cavity and an increased SA that relates to the hydrophobicity (solvent-excluded surface where binding interactions occur). MS is preferred over SA because it better depicts topology (Bhat & Purisima, 2006). The nine amino acid variations between clades B and C are the ones causing this constricted phenomenon. M36I is known to increase the flexibility and cavity, but participations of all the other polymorphisms lead to a contracted cavity. Clade C is more catalytically active than clade B (Bessong, 2008) based on enzyme kinetics (Ali et al., 2010), thus besides other factors, there might be a correlation between the protease architecture and its substrate specificity. Mutations distort the native orientation of the HIV-1 protease residues (Martin et al., 2005) thereby lessening the active site volume, as illuminated in the pocket contraction model, thus large ligands like RTV cannot effectively bind (Dirauf et al., 2010).

10 (35.7%) of the infants had HIV-1 clade C protease structures that did not show any variation of the studied parameters since the viruses that they harboured did not acquire

any new mutations after drug-exposure, hence the protease structures before and after treatment were unequivocally identical. In Figure 3-3, points within the zero-line signify scenarios where the protease topology never changed after drug-exposure. Comparison of this class of proteases with subtype C consensus indicated an increase or decrease of their pockets (relative to consensus C). Visualization using the Mage (11) Java applet showed that this change is influenced by shifts in residues between the two protease chains, or displacement or gain of residues into the pocket. Mutations may also contribute to an enlarged cavity, a theory known as the active site expansion model (Martin et al., 2005), which concurs with the substrate envelope hypothesis (Ali et al., 2010) and binding kinetics from docking (Martin et al., 2005). 9 (32.1%) of the infants had proteases with expanded active sites. The expansion model can be monitored by Cα tracing and superimposition (Logsdon et al., 2004; Martin et al., 2005). Here we used the volume as a measuring criterion. These data were correlated to the mutations previously reported in subsection 2.5.2 Assignment, Frequencies and Pattern Determination of Non-Synonymous Mutations.

Mechanisms contributing to the expanded pocket include substitution of shorter side chained amino acids for longer side chained (Logsdon et al., 2004) and or reshaping of the backbone. For example L90M destabilizes protease geometry (Kandathil et al., 2009) by restructuring the amino acid residues between 23-32 and 86-88 of the protease enzyme (Martin et al., 2005). Sample 50459 had K45R, V82A and L90M mutations. The amino acid at position 82 forms part of the 80-85 protease loop (Dirauf et al., 2010). The change of valine to alanine at this positions leads to an expansion of this outer loop since alanine has a smaller side chain as compared to valine. Changes in the side chains causes rearrangements in the protein structure (Prabu-Jeyabalan et al., 2002). V82A/L90M dual mutation is known to lead to contraction of the pocket (Dirauf et al., 2010) but the 50459 sample had this dual mutation but had an expanded cavity. This is because besides the nine polymorphisms associated with subtype C, it had the K45R rare mutation and E35D conserved mutation, when compared with subtype B. This brings into the limelight that K45R might be strongly augmenting the structural effects of E35D and that subtype C mutations might be affecting protease structure in a different fashion as compared to subtype B. When mutations occur together, it is actually difficult to study the effects of a single mutation due to the possibility of synergism (Dirauf et al., 2010), therefore we implicated the combined role of these

subtype C polymorphisms to the evolution of the binding pocket. One of the strategies to study the impact of each mutation would be to mutate the wild-type with a single missense substitution (Olsen et al., 1999). Given a pattern of mutation, especially those acting in synergism, more point mutations could be introduced and their impacts characterized.

Some proteases had V82A that is expected to lead to an increased pocket, but due to input of other mutations, the net cleft transformation was in some cases a decrease. Proteases 305112 and 305152 were from the same infant but at different time-points. They had the same major and minor mutations but the volume of the latter was large as compared to that of the former structure. This inconsistency is caused by a naturally occurring polymorphism in the latter macromolecule, E35D, causing the pocket-increase. E35D occurs in the flap elbow and alters the conformations of the binding site residues (Kandathil et al., 2009). CASTp is limited to studying the changes in residue-to-residue interactions that later lead to the noticeable architectural adjustments.

Mutations can either lead to (Kandathil et al., 2009) loss or acquisition of forces involved in residue interactions (Logsdon et al., 2004). Examples of such forces include the van der Waals and hydrogen bonds (Prabu-Jeyabalan et al., 2002). These can lead to cross-resistance, as in the case of L90M (Martin et al., 2005). Sample 52076 had a protracted pocket. It had V82A and L10I mutations. Mutations at positions 82 (and 84) extend the S_1 subsites, and lead to loss of affinity. L10I dislodges residues near the position 7, thereby shifting them away from Asp25 (Logsdon et al., 2004). Because of differential residue interaction in the two chains (Kandathil et al., 2009), some side chains of the residues in each of the two protease chains can acquire different conformations (Prabu-Jeyabalan et al., 2002). Therefore shifts of residues of different chains in and out of the pocket are likely to asymmetrically affect the shape and size of protease. Generally, we observe that the size of the MS is often directly related to the volume of the cavity.

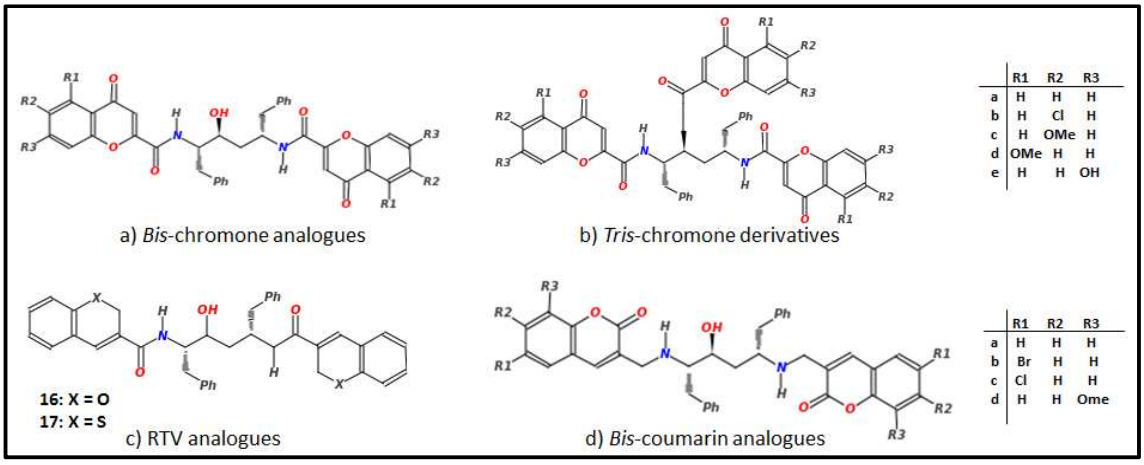
CASTp also generated data of other potential pockets, but these have been ignored because their small sizes. Current algorithms use an invariant probe that might lead to unrealistic pockets (Bhat & Purisima, 2006). The actual binding cleft is supposed to have the largest volume of all. Furthermore, the other possible pockets were disregarded based on the available knowledge about the actual pocket. The Mage imager enabled visualization of the

pocket to which the ligands were to be docked to, thus selection of the actual pocket was much easier. CASTp also predicted the vital residues in the HIV-1 protease. These included residues situated upstream and downstream of the active site floor. They comprise R8 that forms direct contact with substrate (Prabu-Jeyabalan et al., 2002), active site residues (20s-32), residues in 47-50 that comprise part of the flap residues, and residues in loop 80-84 (Ali et al., 2010). Dirauf and colleagues had reported the 80s loop to span until the residue 85.

In summary, mutations affect the size of side chains, orientations of residues and forces involved in residue-residue interactions, which exclusively adjust the pocket volume. From the correlation between mutations and the pocket structure, we integrate both the cavity expansion and contraction phenomena in the subsequent subchapters to explicate the differential patterns in drug predispositions of these viral variants. The consequences of the variations in the architecture of the pockets as a result of naturally occurring and or drug-induced mutations may better be characterized in "real-time" using molecular dynamics.

3.6.2 Construction of a Series of HIV-1 Protease Inhibitors

The Rhodes University Organic Group has for more than a decade used the Baylis-Hillman reaction to synthesize chemical compounds with potential therapeutic expenditure. Among the compounds include a set of RTV analogues that are hydroxyethylene dipeptide isosteres having the chromone (1,4-benzopyrone), coumarin (1-benzopyran-2-one) and chromene functionalities. They mimic the transition-sate and bind to the protease active site (Kaye et al., 2004). The diagrams below show these sets of inhibitors.



	R1	R2	R3
a	H	H	H
b	H	Cl	H
c	H	OMe	H
d	OMe	H	H
e	H	H	OH

	R1	R2	R3
a	H	H	H
b	Br	H	H
c	Cl	H	H
d	H	H	Ome

Figure 3-4: RU-synthesized protease inhibitors drawn using the ZINC[12] molecular drawing interface. Substitution moieties of both *bis*- and *tris*-chromone analogues are shown in the top right. RTV analogues may be replaced with either O or S at the variable groups. RTV analogues and *Bis*-coumarin variable groups are adjacent to them.

The RU-synthesized protease inhibitors (Figure 3-4) had their geometries cleaned after the 3D construction in the Accelrys Discovery Studio Visualizer 3.1 to ensure correct bond angles and lengths. No further optimization using molecular mechanics or *ab initio* methodology was performed.

The FDA approved protease inhibitors (Figure 3-5 and Figure 3-6) were downloaded from the ZINC database (which is a library of at least 727,842 chemical compounds (both in 2D and 3D forms). There are several other web-based databases containing library of drugs for virtual screening. These include Available Chemicals Directory (ACD) (http://www.mdli.com/), ChemNavigator database (http://www.chemnavigator.com/), ChemBank Project (http://chembank.med.harvard.edu/), and Ligand.Info database (http://ligand.info/) among others.

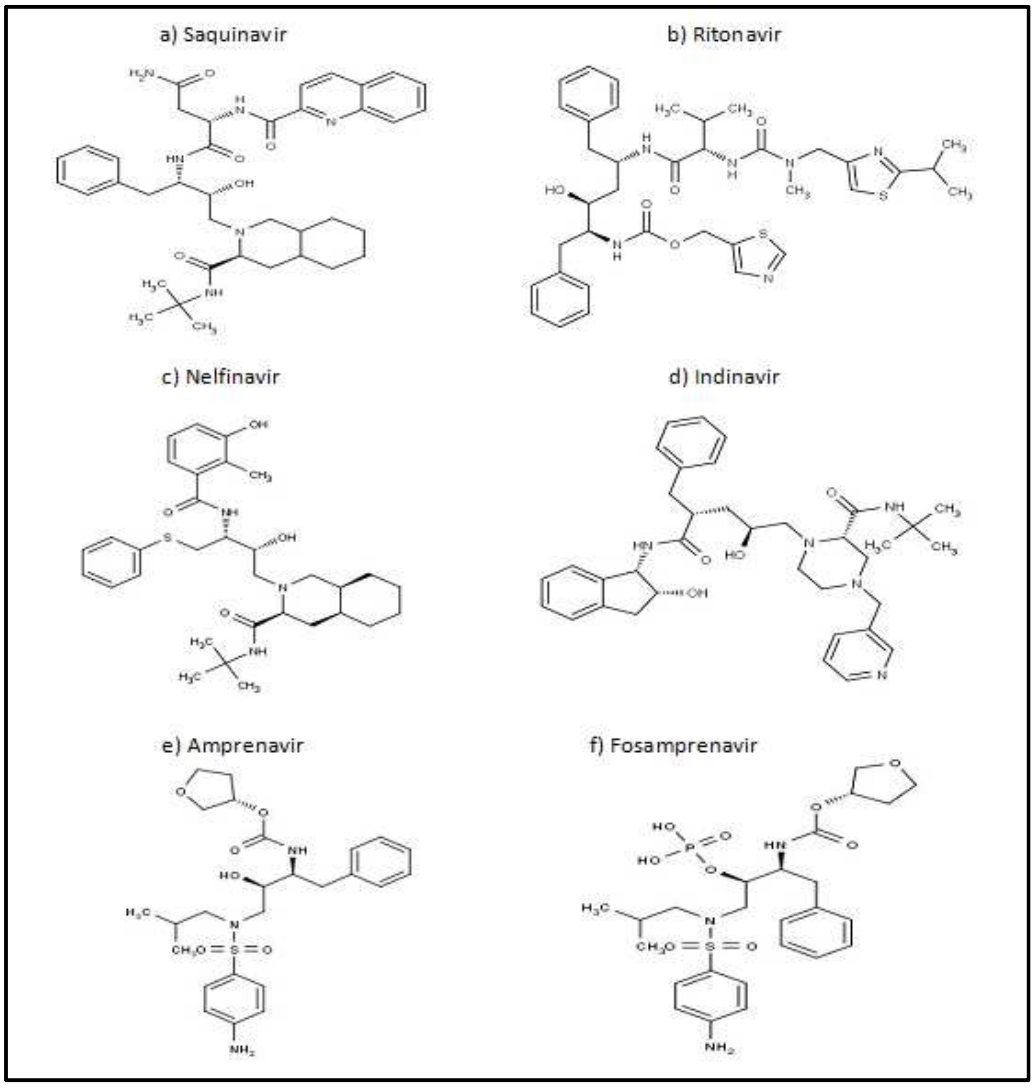


Figure 3-5: Chemical structures of the first generation HIV protease inhibitors, adapted from the drug bank (http://www.drugbank.ca/)

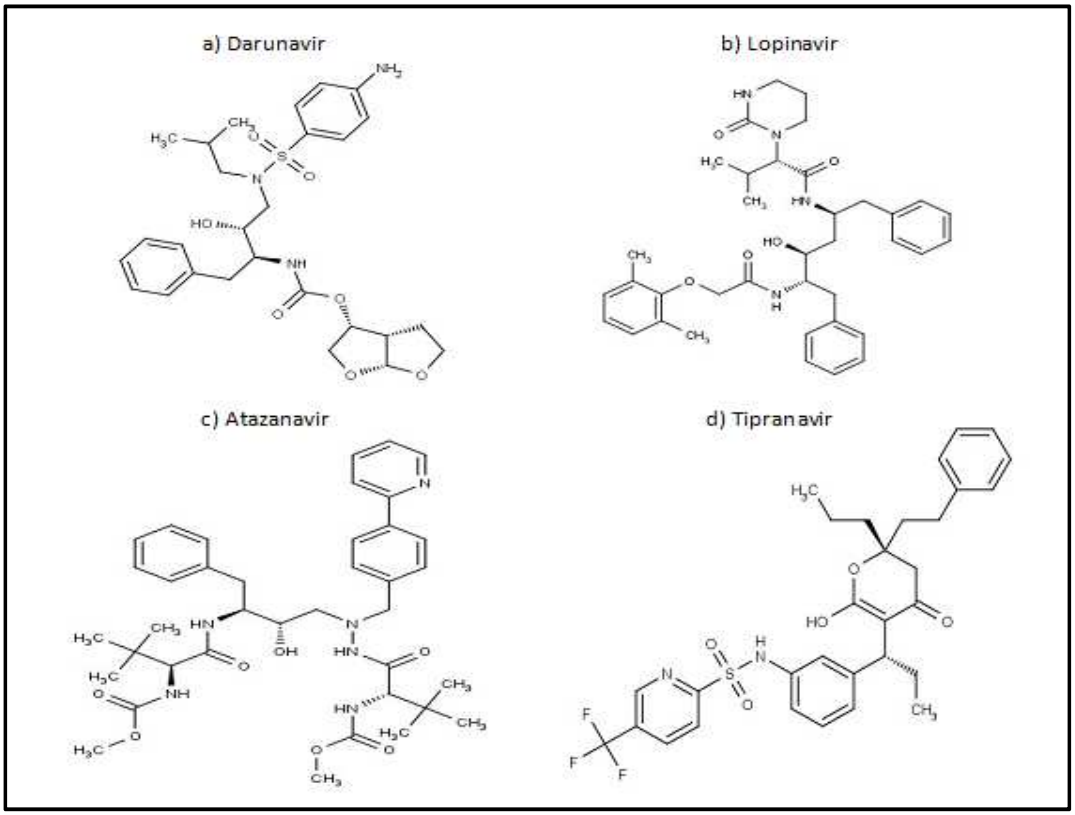


Figure 3-6: Chemical structures of the second generation HIV protease inhibitors, adapted from the drug bank (http://www.drugbank.ca/)

Besides its ready availability and ease of use, the ZINC database was chosen over the other databases because its compounds are expansively postcurated, i.e., they exist in various protonation states, tautomers, stereoisomers and regioisomers, and therefore are ready for docking. Different formats exist for downloading files. These include mol2, 3D SDF, SMILES, and DOCK flexibase (Irwin & Shoichet, 2006) but mol2 was used due to the availability of the mol2-to-pdb OpenBabel interconverter.

An exhaustive conformational search on the initially constructed ligands was neither practical nor absolutely essential prior to successful docking in which a search of conformations takes place. We had performed a pilot study using the VegaZZ v2.4.3 that carries out charge calculation and optimization of the ligand geometry through energy minimization. A systematic search was not possible on systems that presented many torsions, and a random search following 500000 steps per ligand displayed no satisfactory improvement in conformation from that drawn. Thus to circumvent the time cost, the high-quality structures of the FDA inhibitors were selected from the ZINC database. They did not require further modifications.

3.6.3 *In Silico* Molecular Docking: Evaluation of Binding Energies and Interactions

3.6.3.1 Docking Overview and its Validation

During docking, the ligand moves and complexes with the receptor, with the ligand existing at its lowest energy corresponding to the best conformation (Goodsell et al., 1996; Magalhães, et al., 2004). For some receptors, docking is not successful when receptor cavity flexibility is not accounted for (Goodsell et al., 1996), hence this was taken into account in the current study. A particular docking study involving HIV-1 protease receptors has shown that better predicted energies of the bound conformation emanate when Arg8 is modified to have flexible torsions (Morris et al., 2009). In addition, variances in binding residual potentials during docking have also provided the proof that Asp25 (that is usually buried in unbound receptor, but exposed during ligand binding), has minimal flexibility due to its solvation potential (Purohit et al., 2008). Challenges exist in docking mostly if the torsions are engaged (Hetényi & Spoel, 2002). Many torsions increase the computational time and conformational search space, which might lead into a false positive paradox (Morris et al., 2009). To address the problem of computational cost while still maintaining accuracy, strategies previously used include fixing the search grid so that residues outside the search space are kept “frozen” (static) in order to evade their periodic updates (Goodsell et al., 1996). A usual practice is to identify flexible residues and maintain the rest of the protein as rigid during docking. Ideal number of energy evaluations and the conformational search algorithm must be considered (Morris et al., 2009).

LGA is the search algorithm that was used in this study. It is an energy evaluation and optimization method whose principle is based on the natural genetics and natural selection model. It estimates the conformations with the lowest fitness functions (energy values). It starts by building a generation of trial conformations that are selected for fitness (after several energy evaluations), then allowed to perpetuate the filial generation following application of the mutation and crossover operators (Magalhães et al., 2004). The population size of 150 and the total of 100 runs were set based on available evidence as to what is reasonable within AutoDock program. Studies on the factors affecting docking efficiency have revealed that the fraction of the correct binding sites found increases if the population size is set high above the 50 default value, given that the number of energy evaluations and trials are held constant. Beyond 300, it is assumed that the ability to find

the minimum decreases. Similarly, at constant number of evaluations and population size, the number of LGA trials affects the fraction of the correct binding sites attained. The optimal number of is 100 (Hetényi & Spoel, 2002).

For simulated docking, AutoDock searches the conformational space and evaluates conformations using a force field scoring function (Huey et al., 2007) based on AMBER (Toor et al., 2011) that is known to yield accurate results (Goodsell et al., 1996). Specifically, AD4 was utilized. Redocking and cross-docking tests have labelled the AD4 as a more effective method than AD3 due to its provision for selecting and setting the side chain flexibility and protein rigidity, already mentioned as crucial to this study. It uses a semi-empirical energy evaluation approach that integrates intramolecular energy of the bound and unbound forms, specific atom type and charge-based desolvation in its computation (Morris et al., 2009). In AD4, the semi-empirical methodology evaluates the docking energetics in water solvent while the empirical approach computes the solvation effects (Huey et al., 2007).

To ensure confidence in the *in silico* docking, docking was validated prior to the docking of our sets of ligands to the various proteases. Blind docking can be utilized in AutoDock to do this (Laurie & Jackson, 2006). The term “blind” refers to its inability to “see” the pocket but can still discover it (Hetényi & Spoel, 2002). Blind docking has several applications including the ability to predict ligand-receptor complexes (Morris et al., 1998). It requires that both the ligand and receptor are known. It validates docking by assessing correct binding alignment of the ligand to the receptor cavity. Large number of ligands hamper the blind docking computational time (Laurie & Jackson, 2006). In our study, the actual pocket was already defined within the grid. Most FDA approved PIs ware engineered to dock to the closed protease conformation (Martin et al., 2005). Crystallized 1HXB (of closed conformation), to which SQV was bound, was used to check the performance of the AutoDock through reproducibility. The 1TW7 could not be used in validation since its crystallized structure was in the open conformation and did not have any ligand bound to it.

The overall aim of this was to access the degree to which the ligand in virtual docking was in line with the ligand positions in the crystal structure (Goodsell et al., 1996) including orientation (Laurie & Jackson, 2006). Once the *in silico* docking involving 1HXB and modelled SQV was performed, the best SQV with the lowest energy was extracted and redocked to

the crystal structure of protease, 1HXB, and both the ligand position and interactions assessed using the Discovery Studio Visualizer 3.1 (Figure 3-7) for reproducibility.

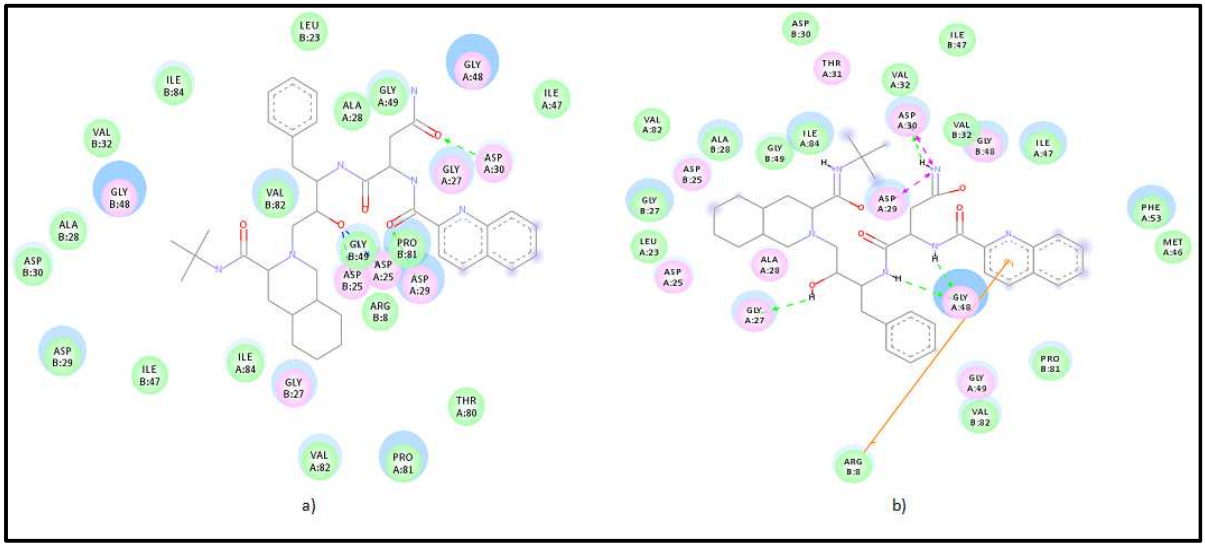


Figure 3-7: Docking validation results: a) shows the interactions of the experimentally crystallized structures (SQV bound to 1HXB), b) indicates the counterpart in simulated docking. Protease residues involved in the interactions with the ligand atoms are shown by coloured discs. Residues with coloured diffuse background indicate the solvent accessible areas.

The global search implemented in the simulated docking enabled the best conformation of the bound ligand in the above validated docking to be established. Although the docked conformation is not identical to the crystal structure, this conformation is not unreasonable due to the computational cost in terms of search space as a result of the torsions present (Morris et al., 2009). If the search space could have been exhausted, then the ligand with the lowest binding energy could have been attained. Energies obtained from the best docked conformation will be an upper bound on the actual best energy possible. Since the position of the active site in protease poses a challenge for docking programs (Magalhães et al., 2004), the correct binding site was defined, explicitly discriminating it from all energetically unfavourable pockets (Hetényi & Spoel, 2002). A view of the Discovery Studio Visualizer 3.1 displays the 2D spatial interactions and contacts between the 1HXB receptor and SQV ligand. The ligand molecule is represented by a skeleton structure whereas discs represent all the receptor residues involved in interactions. 26 1HXB residues were seen in both the ligand-receptor interactions in the experimentally crystallized complex and validated docking. In chains A and B respectively, the crystallized complex had the following residues D25, G27, A28, D29, D30, I47, G48, G49, T80, P81, V82, and I84, and R8', L23', D25', G27', A28', D29', D30', V32', 147', G48', G49', P81', V82', and I84'. These residues endorse

the hydrophobicity nature of the active site (Toor et al., 2011). All these residues were reproduced in the validation with the exception of T80, P81, L23', D29', V82', and I84'. Residues in validated docking that were inexistent in the crystallized complex included T31, M46 and F53. Residues L23, V32 and G48' occurred but were in adjacent chains as when compared to their counterparts in the crystallized complex.

Interactions due to hydrogen, electrostatic and cation-pi bonding are indicated by dashed lines. All these ligand-receptor interactions were within 4 Å distance with the exclusion of pi interactions that can exceed this. In the docking validation, R8' formed a cation-pi interaction with the nitrogen-containing ring of SQV, whereas in the crystallized complex, it was a nonbonded contact. One of the roles of R8/R8' is to occlude access to the active site (Perryman & Lin, 2004). Other nonbonded contacts of the residues with the ligand moieties were apparent. Residues with coloured diffuse backgrounds define the solvent accessible regions, with each region being proportional to the radius of the diffuse disc.

3.6.3.2 Evaluation of FDA-approved Protease Inhibitors

After evaluating the success of the docking validation, the dockings of both the FDA approved and RU-synthesized were also automated using custom scripts in this study. In total, there were 23 ligands and 116 proteases (both closed and open conformations). FVP was eliminated from the docking since its conversion *in vivo*, is comparable to that of APV by removal of a phosphate moiety. It is therefore appropriate to only model the active form of the drug. There were 2667 dockings (excluding the one for the validation) performed during the course of this work. Various stages of the automation included setting search areas, preparing ligands plus flexible residues, combination of rigid protein and flexible parts into a single file, extraction of the best bound ligand conformation and extraction of energies from the docking log files facilitated a smooth flow from input structures to results (Laurie & Jackson, 2006). Thus automated docking eliminated any chance of backlog. All the scripts used for docking any postdocking analyses are in the supplementary information (chapter III, Scripts). The docking scores for both the closed and open protease conformations were extracted (Supplementary Data/Chapter III/Docking/kineticscores_allCLOSED.out and kineticscores_allOPEN.out, respectively). Only the ligands docked to the closed conformation proteases were compared for drug susceptibility responses against the

interpretations from the Stanford HIVdb. The docking scores involving the open conformation will be discussed later.

3.6.3.2.1 Evaluation of Docking: Consensus C Protease in Focus

The docking results for the HIV-1 consensus C and wild-type B were compared in order to establish whether or not there exist disparities in the drug responses. The Stanford HIVdb is tailored to fit drug resistance interpretations for subtype B (Toor et al., 2011). One of its features is that it gives information of the level of drug resistance: levels I to V. Levels II to V reports the extent of the conferred resistance (from potential low-level to high-level resistance). Level I is the susceptibility level (Tang et al., 2012; Zazzi et al., 2004). The Stanford HIVdb indicates that both the clades B and C have the same mutation scoring of zero for all the protease inhibitors. It is surely not true that the polymorphisms in clade C do not perturb the drug susceptibilities. Based on all the docking results, we report the changes in drug susceptibility level. The susceptibility level, this does not necessarily typify that a particular mutant is resistant to a particular inhibitor.

In Table 3-3, indicated are the best energy scores of the subtypes B and C docked to nine protease inhibitors. These energy scores are related to their individual inhibition constants (see kineticscores_allCLOSED.txt). An inhibition constant is inversely proportional to the binding affinity. Hence, a lower binding affinity indicates a greater binding affinity (Wu et al., 2008). The converse is also true. Thus a good drug is one with good binding affinity (Ohtaka & Freire, 2005) besides having commendable pharmacokinetic properties (Kaplan & Hicks, 2005; Wu et al., 2008). Although there is poor characterization of mutations in non-subtype B (Ohtaka & Freire, 2005), we confirm that the nine polymorphisms in clade C (Bessong, 2008; Toor et al., 2011) have impact on the drug responses.

Table 3-3: Minimum estimated binding energy (kcal/mol) of clades B and C docked to protease inhibitors.

	APV	ATV	DRV	IDV	NFV	LPV	RTV	SQV	TPV
Subtype B	-11.84	-13.15	-13.14	-15.02	-16.58	-13.31	-13.58	-14.49	-13.72
Subtype C	-12.72	-13.60	-14.47	-16.99	-17.08	-13.18	-12.83	-15.15	-14.76

ATV, NFV, SQV, APV, DRV, TPV, and IDV, arranged in order of increasing deviation in the minimum estimated binding energies, have better binding energies in consensus C than in clade B. LPV and RTV, also arranged in the same order; depict better binding energies in

clade B than in C. In terms of volume of the binding site, clade C had initially been found to have a contracted site, and a slightly increased MS. This reduced pocket affects the binding of bulky inhibitors like RTV (Dirauf et al., 2010), hence the differences observed. Through the Discovery Studio Visualizer, the ligand fitting in the binding site was analyzed. Some inhibitors fit well while others have poor fitting. In Figure 3-9, LPV has a better fitting in the pocket of clade B than in clade C. TPV, an adaptable inhibitor (Heal et al., 2012) has a more extended form in clade C than in clade B. Inhibitors with extended structures seem to be efficient (Prabu-Jeyabalan et al., 2002; Wu et al., 2008). This could explain its better binding energy and lower inhibition constant in clade C than in clade B.

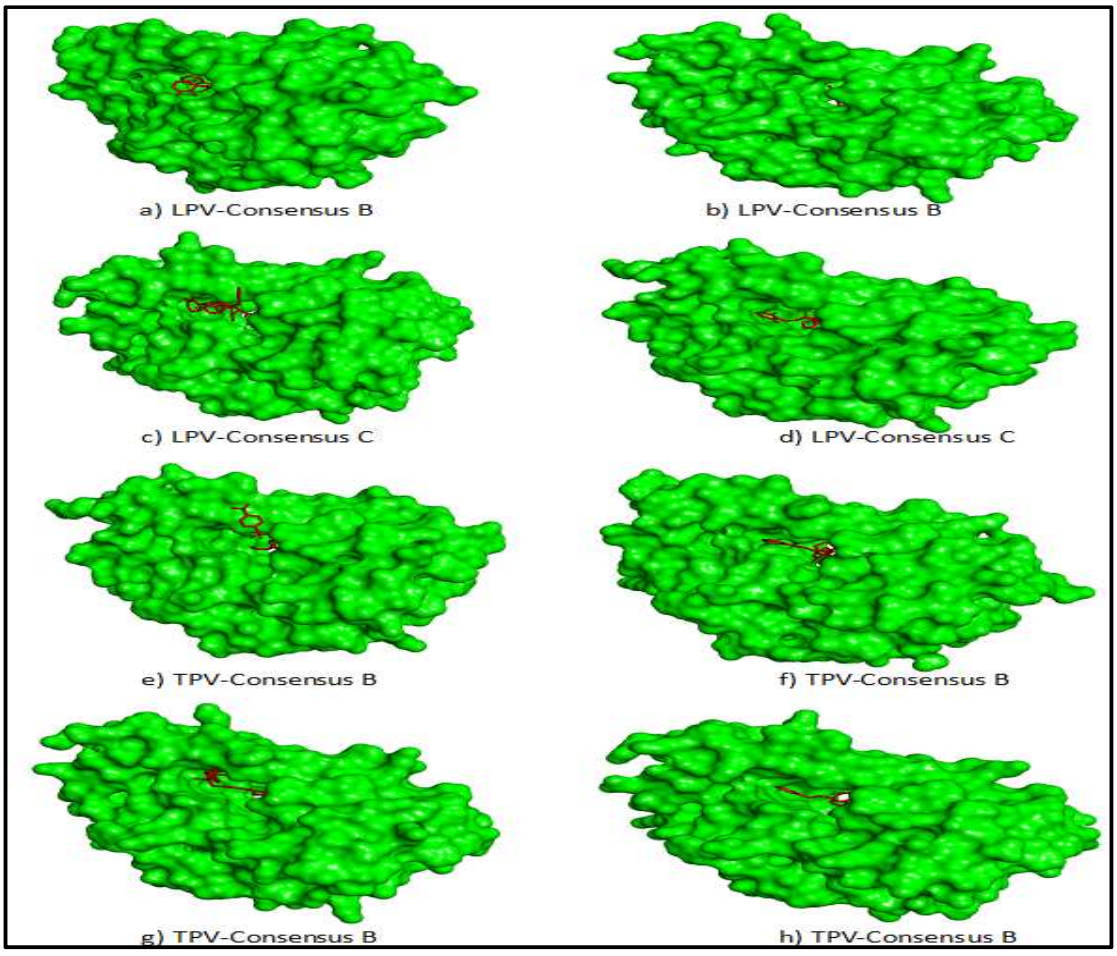


Figure 3-8: Connolly's surface representation depicting LPV and TPV binding fits in consensuses B and C proteases. For each complex, both sides of the protease are shown. Green is the enzyme while red is the ligand.

The same trend is also observed with the RU-synthesized PIs. The binding energies (and inhibition constants) were different in the two proteases. Two factors could account for these. The size of the pocket plays a critical role in the fitting of the inhibitors that vary in

sizes hence good binding energies are obtained when the size of the ligand matches the size of the protein (Ali et al., 2010). Compared to the *bis*-chromone derivatives, the *tris*-chromone derivatives did not fit well in the subtype C pocket due to their sizes (Figure 3-9).

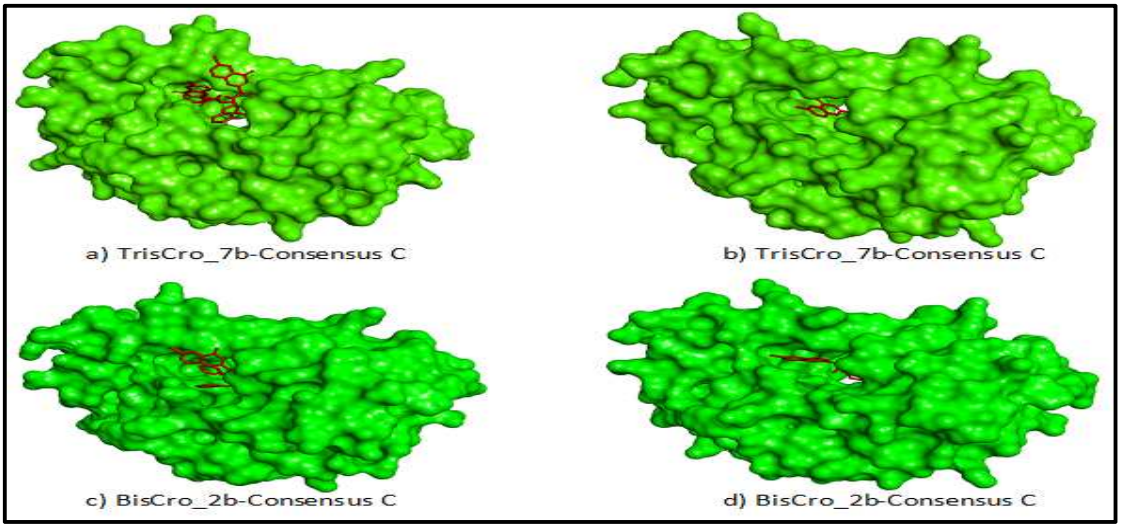


Figure 3-9: Connolly's surface representation depicting TrisCro_7b and BisCro_2b binding fits in consensuses B and C proteases. For each complex, both sides of the protease are shown. Green is the enzyme while red is the ligand.

They are bulkier than *bis*-chromone derivatives. Poor fitting (hence binding energies) (Dirauf et al., 2010) of these inhibitors was observed in clade C consensus model in comparison to clade B. This is still supported by the active site contraction hypothesis. This poses the following question: why is there a difference in binding energies of inhibitors of the same size in the same active site? The structures of these RU-synthesized compounds are shown in Figure 3-4. The RTV analogues, *bis*-chromone and *bis*-coumarin analogues have similar size but different binding energies due to the differences in their functional moieties: chromone, chromene (also sulphur-containing) and coumarin. These functionality groups have different ligand-receptor interaction modes, which contribute to the differences in binding energies and inhibition constants. Behaviour of these RU-synthesized ligands with the different proteases will be discussed later, for now focus is on the FDA approved PIs.

Since we know that differences in HIV subtype are linked to different binding energies and inhibition constants, we used the 2D ligand-receptor interaction profiles of the Discovery Studio Visualizer 3.1 to access the interactions involved in binding of the FDA approved drugs to the clades B and C proteases in order to characterize the associations leading to the various drug responses. Figure 3-10 and Figure 3-11 portray the comparison of interaction profiles in consensuses of clades B and clade C.

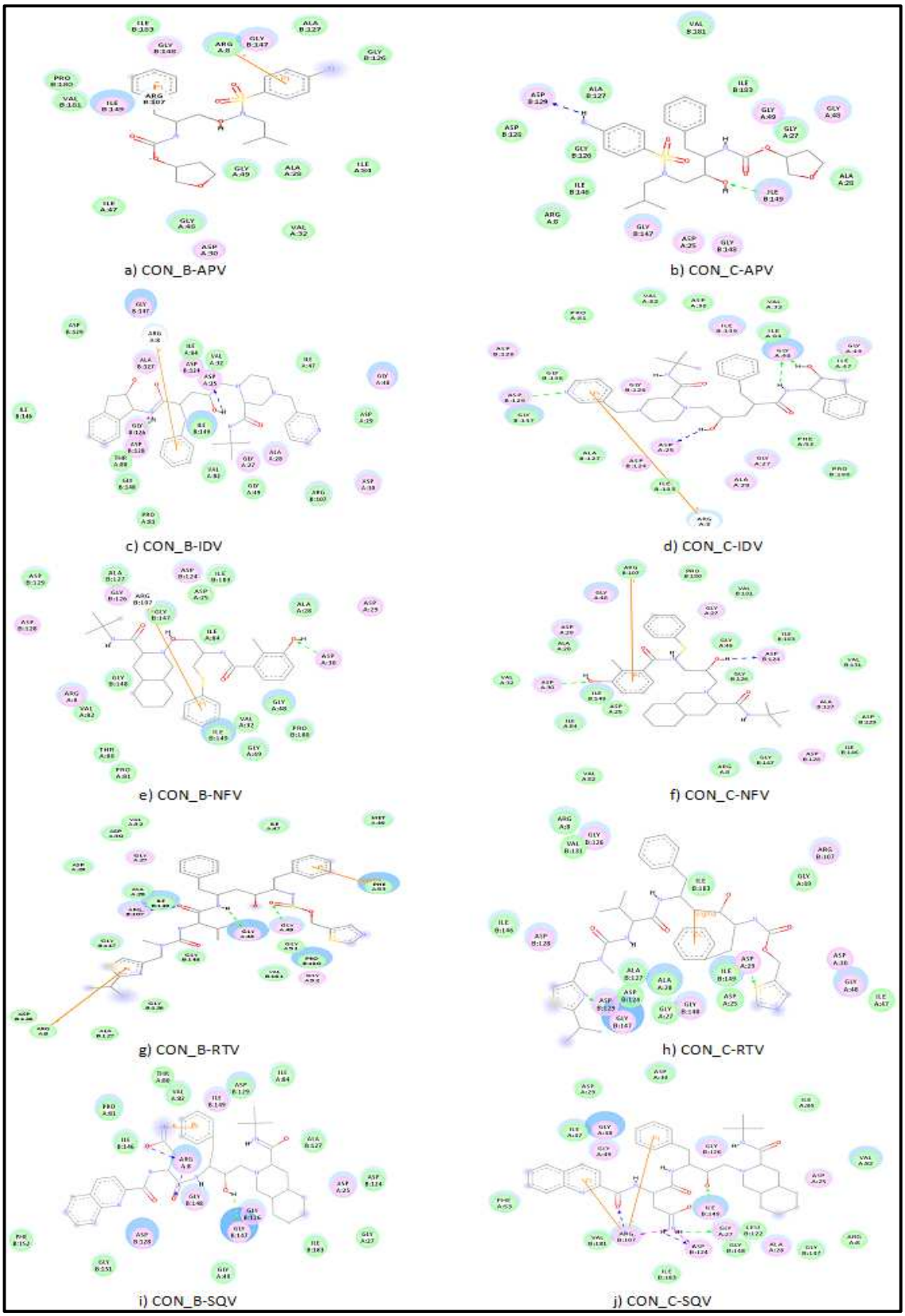


Figure 3-10: 2D interaction profiles between first generation FDA approved PIs and clades B and C proteases. Protease residues involved in the interactions with the ligand atoms are shown by coloured discs. Residues with coloured diffuse background indicate the solvent accessible areas.

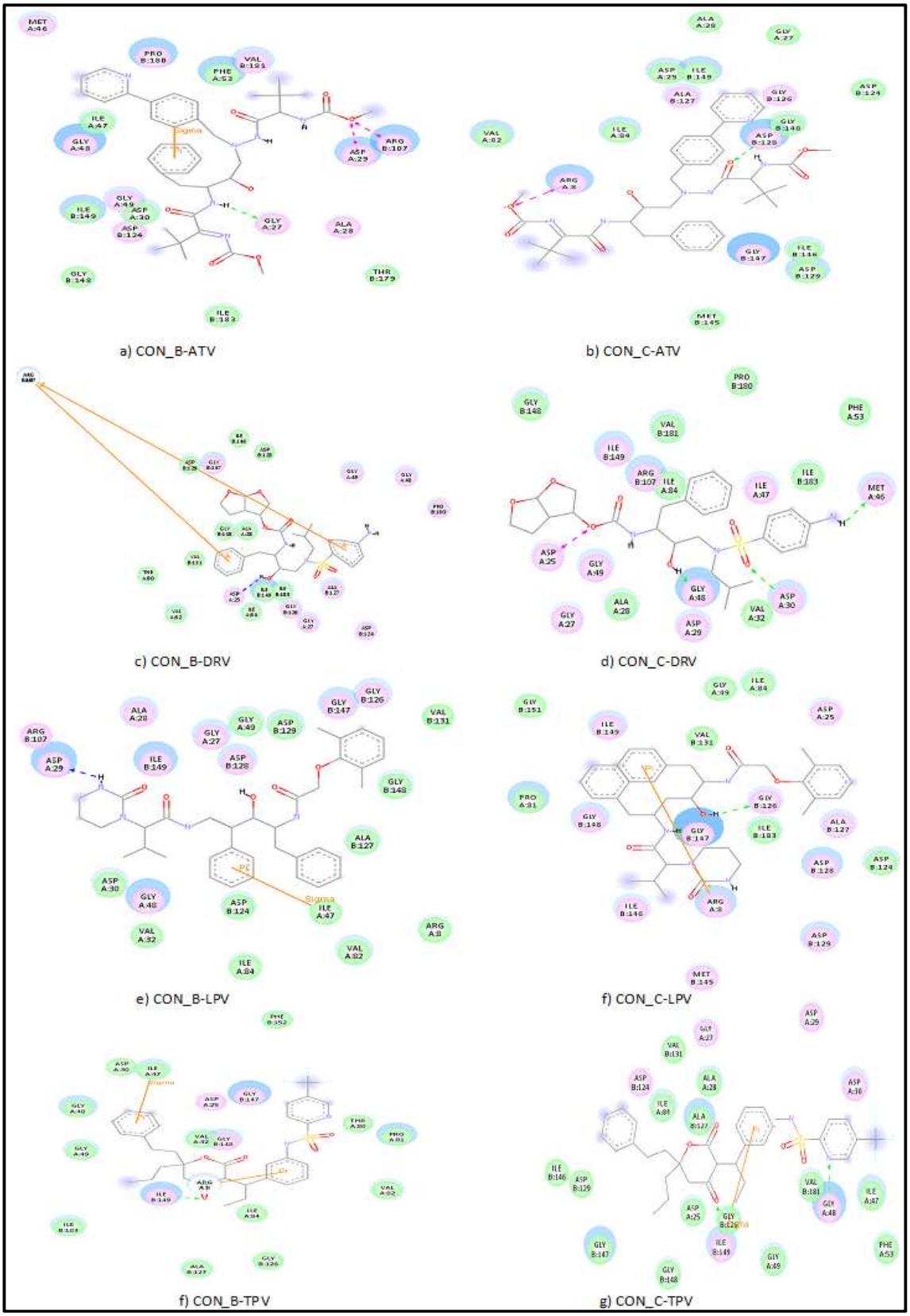


Figure 3-11: 2D interaction profiles between second generation FDA approved PIs and clades B and C proteases. Protease residues involved in the interactions with the ligand atoms are shown by coloured discs. Residues with coloured diffuse background indicate the solvent accessible areas.

Formation of the ligand-receptor complex entails interactions (Chang et al., 2007). Apart from the size of the binding site, data from these interaction profiles show that clades B and C have different binding interactions that result in the differences in the docking scores. Due to symmetry of the protease, an inhibitor can bind from either orientation, but still maintain the binding consistency even though the symmetry will be abolished due to binding of the asymmetrical ligands (Perryman & Lin, 2004). Phenotypically, wild-types B and C are all susceptible to the current FDA approved protease inhibitors, but at the atomic level, these inhibitors behave differently in the two consensus proteases. Generally, some of the residues involved in ligand-receptor interactions in the two clades are conserved. This indicates the specificity of the protease despite the circulating polymorphisms. For instance, most of these inhibitors associated with the Arg8 of both chains. Arg8 of both chains is known to make direct hydrogen contacts with substrates (Prabu-Jeyabalan et al., 2002). For these inhibitors, the contacts were either direct or indirect. These included variation of bonds such as hydrogen bonds, cation-pi bonds and nonbonded contacts. The former two types of bonds are direct bonds. Electrostatic forces are also direct contacts. Ligand-receptor direct or indirect contacts affect binding affinity (Alcaro et al., 2009)

Explanation for differences in susceptibilities observed in the docking of IDV to consensuses B and C, include one cation-pi interaction and four other direct contacts (hydrogen bonding) in active site and flap region. In clade B, only one cation-pi interaction, two hydrogen bonding and one electrostatic interaction were observed (Figure 3-10c and d). Inhibitors that directly interact with Arg8/Arg8' (Prabu-Jeyabalan et al., 2002) and both the active site and flap region residues possess good potency (Heal et al., 2012). For NFV (Figure 3-10e and f), both clades had pi interactions. Consensus B protease interacted with NFV via the nonbonded contacts, hydrogen bonds at Asp30, and sigma-pi interactions at Phe53. Clade C had a hydrogen bond involving Asp25 of the catalytic triad (Brik & Wong, 2003; Kear et al., 2011; Shafer et al., 2001) when docked to NFV. Any active site inhibitor (Hornak et al, 2006a) involving Asp25 (Wu et al, 2008) is likely to exhibit better binding since this enzyme relies heavily on the triad sequence for its catalysis (Brik and Wong, 2003). Hydrogen bonding interactions with Asp30 also contribute to better binding (Prabu-Jeyabalan et al., 2002). This cannot really explain the disparity since in the two clades; this residue is utilized in the NFV-protease interactions, in both B and C consensus sequences.

RTV contains a hydroxyethylene dipeptide isostere (Kaye et al., 2004). Its improved binding in clade B is due to the five direct contacts: a cation-pi, a sigma and three hydrogen bond links. Arg8' participates in one of these hydrogen bonds. In clade C, only two hydrogen bonds take part the association (Figure 3-10g and h). The nature of the RTV size does not favour its effective binding (Figure 3-12). It could not effectively penetrate into the pocket due to the extensive conformational change required.

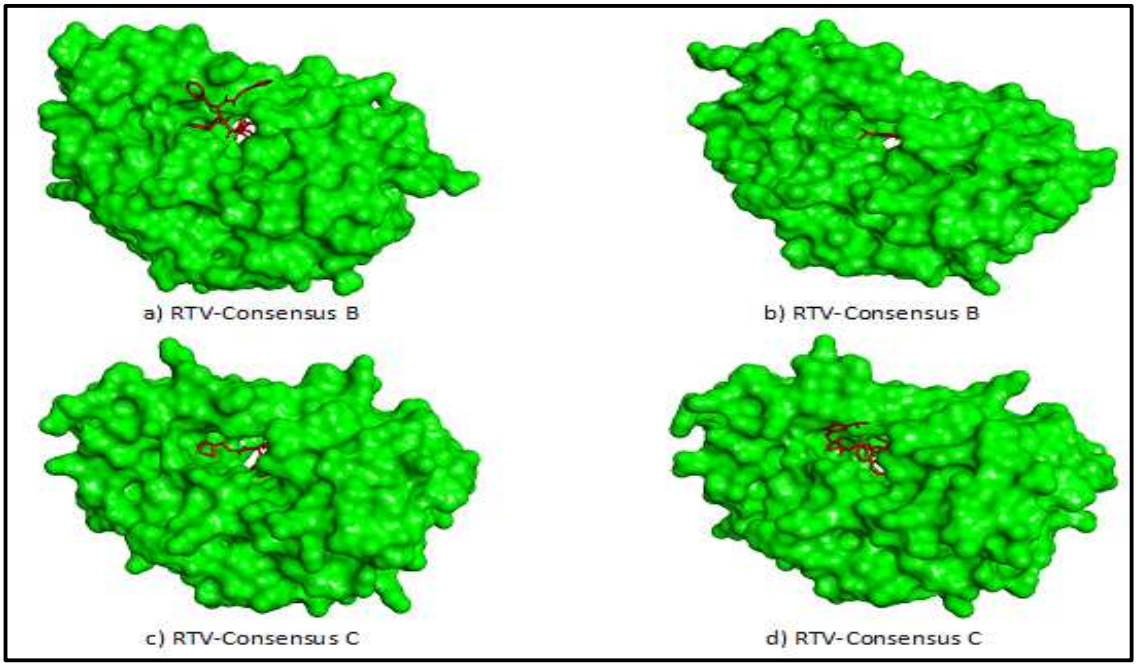


Figure 3-12: Connolly's surface representation depicting RTV binding fits in consensuses B and C proteases. For each complex, both sides of the protease are shown. Green is the enzyme while red is the ligand.

Loss of contacts in mutants has also been reported to affect RTV fitting into the binding site (Dirauf et al., 2010). In subtype C consensus, there was an apparent loss of RTV-protease contacts. Poor binding inhibitors can cause ligands slip off from the binding site (Li et al., 2011).

SQV is a peptidomimetic inhibitor (Alcaro et al., 2009). Even though SQV binding in consensus B protease receptor appeared good (Figure 3-13), our data asserts that wild-type C is more susceptible to SQV therapy than wild-type B. The binding energies were -15.15 kcal/mol and -14.49 kcal/mol respectively. This is due to the extra and favourable bonds involvement in clade C. Clade B showed one cation-pi interaction and three other direct contacts, whereas clade C showed two and six respectively, besides having several nonbonded contacts (Figure 3-10i and j).

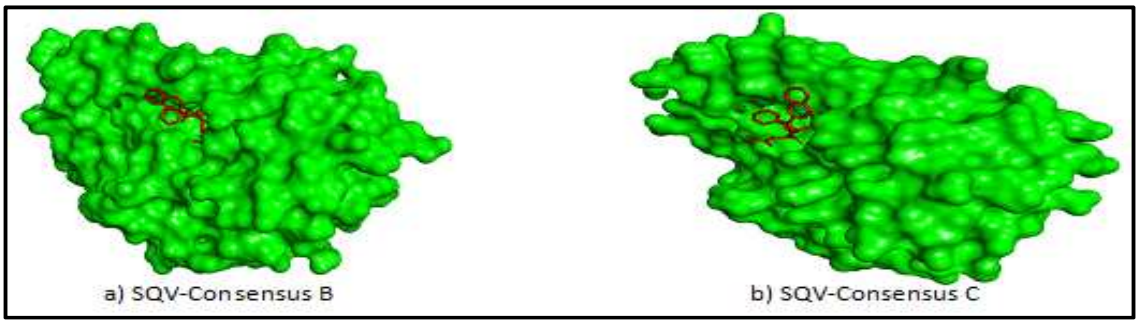


Figure 3-13: Connolly's surface representation depicting SQV binding fits in consensuses B and C proteases. Green is the enzyme while red is the ligand.

APV has a hydroxyethylamine sulphonamide functionality (Alcaro et al., 2009). The binding fit of APV in consensuses B and C protease receptor was reasonably the same (Figure 3-14).

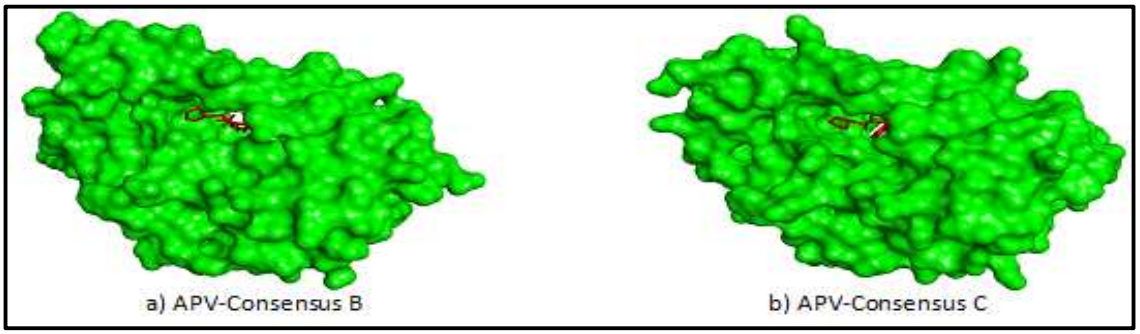


Figure 3-14: Connolly's surface representation depicting APV binding fits in consensuses B and C proteases. Green is the enzyme while red is the ligand.

In clade B, APV formed two cation-pi-interactions and several nonbonded contacts (Figure 3-10a). In subtype C, two hydrogen bonds were formed at Asp30 and Ile50 (Figure 3-10b). Most of the nonbonded contacts seen in subtype B appear in subtype C dockings, but in different chains, due to the symmetrical property of HIV protease enzyme. Asp30 and Ile50 are residues within the active site and flap regions respectively. Inhibitors that interact with both the active site and flap region residues possess good potency (Heal et al., 2012). This explains why APV had better binding energy in clade C than in type B in this study.

DRV was engineered to combat the drug-resistant strains. It displays interactions with the protease through hydrogen bond with the active site residues and not the flap tip residues (Heal et al., 2012). The interaction profiles (Figure 3-11c and d) indicate that unlike in clade B, DRV forms hydrogen interactions with the actives site residues, and M46 and G48 that are part of the flap (Perryman & Lin, 2004), but not flap tip. HIV-1 consensus protease C exhibited more direct contacts with the drug. The nonbonded contacts appeared to be well distributed around the inhibitor than observed in subtype B. DRV had extended better in

consensus C pocket than in B (Figure 3-15). DRV was curled inwards in consensus B active site, hence could not interact with key residues.

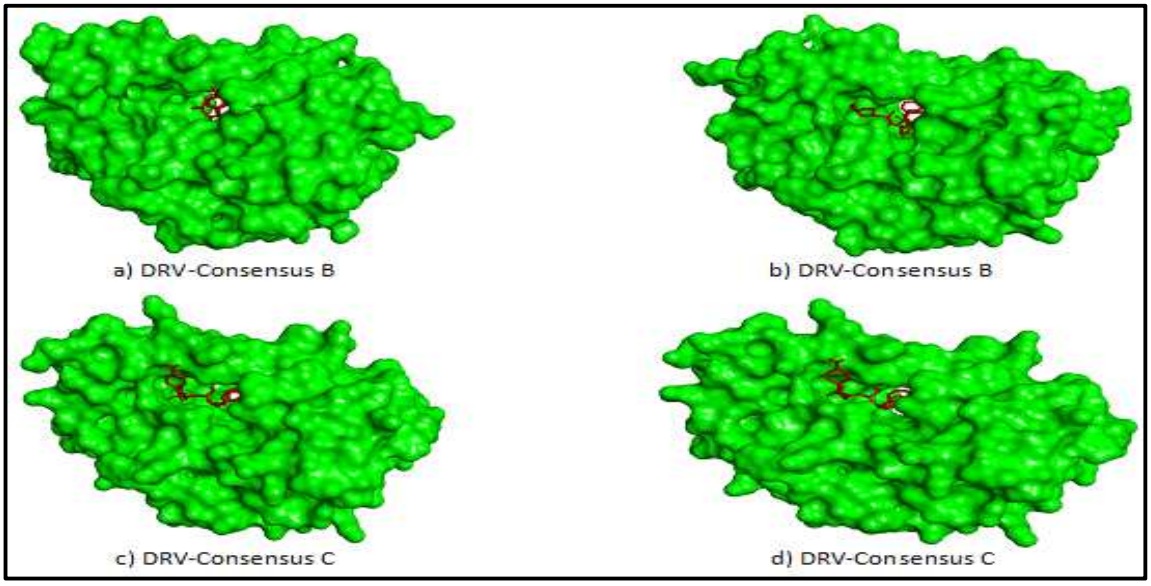


Figure 3-15: Connolly's surface representation depicting DRV binding fits in consensuses B and C proteases. For each complex, both sides of the protease are shown. Green is the enzyme while red is the ligand.

LPV in clade B exhibited a slightly better binding energy than in clade C. Figure 3-11e and f, illustrates that this could be due to the pi-sigma interaction (a strong type of bond) involving the Ile47 residue and one of LPV's benzyl rings. It forms a hydrogen bond with Asp29 and fits well into the binding site and as a result it is surrounded by residues which cluster near it. The arginine at the eighth position in the protease is known to stabilize the LPV binding (Alcaro et al., 2009). Both Arg8/Arg8' are involved in LPV interaction in the consensus B model. Compared to clade C, the interaction of Asp25 in clade B was in close proximity to the inhibitor moiety. In subtype C, only one arginine at position eight participated in the complex formation. The bonding was a cation-pi interaction, which is usually weak. Two hydrogen bonds enhance the interaction between Gly27 and Gly48. The ligand in clade C was curled inwards in the advent of undergoing conformation in order to best fit into the pocket. It could not effectively extend into the binding site (Figure 3-8c and d).

Lastly, TPV is a non-peptidic inhibitor with a broad antiretroviral activity against mutants. It was observed to interact with the Ile50/Ile50' of the flap tips through hydrogen bonding (Heal et al., 2012). A pi-sigma interaction with Ile, cation-pi bond with Arg8, hydrogen bond with Ile50', and nonbonded contacts including Asp25 were observed. TPV binding in consensus C protease was better as indicated by the binding energy. It had a pi interaction

with Gly27', hydrogen bonds with Gly48 and Ile50', and nonbonded contacts including the Asp25/Asp25'. HIV-1 clade C protease had more nonbonded contacts than in consensus B (Figure 3-11f and g). Other factors leading to the difference in the binding energies of TPV in clades B and C include the nature of the inhibitor. Since TPV is an adaptable inhibitor, the elongated form seen in clade C had a better fitting into the binding site (Figure 3-8e, f, g and h). Shape complementarity existed (Prabu-Jeyabalan et al., 2002). An inhibitor with an extended structure has high chance of being an efficient inhibitor because it interacts with several residues, as compared to short-structured inhibitor (Wu et al., 2008).

3.6.3.2.2 Evaluation of Docking: Selected Patient Samples in Focus

A comparison in the drug responses between the docking studies and Stanford HIVdb algorithm were made. Most of the docking results were in agreement with the Stanford HIVdb interpretation. Generally, the samples without mutations after drug-exposures had slight changes when compared against one another (before and after drug-exposure). However, a few had unexpected results. This was due to some of the challenges associated with docking results. These included the number of energy evaluations (Hetényi & Spoel, 2002) and torsions (Morris et al., 2009). Due to the high number of docking experiments, only 4.5 million energy evaluations were set. The number of energy evaluations influences docking simulations. It was found out that below 10 million energy evaluations per trial, the number of correct binding sites reduces. Large number of energy evaluations would therefore be required to attain the lowest energy conformation. However, with the greater than 20 million number of evaluations for some systems, the conformation of the system is rendered static (Hetényi & Spoel, 2002). The four side chain receptor flexibilities (R8, R8', D25 and D25') established in the protease are likely to have had some impact on the results.

The minimum binding energies for the consensus C were used as references to interpret the drug susceptibilities for these infant proteases. They were used to evaluate the extent of susceptibilities. Stanford HIVdb has the susceptibility as level I, of which it has statistical limits (Tang et al., 2012; Zazzi et al., 2004), deducing that a mutant that is susceptible to a particular drug can either be absolutely or incompletely susceptible (i.e., not completely resistant to that particular drug). Based on Stanford HIVdb mutational scores, when susceptibility gradually diminishes, resistance sets in. At baseline, the drug responses from

the docking studies were highly consistent with the Stanford HIVdb interpretation. Natural polymorphisms in non-subtype B are thought to have a negative consequence on the antiretroviral therapy (Nukoolkarn et al., 2004). Docking results have implicated the changes in binding energy to polymorphisms. We expect these findings and contribution to HIV/AIDS treatment to hold true even in validation studies from phenotypic perspectives.

Due to the high-throughput nature of this screening, these data are analyzed in general, but emphases are placed where necessary. This is with respect to the circumstances where drug-linked mutations emerged. Using the consensus C as a reference, we noted a trend in the drug susceptibilities. In most cases there was a significant change. The polymorphisms (frequently occurring, commonly occurring and rarely occurring) are discussed in chapter II. Susceptibility to SQV was favoured by these polymorphisms. Virtually all the binding energies of the docked proteases to SQV had their lowest binding energies lower than that of the consensus C, -15.15 kcal/mol, indicating increase in susceptibilities of these "natural" mutants. It is important to note that the term "increase" here refers to better binding (more negative binding energies), while "decrease" indicates poorer binding (a less negative binding). In sequences 5079, 5086 and 5117, SQV susceptibility decreased. The former sequence had a L10 deletion. Sample 5211 also had a decline in the binding energies. These polymorphisms cumulatively reduced the binding energies of NFV and IDV to a greater extent than of the rest of the drugs. The clade C consensus had -16.99 Kcal/mol and -17.08 kcal/mol binding energies of IDV and NFV respectively. Majority of the sequences had values less than these. In 5114, 5198, 5207, and 5245 the binding energies for NFV increased. 5207 recorded the highest value, -18.02 kcal/mol. For 5211, only LPV showed an increase in susceptibilities. It had an energy of -15.29 kcal/mol against -13.18 kcal/mol that of consensus C. Some sequences did not show any change in susceptibilities. Take sequence 5178 for instance, the SQV binding energy remained at -15.15 kcal/mol.

When these binding energies were matched with those of the Stanford HIVdb, virtually all of the drug-naïve strains viruses would be susceptible to the protease inhibitors. According to this "gold" standard database (Toor et al., 2011), only samples 5014, 5086, 5094, 5198, and 5245, would exhibit low-level resistances due to T74S and or K20R or V77I. The latter two samples indicate contradicting results as per the docking scores. They all have significantly better or worse binding energies. Besides T74S, the other polymorphisms possibly are

involved in this behaviour. Ligand-receptor interactions were generated to assist in unravelling these behaviours. Figure 3-16 and Figure 3-17 are the NFV-protease association profiles. These profiles were compared against those of subtypes B and C in Figure 3-10e and f. Here we used two 2D interaction profiles refer to the 2D interactions. The first one, Figure 3-16, is an in-house ligand-receptor interaction profile that we generated through Python and Perl scripting languages. We refer to it as distance-based fingerprinting profiles (DBFPs). The second one utilized the Discovery Studio Visualizer to indicate how the ligand groups interact with specific atoms of the receptor.

The feasibility of ligand-receptor interaction has been demonstrated by use the fingerprints that measures the surface force distance of interacting ligand and receptor atoms (Luo & Chan, 2012). The fingerprinting maps are able to show the behaviour of the ligand with respect to the entire receptor, e.g., it would be possible to tell when a given ligand has changed its binding orientation. Via DBFPs, changes in interactions can easily be visualized.

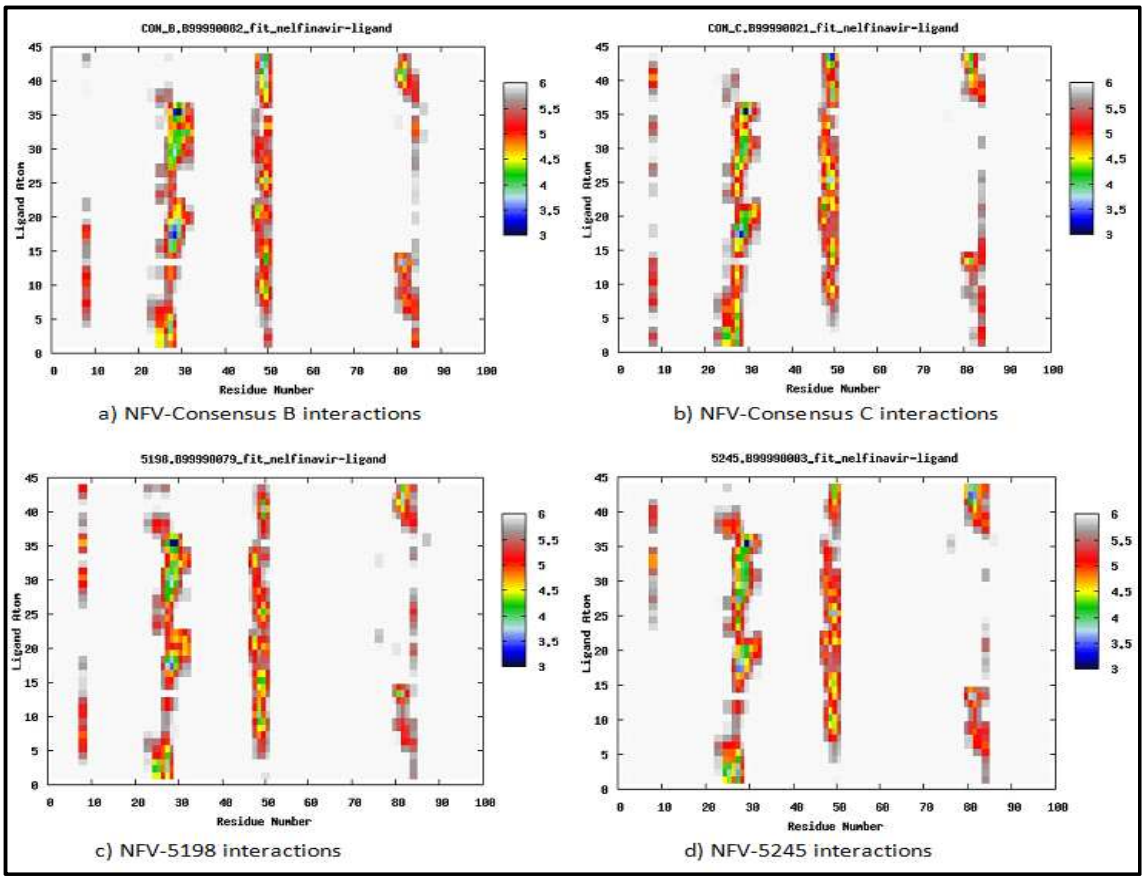


Figure 3-16: Fingerprints of consensuses B and C, and two drug-naïve patient samples with low-level resistance to NFV. The y-axis depicts ligand atoms while the x-axis depicts the overlaid corresponding protease residues. Ligand-receptor interactions within various distances (Å) are shown by a colour code, with black being the closest while white the farthest.

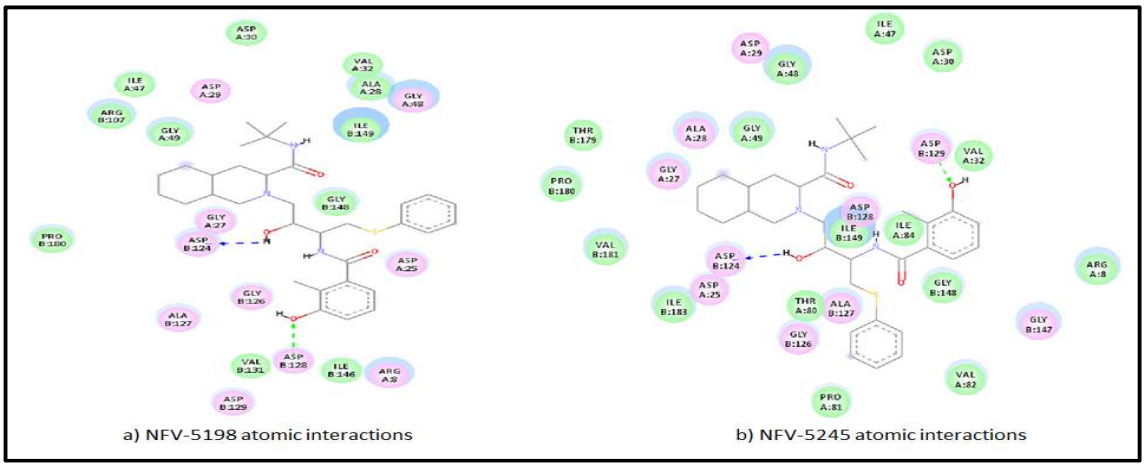


Figure 3-17: 2D interaction profiles between two drug-naïve patient samples and low-level resistance to NFV. Protease residues involved in the interactions with the ligand atoms are shown by coloured discs. Residues with coloured diffuse background indicate the solvent accessible areas.

One thing that can be concluded from Figure 3-16 is that there are similar interaction modes between the ligand and the receptor in all the four cases. There exists some consistency. The y-axis shows the ligand atoms numbered from 1 to n, where n is the number of atoms in a given ligand. For NFV, n is 44. The x-axis shows the number of the protease residues numbered from 1 to 99. Protease is a homodimer containing 99 amino acid residues per chain (Mao, 2011; Ode et al., 2007). In the maps, the two chains were overlaid, so one of limitation of such profile is its inability to specify which chain is actually involved in the ligand atom(s)-receptor residue(s) interaction(s). The colour code discriminates the distances in angstrom between ligand-receptor interactions, with an increasing distance from black to white. This DBFP is able to show any distance even above 6 Å (whereas that of the Discovery Studio Visualizer only captures interactions mainly of less than 4 Å) and hence gives an indication of the positioning of the ligand in relation to the protein.

In all the cases (Figure 3-16), not all the protease residues interact with the ligand. Some protease sites show extensive associations with the ligand while others show minimal interactions. These sites include the nose (Perryman & Lin, 2004), active site floor (Martin et al., 2005), flap tip (Kandathil et al., 2009) and wall turn (Perryman & Lin, 2004), but it is only the 80s loop (Ali et al., 2010; Dirauf et al., 2010) that is extensively involved in the wall turn region. Amino acid residues P79 and I85 are minimally involved. Of the four fingerprints (Figure 3-16), notable changes occur in the nose region. In subtype B, ligand atoms 26-42 are do not interact with the nose residues. Supposing these interactions are favourable, then this would explain the susceptibility variances between clade C and the three subtypes

C. The two infant samples have higher binding energies than consensus C. Their interaction maps indicate that ligand atoms 35-40 are in close proximity with the protease active site residues. These associations could be favourable. Interactions involving the nose and ligand residues 1 to approximately 23 appear not to be favourable because whenever they participated, low binding energies resulted in comparison to sample 5245.

Favourability and non-favourability can only be discerned if both energies and thermodynamics are modified so that the attention is shifted from the DBFPs to energy-based fingerprinting profiles (EBFPs). The following formalism can be utilized:

$$V = W_{vdw}\sum_{i,j}\left(\frac{A_{ij}}{r_{ij}^{12}}-\frac{B_{ij}}{r_{ij}^{6}}\right)+W_{hbond}\sum_{i,j}E(t)\left(\frac{C_{ij}}{r_{ij}^{12}}-\frac{D_{ij}}{r_{ij}^{10}}\right)+W_{elec}\sum_{i,j}\frac{q_i q_j}{\varepsilon(r_{ij})r_{ij}}+W_{sol}\sum_{i,j}(S_i V_j + S_j V_i)e^{\left(-r_{ij}^2/2\sigma^2\right)}$$

where V = Force field due to pair-wise atom terms, *W* = Weighting constants (W_{vdv} = 0.1560, W_{hbond} = 0.0974, W_{elec} = 0.1465, W_{sol} = 0.1159), *A* and *B* = Parameters from Amber force field, *C* and *D* = Maximal depth well of 1 kcal/mol at 2.5 Å for S-H and 5 kcal/mol at 1.9 Å for O-H and N-H, *E(t)* = Hydrogen bonding directionality, $V_{i,j}$ = Volume of surrounding atoms, *S* = Solvation parameter, *q* = Charge, ε = Dielectric constant (78.4) of bulk water at 300 K, σ = Distance weighting factor (3.0 Å) (Huey et al., 2007). In the infant samples, 5198 and 5245, the residues of the protease wall turn, P79 and I85, are weakly complexed with NFV, but appear to result in better binding energies.

The interactions as produced by the Discovery Studio Visualizer indicate that the same protease subsites are still engaged in ligand-receptor interactions (Figure 3-17). The 2D profiles especially of the wild-type B and consensus C show that the cation-pi interactions between the ligand atoms and R8 of the nose region might be partially or totally favourable. All these profiles did not indicate all the protease residues seen to participate in the ligand-receptor interaction. Take protease residues P79 and I85 for instance in cases c) and d). Clade B and C wild-types interactions with NFV are discussed under Figure 3-10, therefore anything discussed here would be a plus to the aforementioned discussion pertaining NFV-protease wild-type B or C interactions. Compared to wild-type B, consensus C had nonbonded contacts with D25 and I50. The other formed a hydrogen bond. D25, D29 and D30 of the active site have a significant binding role (Toor et al., 2011). Sample 5198 had D25 and D29 participating in hydrogen bond whilst the other D25 and D30 participated in

nonbonded contacts. The V82, part of the 80s protease loop (Dirauf et al., 2010), appears to be well positioned for interaction than in consensus C. D25 and D30 of sample 5245 had hydrogen interactions with NFV. D29, I50 and I84 present good steric interactions with NFV, hence high affinity and binding energies. R8 is among the residues clustering without bonding around NFV. R8, D29 and D30 are often involved in substrate binding (Prabu-Jeyabalan et al., 2002), thus an inhibitor utilizing these residues and those aforesaid, in its interactions, is likely to possess high binding energies. The smaller size of this protease pocket has an effect in its binding. NFV fits better as compared to the other three proteases.

Docking also indicate that 3021, 3043, 5086, 5098, 5094, and 5117, would be partially resistant to NFV. Sample 5166 and 5166 had a change in one decimal place in the binding energies when compared to that of the subtype C. None of the patient sequences had all the nine consensus C polymorphisms harboured in them simultaneously. They had some of the wild-type C polymorphic sequences, and other common or rare mutations. The pooling roles of these mutations therefore affect drug responses. Apart from NFV therapy being affected by these polymorphisms in some of the patient samples, RTV also appeared to face similar fate. In most cases, these polymorphisms resulted in a decrease in the binding energies, but still within the susceptibility range of clades B and C. Clade C polymorphisms are therefore likely to lead to emergence of RTV and possibly NFV. Further characterization would be needed to understand the role of each of these polymorphisms in subtype C.

Many effects of most mutations in clade C on drug response hitherto remain a conundrum (Bessong, 2008) due to minimal researches directed towards them. Most of the available data are not in unison (Archary et al., 2010; Ode et al., 2007). The discordances in the available data would also suggest that clade C mutations could be possessing multifaceted-tactics in upholding its fitness and mounting drug-resistance. For many proteases lacking mutational changes we expected them to exhibit similar binding energies and inhibition constants, e.g., 5144-514412 pair, and 5178-517812 pair among others. Any great deviations seen are due to the confounding factors in the docking. Genotyping algorithms do not consider preceding immunomonitoring and toxicological reports (Tang et al., 2012).

The results for the samples with mutations after drug-exposure were compared against the Stanford HIVdb. The discrepancies that were evidenced, particularly in the infants with drug-

induced mutations, are shown in Table 3-4. This could raise the question pertaining to either the accuracy of the Stanford HIVdb or the docking simulations. Accuracy exists in both cases, with Stanford HIVdb having greater weight since its interpretations are generated from clinical studies (Zazzi et al., 2004). To back up our findings, part of the molecular dynamics has been presented in the next subchapter (3.6.4 Molecular Dynamics Simulations). The motive behind this is that docking algorithms are usually optimized for speed rather than accuracy, but molecular dynamics is also optimized for accuracy (Durrant & McCammon, 2011). The docking results of the RU-synthesized ligands when bound to the proteases were not compared to the Stanford HIVdb since the Stanford HIVdb only includes interpretation relating to the FDA approved drugs. This HIVdb does not show the susceptibility responses to RTV as a single dose. It is used as a booster for all the PIs with the exception of NFV. AMP is also not included possibly because of its structural and catabolic relatedness to fAMP.

Discrepancies are due to the clade inequalities. The Stanford HIVdb is tailored for clade B (Toor et al., 2011); other clades had either not emerged or been characterized. We mainly focussed on samples whose inconsistencies with the Stanford HIVdb were likely to occur based on the large deviations of the binding energies prior to and after drug-exposure. Docking results of wild-types B and C were also used in the assessment. The majority of the results between the docking reported here and the information from the Stanford HIVdb corresponded. For example, both Stanford HIVdb and docking algorithms disclosed that the proteases were susceptible to DRV. Protease 301812 has a stretched cavity and is resistant to ATV, unlike 3018. The drug resistance reports were the same from the two algorithms. The ligand-receptor interactions (Figure 3-18) proved to influence the susceptibility.

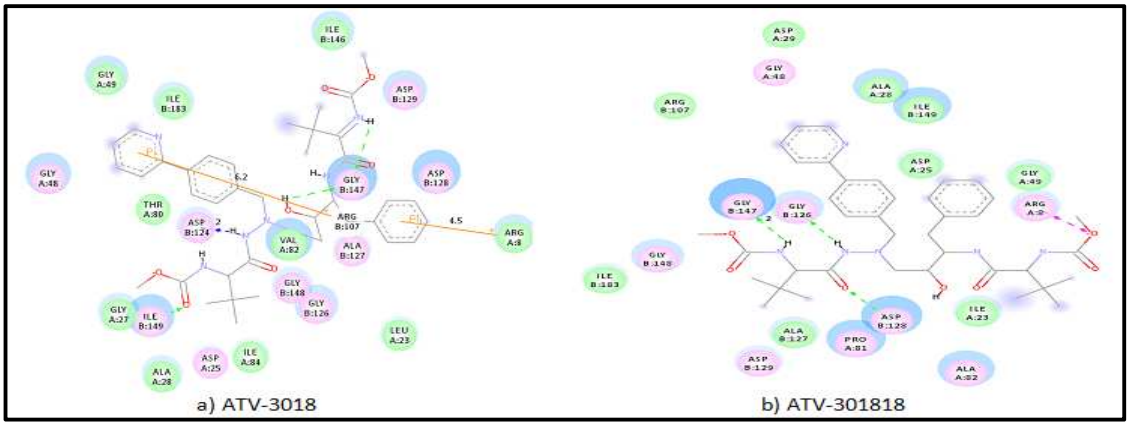


Figure 3-18: 2D interaction profiles of ATV-3018 and ATV-301812. Protease residues involved in the interactions with the ligand atoms are shown by coloured discs. Residues with coloured diffuse background indicate the solvent accessible areas.

Hydrogen network and hydrophobic contacts play crucial roles in the ligand stability. A ligand with weak contacts is likely to dissociate easily (Li et al., 2011). Association of ATV and 301812 has loss of interactions that are apparent in ATV-3018. Hydrogen bond interaction involving one G48' and I50' are absent, hence leading to instability of this ligand (as evidenced from the binding energies).

Our study did not have the levels for defining the degree of drug response, but we assumed that large deviations leading to decrease in binding energies correlates to high-level resistance, as in the case of NFV in 503252, but if such deviations cause increases in binding energies, then these define high-level susceptibilities as in the case of DRV in 305112. With the prevailing computational challenges, it was quite impossible to state potential low-level resistance, especially differentiating it from the low-level resistance. Both the Stanford HIVdb and docking algorithms showed low-level resistances for TPV in 50459, APV in 301812, RTV in 503252, TPV in 521112 among others. In 503252, contraction of the pocket is likely to have affected the binding of the RTV. Some mutants had decline in susceptibilities but were still susceptible when compared against subtype C, e.g., SQV in 51989.

Table 3-4 displays that clades B and C variants respond differently to some of the FDA approved protease inhibitors.

Table 3-4: Comparison of the lowest energy docked conformation from each cluster and drug susceptibility reports (from docking and Stanford HIVdb) of the patient samples (proteases in closed conformation that developed mutation(s) after drug exposure). Only discrepancy reports in therapeutic response are tabulated.

Patient ID	Drug used to Assess Response	DRM According to Stanford HIVdb	Binding Energy (kcal/mol)		Inhibition Constant (Ki) in pM		Drug Response S = Susceptible R = Resistant	
			Before	After	Before	After	Stanford	Docking
301812**	IDV	**I54V**, **V82A**, L23I	-15.18	-15.34	7.44	5.68	Intermediate R	S
	LPV		-13.91	-15.28	64.08	6.28	Intermediate R	S
305112**	ATV	**M46I**, **I54V**, **V82A**, L10F	-12.10	-13.93	1034.00	61.81	High-level R	S
	SQV		-15.21	-15.28	6.27	7.05	Intermediate R	S
305152**	ATV	**M46I**, **I54V, V82A**, L10F	-12.10	-13.95	1034.00	59.44	High-level R	S
	SQV		-15.21	-16.03	6.27	1.79	Intermediate R	S
503252**	LPV	**M46I**, **I54V**, **V82A**, F53FL	-13.64	-14.96	99.97	10.85	Intermediate R	S
	SQV		-14.13	-15.04	43.87	9.48	Intermediate R	Low-level R
	TPV		-13.78	-14.40	79.36	27.92	Low-level R	Intermediate R
50459*	ATV	**V82A**, L90M	-13.33	-13.97	169.82	57.94	Intermediate R	S
	NFV		-16.93	-18.59	0.77	0.02	High-level R	S
	LPV		-14.21	-15.86	38.12	2.36	Intermediate R	S
	SQV		-15.08	-17.34	8.79	0.02	Low-level R	S
51989**	IDV	**I54V**, **V82A**, L33FL	-15.70	-16.86	3.09	0.44	Intermediate R	S

Regimen: **LPV/r, *RTV, **DRM** – Drug Resistance Mutations. Major mutations are shown in **bold.**

The Stanford HIVdb interpretations showed that most of the patients would be resistant to most PIs. It is only DRV and or TPV that could find medicinal worth in most cases in the infants with drug-associated mutations. However, the docking analyses contradicted some of those cases. This is one of the few studies that report these, similar to a study conducted in India by Toor and colleagues. They showed that in the advent of selecting an apt second-line therapy regimen, a subtype C-infected patient failing first-line therapy was reported to have low-level resistance to LPV according to the Stanford HIVdb, contrary to the docking studies that indicated susceptibility. Upon treatment switch to the second-line that included LPV regimen, the patient recovered both clinically and immunologically as the viremia was significantly suppressed (Toor et al., 2011). Misinterpretations of drug resistance testing reports reduce the number of regimens that a patient can actually be started on or switched to. It might also be problematic supposing a patient is prescribed a regimen to which the mutant is resistant to (given that the database had branded the mutant "susceptible"). Therefore there might be misconception, in switching a patient from a "not good" drug to a "worse" or "worst" drug that has been speciously identified as a "better" drug.

Detailed assessments using both the Discovery Studio Visualizer and fingerprinting profiles are included in the Supplementary Data/Chapter III/Docking/Interactions/Discrepancies; only two illustrations are discussed here: Figure 3-19 and Figure 3-20. 3D representation indicated that LPV fits well in 301812. Figure 3-19a and b clarify why this mutant is LPV-susceptible. It forms hydrogen bonds to I50 and both D25 and I84 have a favourable spatial orientation. Residues 82/84 are part of S1 and participate in binding of LPV (Logsdon et al., 2004). 80s loop is flexible (Ali et al., 2010) and 82A caused conformational changes to favour ligand interactions with I84. An efficacious inhibitor may be the one that interacts concurrently with the active site and flap residues (Heal et al., 2012). In 3018, only one hydrogen bond exists with G48. Both D25 and I50 have poor steric effects. Figure 3-19 concurred with these. New contacts between LPV atoms 1 to 25 and protease nose emerge. Other bonds form in the active site and 80s loop. For IDV, the nonbonded contacts involving both I50' and I84 appear to be energetically more favourable than in 3018. More links are in the flap region of the mutant. The fingerprints were found to be sensitive as they demonstrated the participation of ligand-receptor interactions above the 4 Å distance.

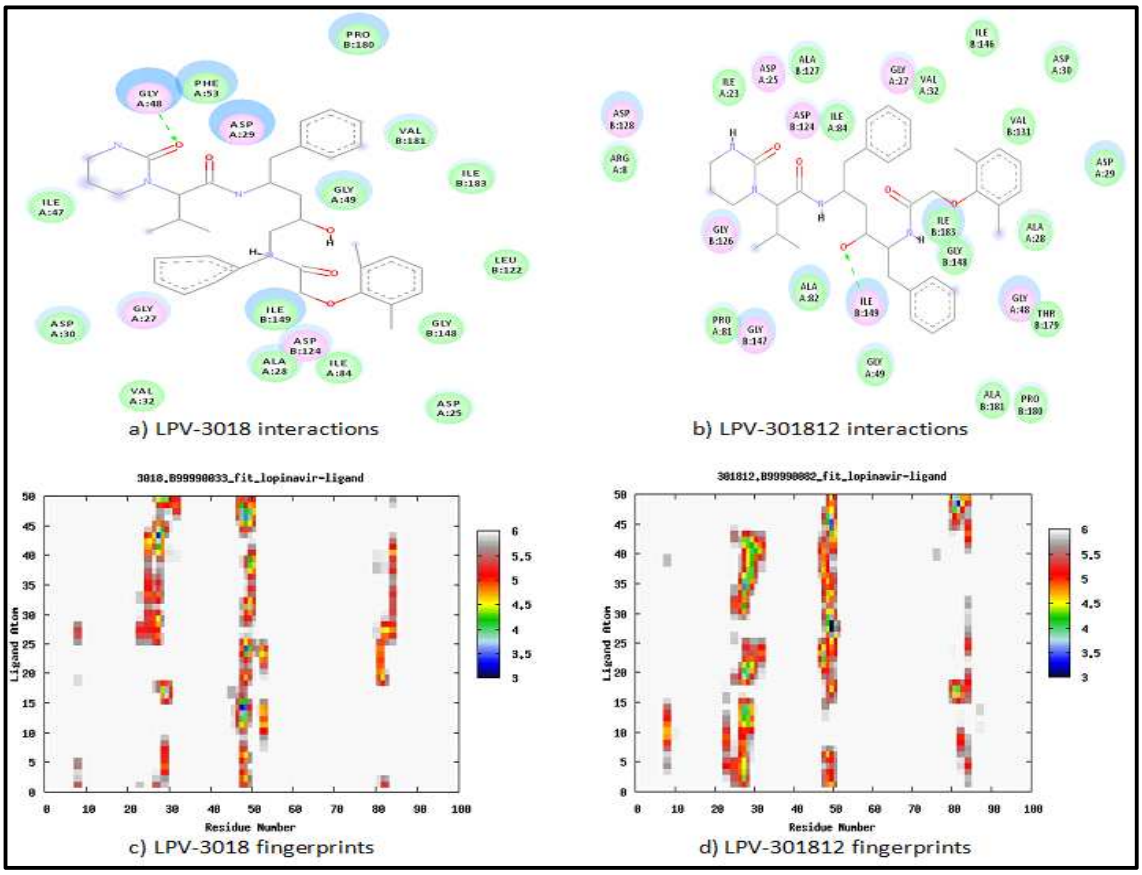


Figure 3-19: Interaction profiles between LPV and 3018 and 301812 using Discovery Studio Visualizer and fingerprints. For a) and b), protease residues involved in the interactions with the ligand atoms are shown by coloured discs. Residues with coloured diffuse background indicate the solvent accessible areas. For c) and d) the y-axis depicts ligand atoms while the x-axis depicts the overlaid corresponding protease residues. Ligand-receptor interactions within various distances (Å) are shown by a colour code, with black being the closest while white the farthest.

Even though some inhibitors and potential inhibitors could be used to target the mutants, some HIV variants appear to be resistant to them based on the binding energies. Mutations within the "substrate envelope" are known to affect inhibitor binding while that of the substrate either increases or remains unchanged (Ali et al., 2010). L23 occurs at active site pocket. Due to its proximity to D25, it rarely undergoes mutation. When L10I emerges, it causes cross resistance. There is a high L10I and V82I potentiation (Johnston et al., 2004). Infant sample 301812 has I54V and V82A as the major mutations. These together with the nine clade C polymorphisms and E35D and M36I, conferred resistance, but not to all protease inhibitors. Table 3-4 defines that the mutant 301812 is susceptible to IDV and SQV. In 3051, both ATV and SQV have poor fit due to the polymorphisms associated with them. Figure 3-20a shows that ATV interacts with the protease via two hydrogen bonds with G48 and fifteen nonbonded contacts with L10, L23, K45, I50', I84, and R87 among others. Both D25 and D25' are not involved.

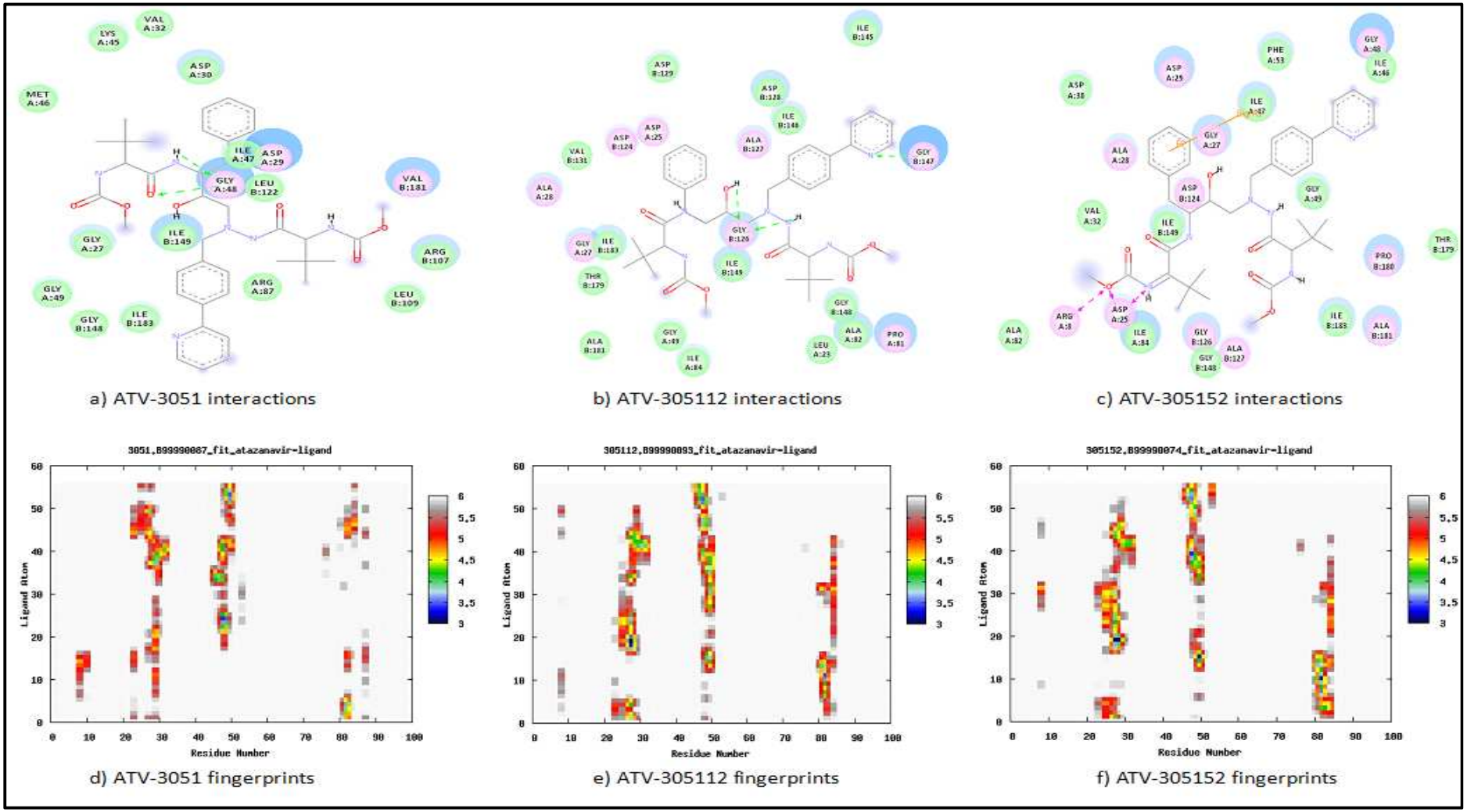


Figure 3-20: Interaction profiles between ATV and 3051, 305112 and 305152 using Discovery Studio Visualizer and fingerprints. For a), b), and c, protease residues involved in the interactions with the ligand atoms are shown by coloured discs. Residues with coloured diffuse background indicate the solvent accessible areas. For d), e), and f), the y-axis depicts ligand atoms while the x-axis depicts the overlaid corresponding protease residues. Ligand-receptor interactions within various distances (Å) are shown by a colour code, with black being the closest while white the farthest.

ATV in 305112 has good fitting. It associates with the receptor through two G27 and two G48 hydrogen bonds, and twenty nonbonded contacts as seen in Figure 3-20. D25 and D25' favourably cluster around the ligand. All these lead to a good binding energy, a higher binding energy than that of 3051. Figure 3-20 indicates that In 305152, the ATV also has good binding, nineteen nonbonded contacts. Both D25' and I50 have good spatial alignments. R8 and D25 electrostatically interact with the ligand. This pool of interactions leads to a better binding energy. The binding energy in 305152 is more enhanced than in 305112 possibly due to the E35D present in the former. The surface interaction profiles in Figure 3-20 reveal that few contacts are involved in 3051.

Interaction with L10 is present in the drug-naïve strain, but absent in the 305112 and 305152 mutants. In the evolution of the virus, L10I emerged in order to evade L10 association with the ligand. This resulted in the mutant exhibiting cross-resistance, but conferred a selective advantage on it worked against SQV since its susceptibility increased. For these mutants, the SQV has good fittings, thus good energetics (Ali et al., 2010). 305152 makes direct contacts with R8 that is crucial in binding (Prabu-Jeyabalan et al., 2002), D25, and G48. In 305112, two hydrogen bonds are involved; D30 and G48. SQV interaction with 3051 involves pi-pi bond with F53, which seems energetically unfavourable. Possibly, there is a steric clash.

LPV has good binding fit in 503252. LPV interacts through hydrogen bonds with the protease D25 and D29. Extensive interactions are seen in at the active site in the fingerprint maps. The slight decrease in binding energy of TPV-503252 compared to TPV-5032 is attributed to unfavourable interactions. Even though there is a direct contact of R8'-F, the R87'-pi bond is likely to be associated with high energy cost that eventually affects the overall binding energy. 5032 has better binding energy due to the hydrogen bond involving the sulfonic moiety of the ligand and A28'of the protease. This mutant interacts with SQV via hydrogen bond at D29. Other key residues participating close range include L23', D30 and I50. 5032 interacted with LPV only through direct G27 contact and nonbonded contacts that included the D25 residue. The maps highlighted better ligand atom-protease residues clustering. From the fingerprinting profiles, strain 51989 susceptibility to IDV was colossally due to the long-range interaction. 5198 had better short-range interactions (within 4 Å) than 51989, but the overall effect encompassing even the long-range interactions (4 - 6 Å) favoured

binding IDV in the mutant. Hydrogen interactions of IDV with the mutant involve R8 and G48 (two bonds). For 5198, they comprise R8', D25, G48 (two bonds) and I50. In spite of the loss of some interactions, new interactions emerge in the nose and 80s loop. At the flap region, interactions below 3.5 Å exist in 51989. In 5198, there are only two.

I54V (even in presence of L90M) confers SQV resistance in subtype B (Alcaro et al., 2009). In subtype C, SQV is ineffectual against mutants with this mutation in addition to non-polymorphisms. V82A/L90M dual mutation in occurrence with polymorphisms conferred SQV susceptibility in 50459 contrary to the Stanford HIVdb. V82A together with polymorphisms conferred susceptibility to SQV in 52423. In the drug-naïve infant, 5207, three hydrogen bonds arising from the active site (D25' at 1.9 Å, G48' at 2.1 Å and 3 Å distances) interact with SQV as visualized from the Discovery Studio Visualizer. Nonbonded contacts including with I50' also contribute to the -16.33 kcal/mol binding energy and 1.08 pM inhibition constant. Two V82 residues are in close proximity and interact with the phenyl rings. V82 often forms van der Waals interactions with aromatic quinolone of SQV (Olsen et al., 1999). The mutant lacks the three hydrogen bonds; the hydrogen bonding involves D25 at 2.1 Å. Only one 82A of which is distantly placed exists within the interaction window; this lessens the SQV-protease hydrophobic clusters. Its binding energy and inhibition constant are -14.79 kcal/mol and 14.38 pM, respectively. Figure 3-21 shows the fingerprints.

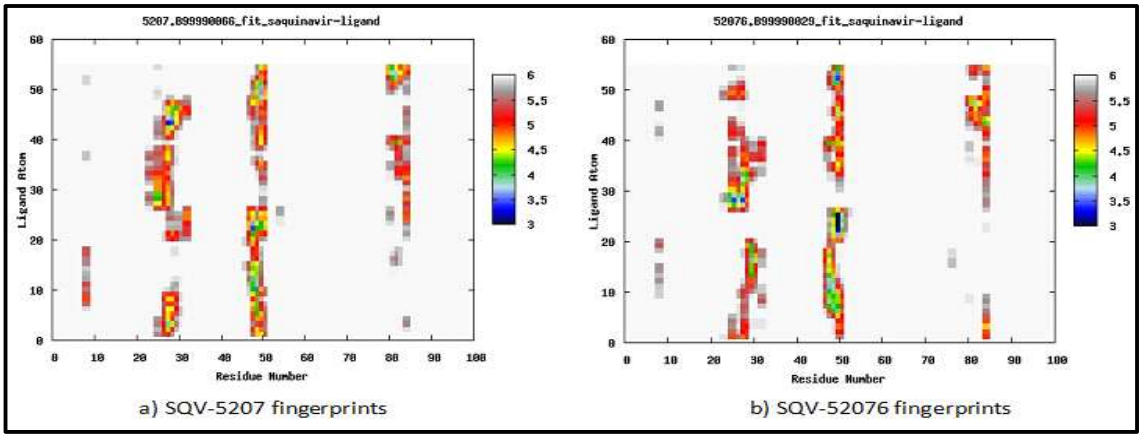


Figure 3-21: Interaction profiles between SQV and 5207 and 52076 using fingerprints. The y-axis depicts ligand atoms while the x-axis depicts the overlaid corresponding protease residues. Ligand-receptor interactions within various distances (Å) are shown by a colour code, with black being the closest while white the farthest.

The fingerprinting profiles depicted loss of interactions in the nose, flap and 80s loop. In V82 γ-methyl group takes part in ligand-receptor interaction, but in 82A, this is erased (Dirauf et al., 2010). Stanford HIVdb interpretation reports that L10I leads does not lead to decreased

susceptibility to SQV. V82A alters shape and size of the pocket in the 80s loop, leading to the loss of interactions (Dirauf et al., 2010). V82A does (Logsdon et al., 2004), but not for NFV and LPV. Here we report that together with V82A and I15V, L19I, K20R, E35D, M36I, R41K, K45KR, C67Y, H69K, L89M in subtype C, it causes high-level resistance, unlike the potential-low level resistance information from Stanford HIV db. The mutations in 52076 had resulted into an expanded active site, due to large Cα backbone shifts (Ali et al., 2010).

V82A/L90M mutation causes contraction of the pocket. Studies have linked this to resistance to RTV but not APV (Dirauf et al., 2010). We observed that in clade C (50459), this synergism together with K20R, E35D, and K45R resulted in an expanded cavity that led to improved RTV binding affinity. E35D occurs in the flap region and causes conformational changes in the residues such as V32 of binding cleft (Kandathil et al., 2009). Binding energy was observed to improve from -12.68 kcal/mol to -14.65 kcal/mol for RTV in 5045 and its mutant. For APV, the score also improved from -12.89 kcal/mol to -13.89 kcal/mol.

Protease mutations can affect the active site and or the non-active sites (Mao, 2011), which have structural and functional effects on the enzyme (van Maarseveen et al., 2012). Both the expansion and the contraction active site hypotheses hold true for some inhibitors. They either favour or hinder the ability of a ligand to effectively bind into the active site. For example, in 301812, 305112, and 315152, the increase in active site volume enhanced the binding of RTV. There is a limit to which RTV can effectively bind in each protease. Polymorphisms, common and rare mutations in clade C appear to have roles. For example, the K70R (in presence of T12S, I15V, L19I, M36I, R41K, H69K, L89M, and I93L) in 526112 leads to a decrease in the volume of the active site. This perturbs the susceptibilities of the protease inhibitors. Susceptibilities levels for LPV, RTV and SQV diminishes, but still within the susceptibility range. With the exception of L63P, the eight wild-type C polymorphisms together with K20R, E35D, I62V, and T74S lead to an increase in pocket volume resistance to IDV according to docking studies, contrary to the Stanford HIVdb. Stanford HIVdb does not underline much attention to the effects of polymorphisms in susceptibilities in non-B clades.

Upset of binding affinity is one of the ways through which mutations affect HIV protease functionality (van Maarseveen et al., 2012). This alteration can either be favourable or unfavourable (Perryman & Lin, 2004) as depicted by increase in changes in binding energies

and inhibition constants (Pazhanisamy et al., 1996). M46I, I54V, and V82A are associated with multiple cross-resistance but not for all the drugs. Accumulation of the polymorphisms in clade C together with specific major and mutations result in susceptibility of the mutant to some inhibitors. Mutations might not be involved directly in upsetting the susceptibilities, but they can act to augment or disrupt the behaviour of residues within certain subsites. Mutations in the substrate binding site alter the binding affinity (Pokorná et al., 2009).

Most of the mutants were also susceptible to both DRV and TPV. TPV is able to target resilient HIV strains. These two inhibitors have different modes of actions. TPV forms hydrogen bonds with the flap tips, I50/I50' and not the active site residues, while the converse pertains to DRV (Heal et al., 2012). This is true for the wild-types B and C (Figure 3-11c, d, f and g). However, both polymorphisms and non-polymorphisms in clade C appear to disrupt these fashions of interactions. 3018, 3051 and 3021 are a few examples of samples with polymorphisms that lack the aforesaid hydrogen interaction manifested in TPV. In the case of DRV in drug-naïve patients, some of the proteases that formed hydrogen bonds with the active site residues include D30 of 3018, D29 of 5086 among others. In drug-exposed patients, they included D29 of 301812 and G27 of 30599 among others. In 3051, D30 and I50' formed hydrogen interactions with the DRV. In 305112 and 305112, the sulfonyl formed hydrogen bond with I50, while in 305152, D25 and D30 participated.

3.6.3.3 General Performance of the FDA-approved and RU-synthesized Ligands

The docking involving these ligands and the various proteases both in closed and open conformations were graphically summarized using energy maps as shown in Figure 3-22 and Figure 3-23 plotted using gnuplot program. The binding energies are represented on a colour-code scale of yellow to black, with black depicting better binding energy scores than yellow. Inhibitors are on the y-axis while proteins are on the x-axis each ordered in the "before and after drug-exposure" arrangement. These maps show the effects of different polymorphisms and or non-polymorphisms on efficacy of two sets of protease inhibitors. The binding energies yield information concerning mutations that confer resistance and viral fitness. Mutations perturb the drug susceptibilities. They may either lead to resistance or susceptibility. For the proteases without mutations after drug-exposure, the energies were reasonably the same, especially in cases where computational challenge was minimized or

evaded. In Figure 3-22, the last column portrays the 1HXB crystallized structure that was utilized to predict the position and orientation of SQV in the protein receptor (Luo & Chan, 2012), and how other inhibitors bound to it. The validation was necessary before the high-throughput simulation could be performed. Docking with other drugs confirmed the capability of *in silico* docking toward locating the active site pocket.

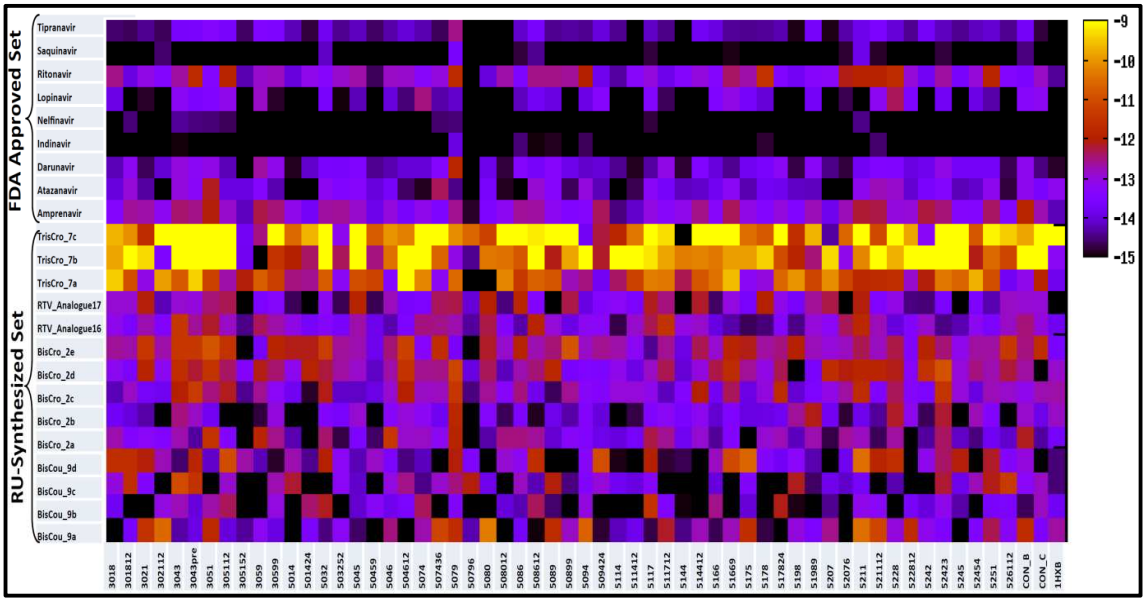


Figure 3-22: Energy maps of 1357 docking results. The binding energies are represented on a colour-code scale of yellow to black, with black being better binding energy scores than yellow. Inhibitors are on the y-axis while proteins are on the x-axis each ordered in the "before and after drug exposure" arrangement.

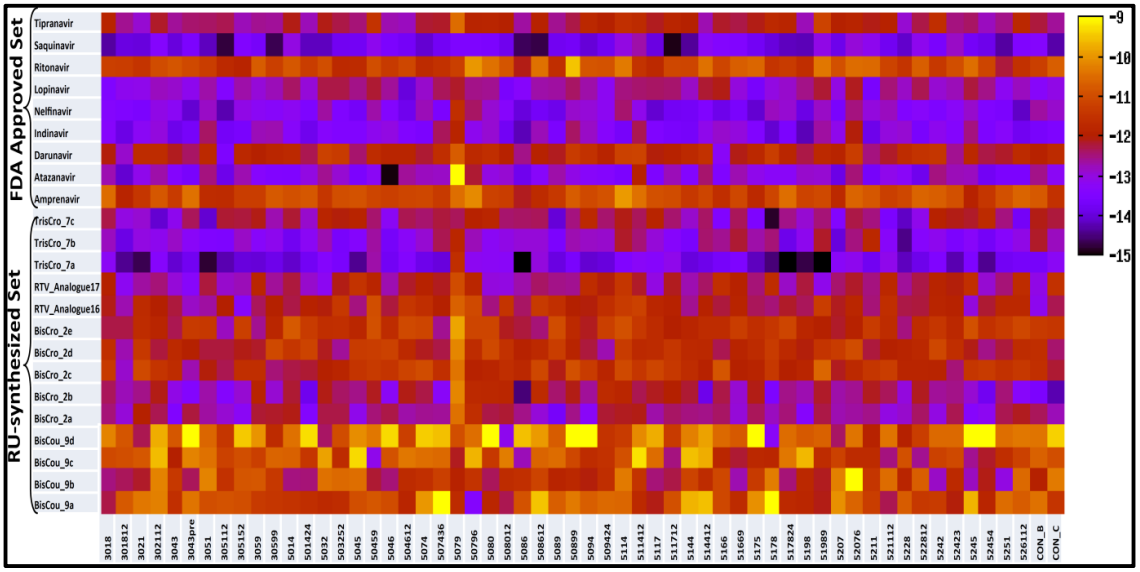


Figure 3-23: Energy maps of 1334 docking results. The binding energies are represented on a colour-code scale of yellow to black, with black being better binding energy scores than yellow. Inhibitors are on the y-axis while proteins are on the x-axis each ordered in the "before and after drug exposure" arrangement.

Docking results of the closed conformation (Figure 3-22) reveal that both the wild-type B and consensus C (on the extreme right excluding the 1HXB) interact quite differently with the inhibitors, due to the already discussed reasons. These lead to different susceptibility degrees between the two clades. The two sets of ligands, FDA approved and RU-synthesized protease inhibitors are associated with different binding energies (and inhibition constants), hence differential binding affinities. This could be due to the different number of energy evaluations used. Due to time limitation, numbers of evaluations were 4.5 million and 450000 for the FDA approved and RU-synthesized ligands. This parameter has a positive influence on the quality of the docking results in terms of the fractions of correct binding sites found (Hetényi & Spoel, 2002).

In spite of the mutations, the FDA approved inhibitors still stand out as good inhibitors due to their consistencies in effectiveness. RTV was the most affected of all among the infant sequences. LPV was better. Compared to LPV, RTV is a bulky inhibitor hence cannot fit into the receptor site (Dirauf et. al., 2010). Again, if the receptor pocket is greatly large, it loses interactions. Among the RU-synthesized potential inhibitors, the *tris*-chromone analogues appear to be the best. But again even within this class of inhibitors, there are efficacious disparities. Those whose substitution groups are either bromide (Br) or chloride (Cl) at position 1 (with the rest of the variable being H), seem to be better than the rest in the class. *Bis*-coumarin can also have good inhibitory effects. The subclass with hydrogen (H) or Cl substitutions at variable group 2 were the best among this coumarin class. The RTV analogues behaved similar to RTV. The RTV_Analogue17 that has a thiol coumarin was associated with better binding energy. *Tris*-chromone analogues were the poorest. This is because they are the bulkiest thus could not fit into the pocket. Compared to *bis*-chromone, it has one extra chromone moiety. In a resource-limited setting, a good protease inhibitor would be the one with broad-spectrum activity. So far, only the *bis*-coumarin and *bis*-chromone analogues might qualify for this. The extended structures of the RU-synthesized inhibitors favour their effective bindings.

The maps pertaining to the ligands docked to the open conformation (Figure 3-23) had lower binding energies compared to those docked to the closed conformation. Among the ligands, the *tris*-chromone had better binding energies in protease with the open conformation than the closed conformation. These ligands are bulky and fit well in

proteases in the open conformation. Despite its size, RTV could not effectively bind (Dirauf et al., 2010) and is significantly affected by mutations. Ligands usually bind to the open conformation before the closed conformation is attained. According to the two maps, the *tris*-chromone derivatives are presently not good inhibitors because they bind poorly in the closed conformation thus might have the tendency to escape from the active site cleft. The *bis*-coumarin analogues have binding energy values that when related to those of the FDA approved protease inhibitors, are termed as "high". The *bis*-chromone with H and Cl groups at the R^2 substitutable groups had binding energies almost comparable to those of the universally acceptable PIs.

A good inhibitor is one with a broad-spectrum activity (Ohtaka & Freire, 2005). In the context of the HIV-1 protease enzyme, a good inhibitor is also expected to have a higher binding energy (more negative) in the closed form than in the open form. The *tris*-chromone analogues therefore are not good inhibitors unless further modified. Because of the mutations, efficacious drugs can be effectuated should they be engineered to mimic the substrate shape, have an extended form, and interact with the invariant protease residues (Prabu-Jeyabalan et al., 2002) without the loss and rapture of hydrogen bond contacts during dimensional alterations of the protease architecture (Li et al., 2011). The formulated energy maps display both the efficacy and cross-reactivity property of the FDA approved PIs against several mutants. Even though some of these RU-synthesized inhibitors can only be utilized in personalized medicine (individualized therapy), in a resource-limited setting like the sub-Saharan subcontinent where the incidence is high, the broad-spectrum medications would be of vast relevance. For example, *tris*-chromone with the Cl R^2-functionality can be used in second-line therapy on infant 3021, but mostly not for others. Modification would be required to reduce the size of the *tris*-chromone class. On the other hand, *bis*-coumarin would require adjustments to improve on its initial binding before the attainment of the closed conformation.

Both consistencies and inconsistencies of the binding mode of the two ligands sets can be evaluated by studying their interactions with the proteases, and noting any common bonds that occur throughout these interactions. We determined whether or not the ability of these ligands to occupy the binding sites were the same or different. This was performed by overlaying the same ligand with the lowest binding energies in across the proteases (Figure

3-24). Superposition of the backbone atoms of the ligands depicted several modes of binding, although a mode of binding may appear quite different to the symmetrical case where the binding is identical. Deviations arose where the ligand failed to effectively bind can due to the mutational effects that affected the behaviour of subsites residues (Ali et al., 2010) or due to the ligand size with respect to the pocket.

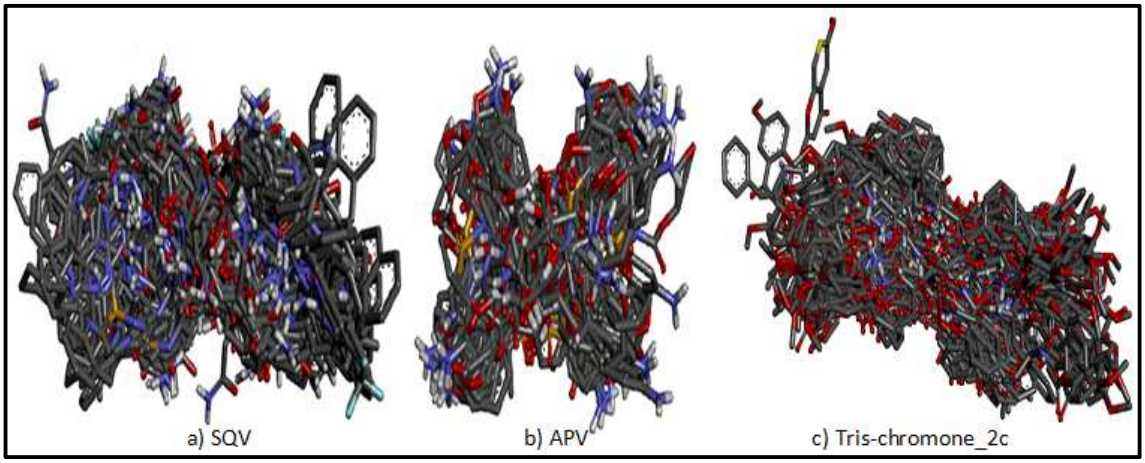


Figure 3-24: Binding of selected ligands showing the binding consistencies. Figures a), b), and c) each represent an overlay of the same ligand (with the lowest energy) each from the different protease dockings. These overlays depicted the symmetrical nature of the protease and the binding consistency of the same ligand in different proteases. Some of the ligands were not in alignment due to mutations in the protease, hence the distorted binding.

The symmetrical nature of the HIV-1 protease (Li et al., 2011) is supported by these overlays that show the occupancies of the binding site among these modes (Ali et al., 2010). HIV protease usually loses its symmetry in its advent of binding to an asymmetric substrate (Prabu-Jeyabalan et al., 2002). Shape complementarity is important in recognition and binding of substrates. But specificity is believed to lie within the asymmetric nature of the substrates which is reinforced via the hydrogen bonds (Prabu-Jeyabalan et al., 2002). This appears to be factual since the ligands are not proteins yet they became bound to the receptor. Binding affinity is enhanced when there is good ligand fitting and hydrogen bonds are involved in the ligand-receptor association, as already seen in some of the docking cases in this study. The RU-synthesized ligands presented similar occupancies of the pocket as those of the FDA approved ligands.

3.6.3.4 Evaluation of RU-synthesized Ligands as Inhibitors

Since the RU-synthesized ligands are among the potential drugs in the drug pipeline, they were analyzed for ligand-receptor interactions. Only a few interactions have been illustrated

here (Supplementary Data/Chapter III/Docking/Interactions/(CLOSED_CON_C-Interactions and OPEN_CON_C-Interactions). Virtual screening by docking allows molecules to be evaluated for better ligand-receptor kinetics with an aim to be used as drugs. It is fast, cheap and often yields reliable results (Laurie & Jackson, 2006). Interaction analyses were to aid in effective development of these potential protease inhibitors. Only interactions with consensus C were used. This was to eliminate bias in future treatment possibly utilizing such inhibitors. If the redesign is tailored within the framework of this consensus, then it would be possible to be used across clade C mutants. After the initial ligand adjustments, further amendments can be done using the ligand-mutant protease interactions. Assessment of potential drugs uses crucial interactions in drug design (Luo & Chan, 2012).

For explanation purposes, the RU-synthesized ligands will by referred by their codes. Full information by Kaye and colleagues (2004) pertaining to these ligands are accessible at http://www.tandfonline.com/doi/pdf/10.1081/SCC-200025617. Among the *bis*-coumarin analogues, it is BisCou_9a that did fit well enough in the active site cleft of the protease in the open conformation. When bound to the closed protease, BisCou_9a and BisCou_9b each had one of its coumarin outside the pocket. Unlike their counterparts, BisCou_9c and BisCou_9d had good bind (with that of BisCou_9c being the best), and extended well in the active site cleft. All the *bis*-chromone analogues had good fittings in the open protease with the exclusion of BisCro_9b whose occupancy in the pocket was not extensive. In the closed conformation of the enzyme, BisCro_9c was “overextended” in the binding site, hence not expected to interact productively (with many residues). For the RTV analogues, RTV_Analogue16 presented a similar scenario like BisCro_9a. Binding of the RTV_Analogue17, a thiol-containing chromene, was good but this ligand did not extend in the active site cleft effectively in the open conformation of the protease. In the closed conformation, they all had good binding. Among the *tris*-chromone derivatives, it is only TriCro_7b that exhibited good binding and extension in the open conformation of the protease. For instance, for TriCro_7c, approximately the entire enzyme pocket was occupied by this inhibitor. All the *tris*-chromone analogues fitted poorly in the pocket of the protease.

Figure 3-25 shows the interaction of BisCou_9a bound to protease C. Atomic interactions unravel its ability to interact with both the active site (G27 and D29) and flap residues (I50’) (of the closed form of protease) via the hydrogen bonds, unlike in the open conformation

where only the nose (R8) and active site (G27') residues form hydrogen bonds. It also forms both cation-pi and sigma bonds. In the closed conformation, the amino acid residues involved in interactions are clustered in close proximity around the ligand. This is can also be visualized in the fingerprints profiles. As the flap closes curls to boost protease functionality (Li et al., 2011), amino acid residues in the protease binding site begin to associate with the ligand, thereby forming new contacts. Some interactions are lost as better ones are formed. The structural water could be useful in augmenting the binding of this ligand.

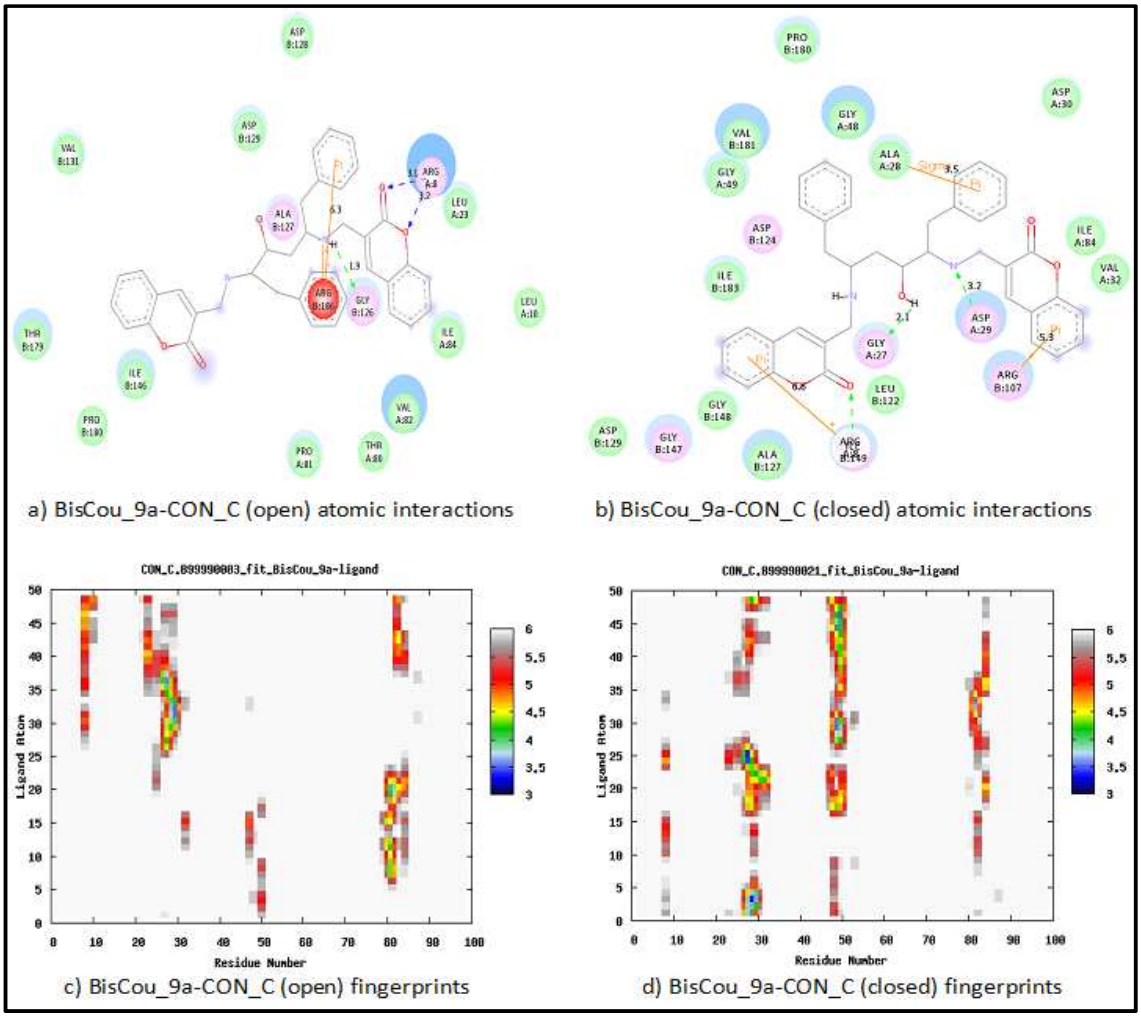


Figure 3-25: Interaction profiles between BisCou_9a and CON_C (both open and closed conformations) using Discovery Studio Visualizer and fingerprints. For a) and b) protease residues involved in the interactions with the ligand atoms are shown by coloured discs. Residues with coloured diffuse background indicate the solvent accessible areas. For c) and d) the y-axis depicts ligand atoms while the x-axis depicts the overlaid corresponding protease residues. Ligand-receptor interactions within various distances (Å) are shown by a colour code, with black being the closest while white the farthest.

The rest of the *bis*-coumarin analogues interact with the active site via hydrogen bonds, with the bonds being enhanced in the closed form of protease. BisCou_9d also forms sigma pi interactions with I50', which is a strong covalent bond, hence the better binding energy.

Bis-chromone analogues especially BisCro_9a and BisCou_9b rarely form hydrogen bonds with the protease, and if they do, then they are few. BisCro_9b only forms a cation-pi bond with the protease with opened flap but only nonbonded contacts exist in the closed form. BisCro_2c has one pi and hydrogen bond involving R8. When in the closed protease structure, sigma bond, pi bonds, hydrogen bonds and nonbonded contacts emerge. For BisCro_9d, nonbonded contacts prevail when bound to the open protease structure. Direct interaction with the closed protease exploits the D29' residue. BisCro_2e has hydrogen bonding with D30 and G48 with the open protease. In closed protease, there a hydrogen bond and a pi contacts involve G48 and R8, respectively. BisCro_2a depends on the nonbonded contacts for interaction (Figure 3-26). Despite this, this ligand appears more promising. Modifications directed toward it should involve developing ways for it to interact via hydrogen (and or covalent bonds). Hydrogen bonds also play a fundamental role in specificity (Prabu-Jeyabalan et al., 2002).

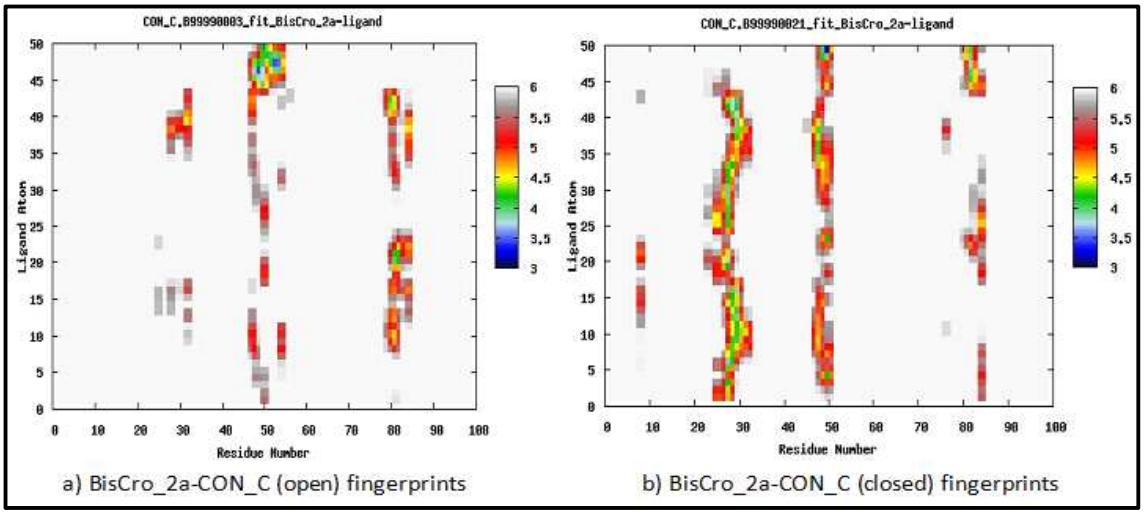


Figure 3-26: Interaction profiles between BisCro_2a and CON_C (both open and closed conformations) using fingerprints. The y-axis depicts ligand atoms while the x-axis depicts the overlaid corresponding protease residues. Ligand-receptor interactions within various distances (Å) are shown by a colour code, with black being the closest while white the farthest.

The RTV analogues are ligands with chromene functionalities are believed to interact with the ligand and structural water via hydrogen bonding or undergo nucleophilic cleavage that will lead to covalent bond formation with the protease receptor (Kaye et al., 2004). The docking lacked structural water. RTV_Analogue16 atoms interact using sigma bond with residues in the 80s loop, either V82 or I84. Interactions in the nose region vary in the two protease conformation. Closed conformation has minimal interactions. In open protease conformation, RTV_Analogue17 interacts with the T80 via hydrogen bond. Minimal contacts

with the nose are visualized. In closed conformation, hydrogen bond with R8 and cation-pi bond with D25 are seen to interact with the ligand. The RTV analogues have similar binding modes. In closed protease, RTV_Analogue17 has more interactions with the nose and active site than RTV_Analogue16.

In *tris*-chromone derivatives, as the complex acquires the closed conformation, there are losses of hydrogen bond and or interactions with the aromatic ring of ligands probably due to the flap forcibly curling into the active site. In TriCro_7a (Figure 3-27), there is overcrowding of some residues (of the closed protease) around the ligand. Such contacts may have high energies. To compensate for this high energy cost due to unfavourable binding, the binding energy has to reduce.

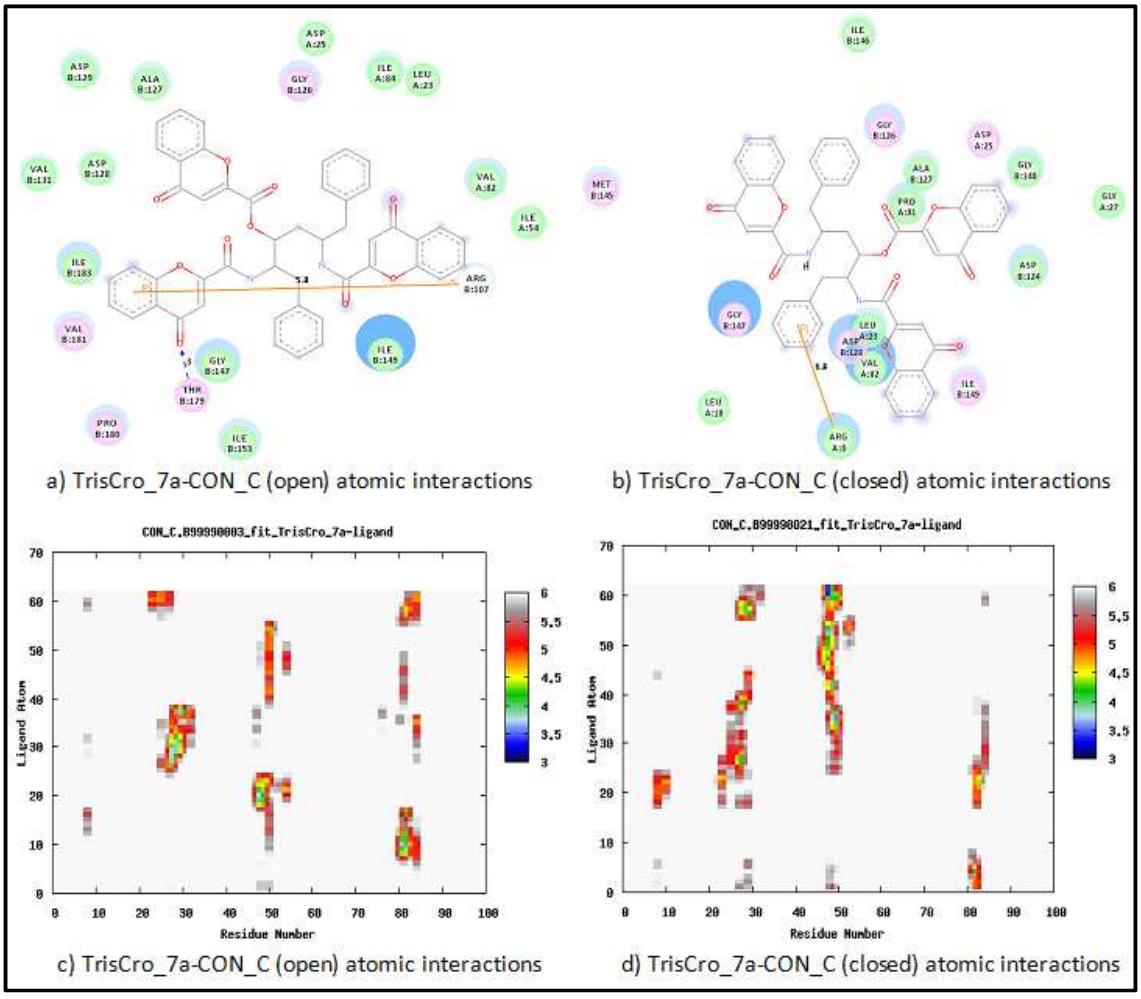


Figure 3-27: Interaction profiles between TrisCro_7a and CON_C (both open and closed conformations) using Discovery Studio Visualizer and fingerprints. For a) and b) protease residues involved in the interactions with the ligand atoms are shown by coloured discs. Residues with coloured diffuse background indicate the solvent accessible areas. For c) and d) the y-axis depicts ligand atoms while the x-axis depicts the overlaid corresponding protease residues. Ligand-receptor interactions within various distances (Å) are shown by a colour code, with black being the closest while white the farthest.

The pharmaceutical chemists should therefore maximize on the ligands with the extended forms that can spread across the active site. Those inhibitors not utilizing hydrogen bonding or have minimal hydrogen bonds in their interactions with the protease should be tailored to manifest these interactions, especially to involve the D25/25', I50/50' and at least two invariant residues in the 80s loop so that they would be still effective against mutants. The BisCro_2a and BisCro_2b are good inhibitors based on their favourable binding energies in both protease conformations. They would need greater tunings prior to their use. *De novo* ligand synthesis can also be utilized from this point, taking different ligand substructure with good interactions (in terms of energies and distances) and assembling them. Due to the adaptability nature of the HIV-1 protease, the ligands should be made adaptable.

3.6.4 Molecular Dynamics Simulations

Mutations do not occur haphazardly and molecular dynamics has endeavoured to demystify how mutations at the structural level affect the thermodynamics (Yu et al., 2011). The HIV-1 molecular flaps have been shown to undergo motion to a completely opened conformation during a 10-ns simulation in solvated system. Compared to the wild-type protease receptor, molecular dynamics using a crystallized HIV-1 multidrug resistant variant, isolate 769 showed that the flap tip residues were more distantly placed in the mutant (Logsdon et al., 2004). Flaps must be able to open up to allow binding to substrate (Chang et al., 2007). HIV-1 protease dynamics is nowadays extensively and intensively studied because of its use in drug engineering and behaviour in mutated proteases (Yu et al., 2011).

In this study, the docked ligands were combined with the protein, followed by dynamics simulations for each protein-ligand ensemble as described in the methodology. Explicit solvation is understood to be essential (Perryman & Lin, 2004), and so simulations included an explicit water model. Figure 3-28 shows the periodic box of one the neutralized systems, ATV-3018, to be used in the stages of minimization, heating to 310 K, equilibration and production dynamics for analysis. The periodic water condition permits the simulation of the dynamics as it sets the platform for mimicking the *in vivo* binding (Dirauf et al., 2010). Insufficient equilibration may cause large conformational changes at the beginning of the HIV-1 simulations (Hornak et al., 2006a). HIV-1 protease function is due to the conformational changes that it undergoes (Mao, 2011).

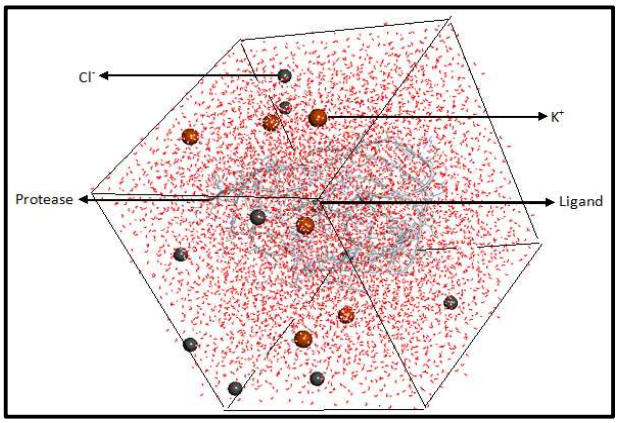


Figure 3-28: Protease structure in periodic boundary condition. Water has been used explicitly. Potassium (K^+) and chloride (Cl^-) have been used to neutralize the system.

As a preliminary analysis, the flap motions during dynamics of 3018 and 30182 with ATV, NFV and IDV were studied and provide a unique insight into the differences between the mutants. Figure 3-29 indicates the appearance of the flap region at the end of the dynamics. The flap in 301812, I54V affected the flap region thereby interfering with its flexibility; hence it could not fully close. Analysis of 3018 and 301812 with ATV showed a wide range of motions associated with the accommodation and stabilization of the inhibitor within the active site (Figure 3-30). The y-axis indicates the two flap distance with respect to G48 whereas the x-axis indicates the course of dynamics (after every 100 steps). When a ligand binds to protease receptor, profound changes occur in the flap region (Alcaro et al., 2009). Flap motions also generate a wide range of conformations (Hornak et al., 2006a). The flap dynamics course showed that the flap region of 301812 was more open compared to 3018.

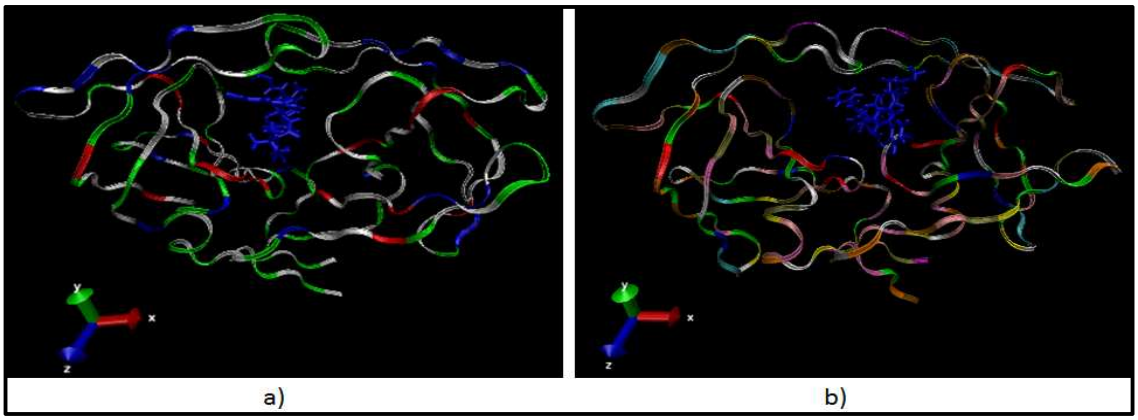


Figure 3-29: Snapshots at the end of molecular dynamics. a) ATV-3018 and b) ATV-301812. In both cases, the ligands are depicted as coloured licorice, while the proteases are represented by chains coloured by residue. The flap regions of the two complexes (a and b) are different, with that of b) being more wide open than that of a), hence the resistance.

At the start of dynamics, the flaps are usually opened but later fall close to the catalytic site (Hornak et al., 2006a).

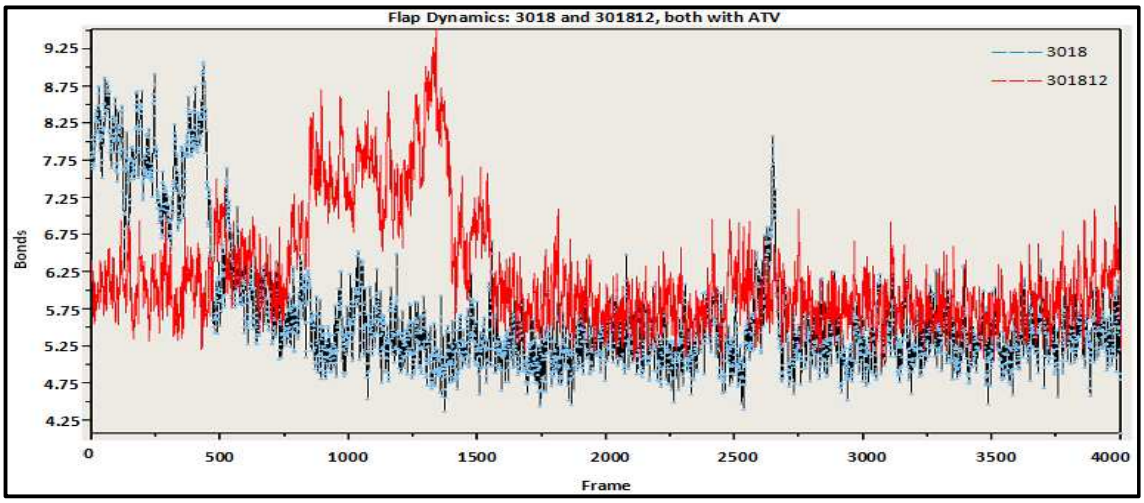


Figure 3-30: Interatomic distance of the Cα of G48 and G48' of the flap region. The graph indicates the distances (y-axis) between G48 and G48' after every 100 steps of molecular dynamics simulations (x-axis). Compared to sample 3018 (in **blue**), sample 301812 (in **red**) has the two molecular flaps widely spaced, hence its resistance to ATV.

In dynamics, changes occur at active site to accommodate the ligand (Zhu et al., 2003). Atomic interactions for these samples have been shown in Figure 3-18 and discussed thereafter. Further studies will involve simulations on the open conformations and observation of the flap regions during dynamics.

3.7 CONCLUSION

HIV-1 protease evolution is caused by mutations. They endow the virus with fitness either by either extending or contracting the binding site through positional shifts. These changes interfere with the drug binding by imposing steric clashes in the protease pocket. Computational drug design is one of the most effective platforms that permit screening of compounds for potential drug use. Good binding fit is not necessarily linked to high binding energies. Even though the RU-synthesized ligands have appreciably good consistencies in the pocket occupancies as compared to the FDA approved protease inhibitors; there is still the need to engineer them so that they can also "evolve" with the virus. Unlike in subtype B, some mutations in subtype C protease respond differently to drugs. In such cases, the virus is either susceptible or resistant to the inhibitors. This is an indication that the two clades behave differently in some instances in order to enforce drug resistance. The Stanford HIVdb does not consider polymorphism in clade C in its interpretations. However, validation, phenotyping and site-directed mutagenesis studies should be performed in future to support this. Surface topographical measurements, docking, molecular dynamics and ligand-receptor interactions are vital in unravelling the mechanisms leading to drug-resistances.

CHAPTER FOUR

4. OVERALL DISCUSSION, CONCLUSION AND FUTURE SCIENTIFIC DIRECTION

Both synonymous and non-synonymous mutations can only be defined at the amino acid level. The degeneracy nature of the genetic code accounts for the silent mutations. Inter-subtype variability in HIV-1 is caused by an accumulation of non-synonymous mutations. Most of these non-synonymous mutations appear to preserve the physicochemical properties of the amino acid residues, thereby ensuring that the viral functionality is still guaranteed. At the moment, patients are still limited to the drug alternatives, hence drugs to which the mutants confer intermediate resistance to can still be used as long as they exhibit a high genetic barrier to resistance (Tang et al., Shafer 2012). Some mutations in clade C have different fashions of acting. If the mutation patterns can be predicted, then the changes in active site too in terms of expansion and contraction can be predicted, hence better drugs with conformational propensities (Ohtaka & Freire, 2005) can be developed to restrict infections even from HIV variants. But when mutations occur together, it is actually difficult to study the effect of a single mutational change due to the possibility of synergism (Dirauf et al., 2010). Limited information subsists for the drug resistance mechanisms in non-clade B strains, especially for the PIs (Martinez-Cajas et al., 2012), therefore we believe that we have added value existing data.

With continued success in HAART, drug resistance mutations still surface (Morand-Joubert et al., 2006). Protease mutations can occur at the active site and or non-active sites (Mao, 2011). Active site mutations perturb binding affinity (Morand-Joubert et al., 2006) via conformational changes (Perryman & Lin, 2004). Non-active site non-synonymous mutations affect binding affinity and enzyme turnover rate (Olsen et al., 1999). They cause structural changes through side chains rearrangements, leading to multidrug resistance (Prabu-Jeyabalan et al., 2002). Mutations emerge gradually but rapidly increase upon therapy switch where pre-existing minority population may be selectively favoured (Morand-Joubert et al., 2006). The stoppage antiretroviral therapy models recommended by Taylor and colleagues (2007) should be implemented to maximize on the limited HAART.

Docking and molecular dynamics in conjunction with fingerprints can be utilized as discovery tools to show critical ligand-receptor interaction patterns that can be used to guide drug

design (Luo & Chan, 2012). The flaps have minimal influence on substrate specificity, they majorly aid in protease dynamics (Perryman & Lin, 2004). These findings on characterization of drug resistance molecular mechanisms are not only crucial to the medicinal chemists dealing with HIV-1 subtype C proteases, but can be employed by other drug designers as heuristic models to engineer regimens that target other structural entities of tenacious microbes. Current inhibitors usually interact via hydrogen bond with the active site and or flap residues (Heal et al., 2012). To synthesize novel inhibitors, the other invariant subsite residues of the protease should be utilized. Dimerization enhances the protease stability, therefore an inhibitor also distorting this interface should be a commendable drug (Kandathil et al., 2009). Paracelsus, the father of toxicology, once said, *"Everything is poison, there is poison in everything, only the dosage makes a thing not to be poison."* Therefore, before the adjusted RU-synthesized ligands can eventually be dispensed affordably into the patient shelves, a lot of efforts need to be experimentally inputted in order to establish their effectiveness, residual effects and toxicities.

Finally, we suggest that docking algorithm can be an effective and affordable platform for carrying out drug resistance measurements once it has clear cut-offs for defining susceptibilities. These can be used in surveillances of mutations in resource-limited settings. Future endeavours should revolve around performance of high-throughput molecular dynamics, validation and phenotyping studies. For the prediction(s) of pattern(s) of occurrence(s) of mutations, prospective or retrospective studies entailing relatively long follow-ups on the HIV-1 clade C-infected subjects (maximizing on the same therapy) should be carried out, and better models such as the Bayesian model used to compute these predictions. In concurrent with vaccine and drug design and development, the sub-Saharan region ought to construct a database customized for its use, if it must have confidence in its drug resistance mutations surveillances, determination of resistance baselines and management of the deadly scourge, which continues to claim scores of lives, in a world that has failed to fully put into practice sexual abstinence by choice until marriage, or being faithful to one's partner.

5. REFERENCES

Abbas, W., & Herbein, G. (2012). Molecular Understanding of HIV-1 Latency. *Advances in Virology*, 1-14.

Abecasis, A. B., Lemey, P., Vidal, N., De Oliveira, T., Peeters, M., Camacho, R., Shapiro, B., et al. (2007). Recombination Confounds the Early Evolutionary History of Human Immunodeficiency Virus Type 1: Subtype G is a Circulating Recombinant Form. *Journal of Virology*, *81*(16), 8543-51.

Africa, S. S. (2009). *South African Statistics, 2009* (pp. 1-202).

Africa, S. S. (2011). *Statistical Release - Mid-Year Population Estimates 2010* (pp. 1-16).

Africa, S. S. (2012). *Statistical Release - Mid-Year Population Estimates* 2011 (pp. 1-18).

Alcaro, S., Artese, A., Ceccherini-Silberstein, F., Ortuso, F., Perno, C. F., Sing, T., & Svicher, V. (2009). Molecular Dynamics and Free Energy Studies on the Wild-Type and Mutated HIV-1 Protease Complexed with Four Approved Drugs: Mechanism of Binding and Drug Resistance. *J Chem Inf Model*, *49*(7), 1751–61.

Ali, A., Bandaranayake, R. M., Cai, Y., King, N. M., Kolli, M., Mittal, S., Murzycki, J. F., et al. (2010). Molecular Basis for Drug Resistance in HIV-1 Protease. *Viruses*, *2*, 2509-2535.

Araya, T., Tensou, B., Davey, G., & Berhane, Y. (2011). Burial Surveillance Detected Significant Reduction in HIV-related deaths in Addis Ababa, Ethiopia. *Tropical Medicine & International Health*, *16*(12), 1483-1489.

Archana, P., Sathishkumar, N., & Bharathi, N. (2010). In Silico Docking Analysis of Curcumin - An Inhibitor for Obesity. *International Journal of Pharma and Bio Sciences*, *1*(4), 224-235.

Archary, D., Gordon, M. L., Green, T. N., Coovadia, H. M., Goulder, P. J. R., & Ndung'u, T. (2010). HIV-1 Subtype C Envelope Characteristics Associated with Divergent Rates of Chronic Disease Progression. *Retrovirology*, *7*(1), 92.

Arnott, A., Jardine, D., Wilson, K., Gorry, P. R., Merlin, K., Grey, P., Law, M. G., et al. (2010). High Viral Fitness during Acute HIV-1 Infection. *PloS One*, *5*(9).

Artimo, P., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, G., De Castro, E., Duvaud, S., et al. (2012). ExPASy: SIB Bioinformatics Resource Portal. *Nucleic Acids Research*, *40*(Web Server Issue), W597-603.

Arts, E. J., & Hazuda, D. J. (2012). HIV-1 Antiretroviral Drug Therapy. *Cold Spring Harbor Perspectives in Medicine*, *2*(4), a007161.

Baldwin, E. T., Bhat, T. N., Gulnik, S., Liu, B., Topol, I. A., Kiso, Y., Mimoto, T., et al. (1995). Structure of HIV-1 Protease with KNI-272, a Tight-Binding Transition-State Analog Containing Allophenylnorstatine. *Structure London England 1993*, *3*(6), 581-590.

Barbour, J. D., & Grant, R. M. (2005). The Role of Viral Fitness in HIV Pathogenesis. *Current HIV/AIDS Reports*, *2*(1), 29-34.

Benkert, P., Künzli, M., & Schwede, T. (2009). QMEAN server for Protein Model Quality Estimation. *Nucleic Acids Research*, *37*(Web Server Issue), W510-4.

Benko, D. M., Schwartz, S., Pavlakis, G. N., & Felber, B. K. (1990). A Novel Human Immunodeficiency Virus Type 1 Protein, tev, Shares Sequences with tat, env, and rev Proteins. *Journal of Virology*, *64*(6), 2505-18.

Bessong, P. O. (2008). Polymorphisms in HIV-1 subtype C proteases and the Potential Impact on Protease Inhibitors. *Tropical Medicine and International Health*, *13*(2), 144-151.

Bhat, S., & Purisima, E. O. (2006). Molecular Surface Generation using a Variable-Radius Solvent Probe. *Proteins*, *62*(1), 244-61.

Binkowski, A. T., Naghibzadeh, S., & Liang, J. (2003). CASTp: Computed Atlas of Surface Topography of proteins. *Nucleic Acids Research*, *31*(13), 3352-3355.

Boudová, S., Li, H., Sajadi, M. M., Redfield, R. R., & Pauza, C. D. (2012). Impact of Persistent HIV Replication on CD4 Negative Vγ2Vδ2 T Cells. *The Journal of Infectious Diseases*, *205*(May), 1448-1455.

Brik, A., & Wong, C. (2003). HIV-1 Protease: Mechanism and Drug Discovery. *Org. Biomol. Chem.*, *3*(1), 5-14.

Briz, V., Poveda, E., & Soriano, V. (2006). HIV Entry Inhibitors: Mechanisms of Action and Resistance Pathways. *The Journal of Antimicrobial Chemotherapy*, *57*(4), 619-27.

Bulla, I., Schultz, A. K., Schreiber, F., Zhang, M., Leitner, T., Korber, B., Morgenstern, B., et al. (2010). HIV Classification using the Coalescent Theory. *Bioinformatics (Oxford, England)*, *26*(11), 1409-15.

CARMMA. (2012). *South Africa's National Strategic Plan For a Campaign On Accelerated Mortality In Africa (CARMMA)* (pp. 1-18).

Caroline, A., & Andrew, N. (2009). Should HIV Therapy be Started at a CD4 Cell Count above 350 cells/µl in Asymptomatic HIV-1-Infected Patients? [Special commentary]. *Current Opinion in Infectious Diseases*, *22*(April), 1-11.

Ceballos, A., Remes Lenicov, F., Sabatté, J., Rodríguez Rodrígues, C., Cabrini, M., Jancic, C., Raiden, S., et al. (2009). Spermatozoa Capture HIV-1 through Heparan Sulfate and Efficiently Transmit the Virus to Dendritic Cells. *The Journal of Experimental Medicine*, *206*(12), 2717-2733.

Chandwani, A., & Shuter, J. (2008). Lopinavir/Ritonavir in the Treatment of HIV-1 Infection: A Review. *Therapeutics and Clinical Risk Management*, *4*(5), 1023-1033.

Chang, C. E. a, Trylska, J., Tozzini, V., & McCammon, J. A. (2007). Binding Pathways of Ligands to HIV-1 Protease: Coarse-Grained and Atomistic Simulations. *Chemical Biology & Drug Design*, *69*(1), 5-13.

Coffin, J. M. (1995). HIV Population Dynamics *in vivo*: Implications for Genetic Variation, Pathogenesis, and Therapy. *Science*, *267*(5197), 483-489.

Cohen, M. S., Shaw, G. M., McMichael, A. J., & Haynes, B. F. (2011). Acute HIV-1 Infection. *The New England Journal of Medicine*, *364*(20), 1943-1954.

Collins, J. R., Burt, S. K., & Erickson, J. W. (1995). Flap Opening in HIV-1 Protease Simulated by "Activated" Molecular Dynamics. *Nature Structural Biology*, *2*(4), 334-338.

Cunningham, A. L., Harman, A. N., & Donaghy, H. (2007). DC-SIGN "AIDS" HIV Immune Evasion and Infection. *Nature Immunology, 8*(6), 556-8.

Dalai, S. C., De Oliveira, T., Harkins, G. W., Kassaye, S. G., Lint, J., Manasa, J., Johnston, E., et al. (2009). Evolution and Molecular Epidemiology of Subtype C HIV-1 in Zimbabwe. *AIDS (London, England)*, *23*(18), 2523-32.

Das, A., Prashar, V., Mahale, S., Serre, L., Ferrer, J. L., & Hosur, M. V. (2006). Crystal Structure of HIV-1 Protease *in situ* Product Complex and Observation of a Low-Barrier Hydrogen Bond between Catalytic Aspartates. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(49), 18464-9.

De Oliveira, T., Deforche, K., Cassol, S., Salminen, M., Paraskevis, D., Seebregts, C., Snoeck, J., et al. (2005). An Automated Genotyping System for Analysis of HIV-1 and other Microbial Sequences. *Bioinformatics (Oxford, England)*, *21*(19), 3797-800.

Deeb, O., & Jawabreh, M. (2012). Exploring QSARs for Inhibitory Activity of Cyclic Urea and Nonpeptide-Cyclic Cyanoguanidine Derivatives HIV-1 Protease Inhibitors by Artificial Neural Network. *Advances in Chemical Engineering and Science*, *2012*(January), 82-100.

Di Luccio, E., & Koehl, P. (2011). A Quality Metric for Homology Modeling: the H-factor. *BMC Bioinformatics*, *12*(1), 1-19.

Dibrov, A., Myal, Y., & Leygue, E. (2009). Computational Modelling of Protein Interactions: Energy Minimization for the Refinement and Scoring of Association Decoys. *Acta Biotheoretica*, *57*(4), 419-428.

Dirauf, P., Meiselbach, H., & Sticht, H. (2010). Effects of the V82A and I54V Mutations on the Dynamics and Ligand Binding Properties of HIV-1 Protease. *Journal of Molecular Modeling*, *16*(10), 1577-83.

Douek, D. C., Roederer, M., & Koup, R. A. (2009). Emerging Concepts in the Immunopathogenesis of AIDS. *Annual Review of Medicine*, *60*(October 2008), 471-484.

Dragic, T., Litwin, V., Allaway, G. P., Martin, S. R., Huang, Y., Nagashima, K. A., Cayanan, C., et al. (1996). HIV-1 Entry into $CD4^+$ Cells is Mediated by the Chemokine Receptor CC-CKR-5. *Nature*, *381*(6584), 667-673.

Dundas, J., Ouyang, Z., Tseng, J., Binkowski, A., Turpaz, Y., & Liang, J. (2006). CASTp: Computed Atlas of Surface Topography of Proteins with structural and Topographical Mapping of Functionally Annotated Residues. *Nucleic Acids Research*, *34*(Web Server issue), W116-8.

Durrant, J. D., & McCammon, J. A. (2011). Molecular Dynamics Simulations and Drug Discovery. *BMC Biology*, *9*(71), 2-9.

Freedberg, D. I., Ishima, R., Jacob, J., Wang, Y., Kustanovich, I., Louis, J. M., & Torchia, D. A. (2002). Rapid Structural Fluctuations of the Free HIV Protease Flaps in Solution: Relationship to Crystal Structures and Comparison with Predictions of Dynamics Calculations. *Protein Science*, *11*, 221-232.

Frohoff, C., Moodley, M., Fairlie, L., Coovadia, A., Moultrie, H., Kuhn, L., & Meyers, T. (2011). Antiretroviral Therapy Outcomes in HIV-Infected Children after Adjusting Protease Inhibitor Dosing during Tuberculosis Treatment. *PloS One*, *6*(2), 1-8.

Gao, F., Vidal, N., Li, Y., Trask, S. a, Chen, Y., Kostrikis, L. G., Ho, D. D., et al. (2001). Evidence of Two Distinct Subsubtypes within the HIV-1 Subtype A Radiation. *AIDS Research and Human Retroviruses*, *17*(8), 675-88.

Gilbert, P. B., McKeague, I. W., Eisen, G., Mullins, C., Guéye-NDiaye, A., Mboup, S., & Kanki, P. J. (2003). Comparison of HIV-1 and HIV-2 Infectivity from a Prospective Cohort Study in Senegal. *Statistics in Medicine*, *22*(4), 573-93.

Goodsell, D. S., Morris, G. M., & Olson, A. J. (1996). Automated Docking of Flexible Ligands: Applications of AutoDock. *Journal of Molecular Recognition*, *9*(November 1995), 1-5.

Gowthaman, U., Jayakanthan, M., & Sundar, D. (2008). Molecular Docking Studies of Dithionitrobenzoic Acid and its Related Compounds to Protein Disulfide Isomerase: Computational Screening of Inhibitors to HIV-1 Entry. *BMC Bioinformatics*, *9 Suppl 12*, S14.

Guex, N., Peitsch, M. C., & Schwede, T. (2009). Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: A Historical Perspective. *Electrophoresis*, *30*(S1), S162-S173.

Habasque, C., Aubry, F., Jégou, B., & Samson, M. (2002). Study of the HIV-1 Receptors CD4, CXCR4, CCR5 and CCR3 in the Human and Rat Testis. *Molecular Human Reproduction*, *8*(5), 419-425.

Hahn, B. H., Shaw, G. M., De Cock, K. M., & Sharp, P. M. (2000). AIDS as a Zoonosis: Scientific and Public Health Implications. *Science (New York, N.Y.)*, *287*(5453), 607-14.

He, J., Chen, Y., Farzan, M., Choe, H., Ohagen, A., Gartner, S., Busciglio, J., et al. (1997). CCR3 and CCR5 are Co-Receptors for HIV-1 Infection of Microglia. *Nature*, *385*(6617), 645-649.

Heal, J. W., Wells, S. a, Jimenez-Roldan, E., Freedman, R. F., & Römer, R. a. (2011). Rigidity Analysis of HIV-1 Protease. *Journal of Physics: Conference Series*, *286*, 012006.

Heal, J. W., Wells, S. A., Freedman, R. B., & Römer, R. A. (2012). Inhibition of HIV-1 Protease: the Rigidity Perspective. *Structural Bioinformatics*, *28*(3), 350-357.

Henikoff, S., & Henikoff, J. G. (1992). Amino Acid Substitution Matrices from Protein Blocks. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(22), 10915-9.

Hetényi, C., & Spoel, D. V. an Der. (2002). Efficient Docking of Peptides to Proteins without Prior Knowledge of the Binding Site. *Protein Science*, *11*, 1729-1737.

Hioe, C. E., Tuen, M., Vasiliver-Shamis, G., Alvarez, Y., Prins, K. C., Banerjee, S., Nádas, A., et al. (2011). HIV Envelope gp120 Activates LFA-1 on CD4 T-Lymphocytes and Increases Cell Susceptibility to LFA-1-Targeting Leukotoxin (LtxA). *PloS One*, *6*(8), e23202.

Hirsch, M. S., Brun-Vézinet, F., Clotet, B., Conway, B., Kuritzkes, D. R., D'Aquila, R. T., Demeter, L. M., et al. (2003). Antiretroviral Drug Resistance Testing in Adults Infected with Human Immunodeficiency Virus Type 1: 2003 Recommendations of an International AIDS Society-USA Panel. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, *37*(1), 113-28.

Hladik, F., & McElrath, M. J. (2008). Setting the Stage: Host Invasion by HIV. *Nature Reviews Immunology*, *8*(6), 447-457.

Hornak, V., Okur, A., Rizzo, R. C., & Simmerling, C. (2006a). HIV-1 Protease Flaps Spontaneously Open and Reclose in Molecular Dynamics Simulations. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(4), 915-20.

Hornak, V., Okur, A., Rizzo, R. C., & Simmerling, C. (2006b). HIV-1 Protease Flaps Spontaneously Close to the Correct Structure in Simulations Following Manual Placement of an Inhibitor into the Open State. *Journal of the American Chemical Society*, *128*(9), 2812-2813.

Huey, R., Morris, G. M., Olson, A. J., & Goodsell, D. S. (2007). Software News and Update - A Semiempirical Free Energy Force Field with Charge-Based Desolvation. *Journal of Computational Chemistry*, *28*(6), 1145-1152.

Höglund, S., Su, J., Reneby, S. S., Végvári, Á., Hjertén, S., Sintorn, I., Foster, H., et al. (2002). Tripeptide Interference with Human Immunodeficiency Virus Type 1 Morphogenesis. *Antimicrobial agents and chemotherapy*, *46*(11), 3597-3605.

Ibe, S., Yokomaku, Y., Shiino, T., Tanaka, R., Hattori, J., Fujisaki, S., Iwatani, Y., et al. (2010). HIV-2 CRF01_AB: First Circulating Recombinant Form of HIV-2. *Journal of Acquired Immune Deficiency Syndromes, 54*(3), 241-7.

Ingr, M., Uhlíková, T. A. T. Á. N. A., & Str, K. (2003). Kinetics of the Dimerization of Retroviral Proteases: The " fireman ' s grip " and Dimerization. *Protein Science*, *12*, 2173-2182.

Irwin, J. J., & Shoichet, B. K. (2006). ZINC - A Free Database of Commercially Available Compounds for Virtual Screening. *J Chem Inf Model*, *45*(1), 177-182.

Jetzt, A. E., Yu, H., Klarmann, G. J., Ron, Y., Preston, B. D., & Dougherty, J. P. (2000). High Rate of Recombination throughout the Human Immunodeficiency Virus Type 1 Genome High Rate of Recombination throughout the Human Immunodeficiency Virus Type 1 Genome. *Journal of Virology*, *74*(3), 1234-1240.

Johnson, V. a, Calvez, V., Günthard, H. F., Paredes, R., Pillay, D., Shafer, R., Wensing, A. M., et al. (2011). 2011 Update of the Drug Resistance Mutations in HIV-1. *Topics in Antiviral Medicine*, *19*(4), 156-64.

Johnston, E., Winters, M. A., Rhee, S., Merigan, T. C., Schiffer, C. A., & Shafer, R. W. (2004). Association of a Novel Human Immunodeficiency Virus Type 1 Protease Substrate Cleft Mutation, L23I, with Protease Inhibitor Therapy and *in vitro* Drug Resistance Association of a Novel Human Immunodeficiency Virus Type 1 Protease Substrate Cleft Mutation. *Antimicrobial Agents and Chemotherapy*, *48*(12), 4864-4868.

Jones, K. L., Roche, M., Gantier, M. P., Begum, N. a, Honjo, T., Caradonna, S., Williams, B. R. G., et al. (2010). X4 and R5 HIV-1 have Distinct Post-Entry Requirements for Uracil DNA Glycosylase during Infection of Primary Cells. *The Journal of Biological Chemistry*, *285*(24), 18603-14.

Kandathil, A J, Ramalingam, S., Kannangai, R., David, S., & Sridharan, G. (2005). Molecular Epidemiology of HIV. *The Indian Journal of Medical Research*, *121*(4), 333-344.

Kandathil, Abraham Joseph, Joseph, A. P., Kannangai, R., Srinivasan, N., Abraham, O. C., Pulimood, S. A., & Sridharan, G. (2009). Structural Basis of Drug Resistance by Genetic Variants of HIV Type 1 clade C Protease from India. *AIDS Research and Human Retroviruses*, *25*(5), 511-519.

Kantor, R., & Katzenstein, D. (2004). Drug Resistance in Non-Subtype B HIV-1. *Journal of Clinical Virology the Official Publication of the Pan American Society for Clinical Virology*, *29*(3), 152-159.

Kantor, R., Zijenah, L. S., Shafer, R. W., Mutetwa, S., Johnston, E., Lloyd, R., Lieven, A. V. O. N., et al. (2002). HIV-1 Subtype C Reverse Transcriptase and Protease Genotypes in

Zimbabwean Patients Failing Antiretroviral Therapy. *AIDS Research and Human Retroviruses*, *18*(18), 1407-1413.

Kaplan, S. S., & Hicks, C. B. (2005). Safety and Antiviral Activity of Lopinavir/Ritonavir-Based Therapy in Human Immunodeficiency Virus Type 1 (HIV-1) Infection. *Journal of Antimicrobial Chemotherapy*, *56*(June), 273-276.

Kaye, P. T., Musa, M. A., Nchinda, A. T., & Nocanda, X. W. (2004). Novel Heterocyclic Analogues of the HIV-1 Protease Inhibitor, Ritonavir. *Synthetic Communications*, *34*(14), 2575-2589.

Kear, J. L., Galiano, L., Veloro, A. M., Busenlehner, L. S., & Fanucci, E. (2011). Monitoring the Autoproteolysis of HIV-1 protease by site-directed spin-labeling and Electron Paramagnetic Resonance Spectroscopy. *Journal of Biophysical Chemistry*, *2*(2), 136-145.

Korber, B., Gaschen, B., Yusim, K., Thakallapally, R., Kesmir, C., & Detours, V. (2001). Evolutionary and Immunological Implications of Contemporary HIV-1 Variation. *British Medical Bulletin*, *58*, 19-42.

Kuznetsov, Y. G., Victoria, J. G., Robinson, W. E., & McPherson, A. (2003). Atomic Force Microscopy Investigation of Human Immunodeficiency Virus (HIV) and HIV-Infected Lymphocytes. *Journal of Virology*, *77*(22), 11896-11909.

Laurie, A. T. R., & Jackson, R. M. (2006). Methods for the Prediction of Protein-Ligand Binding Sites for Structure-based Drug Design and Virtual Ligand Screening. *Current Protein & Peptide Science*, *7*(5), 395-406.

Le Rouzic, E., & Benichou, S. (2005). The Vpr Protein from HIV-1: Distinct Roles along the Viral Life Cycle. *Retrovirology*, *2*(1), 11.

Li, D., Ji, B., Hwang, K. C., & Huang, Y. (2011). Strength of Hydrogen Bond Network takes crucial roles in the dissociation process of inhibitors from the HIV-1 Protease Binding Pocket. *PloS One*, *6*(4), 1-13.

Li, H., & Pauza, C. D. (2011). HIV Envelope-Mediated, CCR5/alpha4beta7-Dependent Killing of CD4-Negative Gammadelta T Cells which are Lost during Progression to AIDS. *Journal of the Society of Hematology*, *188*(22), 5824-5831.

Li, S. C., Bu, D., Xu, J., & Li, M. (2011). Finding Nearly Optimal GDT Scores. *Journal of Computational Biology*, *18*(5), 693-704.

Lobritz, M. a, Ratcliff, A. N., & Arts, E. J. (2010). HIV-1 Entry, Inhibitors, and Resistance. *Viruses*, *2*(5), 1069-105.

Logsdon, B. C., Vickrey, J. F., Martin, P., Proteasa, G., Koepke, J. I., Terlecky, S. R., Wawrzak, Z., et al. (2004). Crystal Structures of a Multidrug-Resistant Human Immunodeficiency Virus Type 1 Protease Reveal an Expanded Active-Site Cavity. *Journal of Virology*, *78*(6), 3123-3132.

Luo, W., & Chan, K. C. C. (2012). Discovering Patterns in Drug-Protein Interactions based on their Fingerprints. *BMC bioinformatics*, *13 Suppl 9*(Suppl 9), S4.

Umamaheswari, A. M., Asokkumar, A. T. K., Sivashanmugam, V. S., & Jagannath, P. (2012). Docking Studies: Search for Possible Phytoconstituents for the Treatment of Gout. *International Journal of Biological & Pharmceutical Research*, *3*(1), 6-11.

Magalhães, C. S. De, Barbosa, H. J. C., & Dardenne, L. E. (2004). A genetic Algorithm for the Ligand-Protein Docking Problem. *Genetics and Molecular Biology*, *27*(4), 605-610.

Malim, M. H., & Emerman, M. (2001). HIV-1 Sequence Variation: Drift, Shift, and Attenuation. *Cell*, *104*(4), 469-472.

Mao, Y. (2011). Dynamical basis for Drug Resistance of HIV-1 Protease. *BMC Structural Biology*, *11*(1), 31.

Martin, P., Vickrey, J. F., Proteasa, G., Jimenez, Y. L., Wawrzak, Z., Winters, M. a, Merigan, T. C., et al. (2005). "Wide-open" 1.3 A Structure of a Multidrug-Resistant HIV-1 Protease as a Drug Target. *Structure (London, England: 1993)*, *13*(12), 1887-95.

Martinez-Cajas, J. L., Wainberg, M. a, Oliveira, M., Asahchop, E. L., Doualla-Bell, F., Lisovsky, I., Moisi, D., et al. (2012). The Role of Polymorphisms at Position 89 in the HIV-1 Protease Gene in the Development of Drug Resistance to HIV-1 Protease Inhibitors. *The Journal of Antimicrobial Chemotherapy*, *67*(4), 988-94.

Marx, P. a, Alcabes, P. G., & Drucker, E. (2001). Serial human passage of Simian Immunodeficiency Virus by Unsterile Injections and the Emergence of Epidemic Human Immunodeficiency Virus in Africa. *Philosophical Transactions of the Royal Society of London. Series B, Biological sciences*, *356*(1410), 911-20.

Maréchal, V., Arenzana-seisdedos, F., Heard, J., & Schwartz, O. (1999). Opposite Effects of SDF-1 on Human Immunodeficiency Virus Type 1 Replication Opposite Effects of SDF-1 on Human Immunodeficiency Virus Type 1 Replication. *Journal of Virology*, *73*(5), 3608-3615.

Mathu, A. M. N. (2012). Structural Analysis of Effects of Mutations on HIV-1 Subtype C Protease Active Site, (March), 1-91.

McCutchan, F. E. (2000). Understanding the Genetic Diversity of HIV-1. *AIDS*, *14 Suppl 3*(SUPPL. 3), S31-S44.

Miller, M., Schneider, J., Sathyanarayana, B. K., Toth, M. V, Marshall, G. R., Clawson, L., Selk, L., et al. (1989). Structure of Complex of Synthetic HIV-1 Protease with a Substrate-based Inhibitor at 2.3 A Resolution. *Science*, *246*(4934), 1149-1152.

Mishra, G. R., Suresh, M., Kumaran, K., Kannabiran, N., Suresh, S., Bala, P., Shivakumar, K., et al. (2006). Human Protein Reference Database - 2006 Update. *Nucleic Acids Research*, *34*(Database issue), D411-D414.

Morand-Joubert, L., Charpentier, C., Poizat, G., Chêne, G., Dam, E., Raguin, G., Taburet, A. M., et al. (2006). Low Genetic Barrier to Large Increases in HIV-1 Cross-Resistance to Protease Inhibitors during Salvage Therapy. *Antiviral Therapy*, *11*(2), 143-54.

Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K., Olson, A. J., et al. (1998). Automated Docking using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function. *Journal of Computational Chemistry*, *19*(14), 1639-1662.

Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., & Olson, A. J. (2009). AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *Journal of Computational Chemistry*, *30*(16), 2785-2791.

Mthunzi, P., & Meyer, D. (2004). Limited Cross-Reactivity between Different HIV-1 Clades. *Journal of Clinical Virology: The Official Publication of the Pan American Society for Clinical Virology*, *31 Suppl 1*(2004), S88-91.

Nicastri, E., Sarmati, L., d'Ettorre, G., Palmisano, L., Parisi, S. G., Uccella, I., Rianda, A., et al. (2003). Replication Capacity, Biological Phenotype, and Drug Resistance of HIV strains Isolated from Patients Failing Antiretroviral Therapy. *Journal of Medical Virology*, *69*(1), 1-6.

Nielsen, M. H., Pedersen, F. S., & Kjems, J. (2005). Molecular Strategies to Inhibit HIV-1 Replication. *Retrovirology*, *2*(10), 1-20.

Nukoolkarn, S., Pongthapisith, V., Panyim, S., & Leelamanit, W. (2004). Sequence Variability of the HIV Type 1 Protease Gene in Thai Patients Experienced with Antiretroviral Therapy. *AIDS Research and Human Retroviruses*, *20*(12), 1368-72.

Ode, H., Matsuyama, S., Hata, M., Neya, S., Kakizawa, J., Sugiura, W., & Hoshino, T. (2007). Computational Characterization of Structural Role of the Non-Active Site Mutation M36I of Human Immunodeficiency Virus Type 1 Protease. *Journal of Molecular Biology*, *370*(3), 598-607.

Ohtaka, H., & Freire, E. (2005). Adaptive Inhibitors of the HIV-1 Protease. *Progress in Biophysics and Molecular Biology*, *88*(2), 193-208.

Olsen, D. B., Stahlhut, M. W., Rutkowski, C. A., Schock, H. B., Aimee, L., & Kuo, L. C. (1999). Non-Active Site Changes Elicit the HIV-1 Protease to Inhibitors. *Journal of Biological Chemistry*, *274*(6), 23699-23701.

Ortega, K. L., Vale, D. a, & Magalhães, M. H. C. G. (2009). Impact of PI and NNRTI HAART-based Therapy on Oral Lesions of Brazilian HIV-infected patients. *Journal of Oral Pathology & Medicine: Official Publication of the International Association of Oral Pathologists and the American Academy of Oral Pathology*, *38*(6), 489-94.

Panther, L. A., & Libman, H. (2005). Medical overview. *Cambridge University Press*.

Papathanasopoulos, M. a, Cilliers, T., Morris, L., Mokili, J. L., Dowling, W., Birx, D. L., & McCutchan, F. E. (2002). Full-Length Genome Analysis of HIV-1 Subtype C Utilizing CXCR4 and Intersubtype Recombinants Isolated in South Africa. *AIDS Research and Human Retroviruses*, *18*(12), 879-86.

Pasquato, A., Dettin, M., Basak, A., Gambaretto, R., Tonin, L., Seidah, N. G., & Di Bello, C. (2007). Heparin Enhances the Furin Cleavage of HIV-1 gp160 Peptides. *FEBS Letters*, *581*(30), 5807-5813.

Pazhanisamy, S., Stuver, C. M., Cullinan, A. B., Margolin, N., Rao, B. G., & Livingston, D. J. (1996). Kinetic Characterization of Human Immunodeficiency Virus Type-1 Protease-Resistant Variants. *The Journal of Biological Chemistry*, *271*(30), 17979-85.

Peeters, M., Gueye, a, Mboup, S., Bibollet-Ruche, F., Ekaza, E., Mulanga, C., Ouedrago, R., et al. (1997). Geographical Distribution of HIV-1 Group O Viruses in Africa. *AIDS (London, England)*, *11*(4), 493-8.

Perryman, A. L., & Lin, J. (2004). HIV-1 Protease Molecular Dynamics of a Wild-Type and of the V82F / I84V mutant: Possible Contributions to Drug Resistance and a Potential New Target Site for Drugs. *Protein Science*, *13*, 1108-1123.

Petrokova, H. (2006). On the Role of the R Configuration of the Reaction-Intermediate Isostere in HIV-1 Protease-Inhibitor Resolution Binding: X-ray Structure at 2.0 A. *Acta Crys.*, *62*, 489-497.

Piana, S., Carloni, P., & Rothlisberger, U. (2002). Drug resistance in HIV-1 Protease: Flexibility-Assisted Mechanism of Compensatory Mutations. *Protein Science*, *11*, 2393-2402.

Plantier, J. C., Leoz, M., Dickerson, J. E., De Oliveira, F., Cordonnier, F., Lemée, V., Damond, F., et al. (2009). A New Human Immunodeficiency Virus Derived from Gorillas. *Nature Medicine*, *15*(8), 871-2.

Pokorná, J., Machala, L., Řezáčová, P., & Konvalinka, J. (2009). Current and Novel Inhibitors of HIV Protease. *Journal of Medicinal Chemistry*, *1*(3), 1209–1239.

Pomerantz, R. J., & Horn, D. L. (2003). Twenty Years of Therapy for HIV-1 Infection. *Nature Medicine*, *9*(7), 867-73.

Prabu-Jeyabalan, M., Nalivaika, E., & Schiffer, C. a. (2002). Substrate Shape Determines Specificity of Recognition for HIV-1 Protease: Analysis of Crystal Structures of Six Substrate Complexes. *Structure (London, England: 1993)*, *10*(3), 369-81.

Purohit, R., Rajasekaran, R., Sudandiradoss, C., George Priya Doss, C., Ramanathan, K., & Rao, S. (2008). Studies on Flexibility and Binding Affinity of Asp25 of HIV-1 Protease Mutants. *International Journal of Biological Macromolecules*, *42*(4), 386-91.

Quesnel-Vallières, M., Kouzayha, I., Tran, E., Barry, I., Lasgi, C., Merindol, N., Monteil, V., et al. (2011). Novel HIV-1 Recombinant Forms in Antenatal Cohort, Montreal, Quebec, Canada. *Emerging Infectious Diseases*, *17*(2), 271-4.

Reeves, J. D., & Doms, R. W. (2002). Human Immunodeficiency Virus Type 2. *The Journal of General Virology*, *83*(Pt 6), 1253-65.

Regoes, R. R., & Bonhoeffer, S. (2005). The HIV Coreceptor Switch: A Population Dynamical Perspective. *Trends in Microbiology*, *13*(6), 269-77.

Rhodes, T., Wargo, H., Hu, W., & Virginia, W. (2003). High Rates of Human Immunodeficiency Virus Type 1 Recombination: Near-Random Segregation of Markers One Kilobase Apart in One Round of Viral Replication. *Journal of Virology*, *77*(20), 11193-11200.

Rodriguez, R., Chinea, G., Lopez, N., Pons, T., & Vriend, G. (1998). Homology Modeling, Model and Software Evaluation: Three Related Resources. *Bioinformatics*, *14*(6), 523-528.

Sali, A., & Blundell, T. L. (1993). Comparative Protein Modelling by Satisfaction of Spatial Restraints. *Journal of Molecular Biology*, *234*(3), 779-815.

Santiago, M. L., Range, F., Keele, B. F., Li, Y., Bailes, E., Bibollet-Ruche, F., Fruteau, C., et al. (2005). Simian Immunodeficiency Virus Infection in Free-Ranging Sooty Mangabeys (Cercocebus atys atys) from the Taï Forest, Côte d'Ivoire: Implications for the Origin of Epidemic Human Immunodeficiency Virus Type 2. *Journal of Virology*, *79*(19), 12515-12527.

Shafer, R. W., Dupnik, K., Winters, M. A., & Eshleman, S. H. (2001). A Guide to HIV-1 Reverse Transcriptase and Protease Sequencing for Drug Resistance Studies. *Reviews*, 1-51.

Shafer, R. W., Jung, D. R., Betts, B. J., Xi, Y., & Gonzales, M. J. (2000). Human Immunodeficiency Virus Reverse Transcriptase and Protease Sequence Database. *Nucleic Acids Research*, *28*(1), 346-348.

Sham, H. L., Kempf, D. J., Molla, A., Marsh, K. C., Kumar, G. N., Chen, C., Kati, W., et al. (1998). ABT-378 , a Highly Potent Inhibitor of the Human Immunodeficiency Virus Protease. *Antimicrobial Agents and Chemotherapy*, *42*(12), 3218-3224.

Shen, R., Smythies, L. E., Clements, R. H., Novak, L., & Smith, P. D. (2010). Dendritic cells Transmit HIV-1 through Human Small Intestinal Mucosa. *Journal of Leukocyte Biology*, *87*(4), 663-670.

Smith, R. A, Loeb, L. A, & Preston, B. D. (2005). Lethal Mutagenesis of HIV. *Virus Research*, *107*(2), 215-28.

Song, L., Sun, Z. Y. J., Coleman, K. E., Zwick, M. B., Gach, J. S., Wang, J., Reinherz, E. L., et al. (2009). Broadly Neutralizing Anti-HIV-1 Antibodies Disrupt a Hinge-Related Function of

gp41 at the Membrane Interface. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(22), 9057-62.

Sánchez, R., & Sali, A. (1997). Evaluation of Comparative Protein Structure Modeling by MODELLER-3. *Proteins*, *Suppl 1*(s 1), 50-58.

Takehisa, J., Kraus, M. H., Ayouba, A., Bailes, E., Van Heuverswyn, F., Decker, J. M., Li, Y., et al. (2009). Origin and Biology of Simian Immunodeficiency Virus in Wild-Living Western Gorillas. *Journal of Virology*, *83*(4), 1635-1648.

Tang, M. W., Liu, T. F., & Shafer, R. W. (2012). The HIVdb System for HIV-1 Genotypic Resistance Interpretation. *Intervirology*, *5107*, 98-101.

Taştan Bishop, A. Ö., Beer, T. A. P. de, & Joubert, F. (2008). Protein Homology Modelling and its use in South Africa. *South African Journal of Science*, *104*(1-2), 2-6.

Taylor, B. S., Sobieszczyk, M. E., McCutchan, F. E., & Hammer, S. M. (2008). The Challenge of HIV-1 Subtype Diversity. *The New England Journal of Medicine, 358*(15), 1590-602.

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), 213-8.

Triques, K., Bourgeois, A., Saragosti, S., Vidal, N., Mpoudi-Ngole, E., Nzilambi, N., Apetrei, C., et al. (1999). High diversity of HIV-1 subtype F strains in Central Africa. *Virology*, *259*(1), 99-109.

UNAIDS. (2011). *UNAIDS World AIDS Day Report 2011* (pp. 1–52).

Vallari, A., Holzmayer, V., Harris, B., Yamaguchi, J., Ngansop, C., Makamche, F., Mbanya, D., et al. (2011). Confirmation of Putative HIV-1 Group P in Cameroon. *Journal of virology*, *85*(3), 1403-7.

Van der Kuyl, A. C. (2012). HIV Infection and HERV Expression: A Review. *Retrovirology*, *9*(1), 6.

Van Heuverswyn, F., Li, Y., Neel, C., Bailes, E., Keele, B. F., Liu, W., Loul, S., et al. (2006). Human Immunodeficiency Viruses: SIV Infection in Wild Gorillas. *Nature*, *444*(7116), 164.

Van Maarseveen, N. M., Andersson, D., Lepšík, M., Fun, A., Schipper, P. J., De Jong, D., Boucher, C. A. B., et al. (2012). Modulation of HIV-1 Gag NC/p1 Cleavage Efficiency affects Protease Inhibitor Resistance and Viral Replicative Capacity. *Retrovirology*, *9*(1), 29.

Vivès, R. R., Imberty, A., Sattentau, Q. J., & Lortat-Jacob, H. (2005). Heparan Sulfate Targets the HIV-1 Envelope Glycoprotein gp120 Coreceptor Binding Site. *The Journal of Biological Chemistry*, *280*(22), 21353-21357.

Wang, Y., Klock, H., Yin, H., Wolff, K., Bieza, K., Niswonger, K., Matzen, J., et al. (2005). Homogeneous High-Throughput Screening Assays for HIV-1. *Journal of Biomolecular Screening*, *10*(5), 456-462.

Watson Buckheit, K., Yang, L., & Buckheit, R. W. (2011). Development of Dual-Acting Pyrimidinediones as Novel and Highly Potent Topical Anti-HIV Microbicides. *Antimicrobial Agents and Chemotherapy*, *55*(11), 5243-54.

Weiss, R. A. (1993). How does HIV cause AIDS? *Science*, *260*(5112), 1273-1279.

WHO, UNAIDS, U. (2011). *Global HIV/AIDS Response: Epidemic Update and Health Sector Progress towards Universal Access* (pp. 1-229).

Wideburg, E., Burt, S. K., Erickson, J. W., Singh, M. K., Chen, C., Wideburg, N. E., Burt, S. K., et al. (1994). Characterization of Human Immunodeficiency Virus Type 1 Variants with Increased Resistance to a C2-Symmetric Protease Inhibitor. *Journal of Virology*, *68*(3), 2016-2020.

Wlodawer, a, & Vondrasek, J. (1998). Inhibitors of HIV-1 Protease: A Major Success of Structure-Assisted Drug Design. *Annual Review of Biophysics and Biomolecular Structure*, *27*, 249-84.

Wlodawer, A., Miller, M., Jaskólski, M., Sathyanarayana, B. K., Baldwin, E., Weber, I. T., Selk, L. M., et al. (1989). Conserved Folding in Retroviral Proteases: Crystal Structure of a Synthetic HIV-1 Protease. *Science*, *245*(4918), 616-621.

Wu, L., LaRosa, G., Kassam, N., Gordon, C. J., Heath, H., Ruffing, N., Chen, H., et al. (1997). Interaction of Chemokine Receptor CCR5 with its Ligands: Multiple Domains for HIV-1 gp120 Binding and a Single Domain for Chemokine Binding. *The Journal of Experimental Medicine*, *186*(8), 1373-1381.

Xiongyu Wu, Per Öhrngren, Jenny K. Ekegren, Johan Unge, Torsten Unge, Hans Wallberg, B. S., & Anders Hallberg, and M. L. (2008). Two-Carbon-Elongated HIV-1 Protease Inhibitors with a Tertiary-Alcohol-Containing Transition-State Mimic. *Journal of Medicinal Chemistry*, *51*(4), 1053-1057.

Yamaguchi, J., Coffey, R., Vallari, A., Ngansop, C., Mbanya, D., Ndembi, N., Kaptué, L., et al. (2006). Identification of HIV type 1 Group N Infections in a Husband and Wife in Cameroon: Viral Genome Sequences provide Evidence for Horizontal Transmission. *AIDS Research and Human Retroviruses*, *22*(1), 83-92.

Yu, V., Raghuraman, H., Hamelberg, D., Tonelli, M., & Westler, W. M. (2011). Protein Conformational Dynamics in the Mechanism of HIV-1 Protease Catalysis. *PNAS*, *108*(52), 20982-20987.

Zazzi, M., Romano, L., Venturi, G., Shafer, R. W., Reid, C., Bello, F. D., Parolin, C., et al. (2004). Comparative Evaluation of Three Computerized Algorithms for Prediction of Antiretroviral Susceptibility from HIV Type 1 Genotype. *Journal of Antimicrobial Chemotherapy*, *53*(2), 356-360.

Zeldin, R. K., & Petruschke, R. A. (2004). Pharmacological and Therapeutic Properties of Ritonavir-Boosted Protease Inhibitor Therapy in HIV-Infected Patients. *Journal of Antimicrobial Chemotherapy*, *53*(1), 4-9.

Zhu, T., Korber, B. T., Nahmias, A. J., Hooper, E., Sharp, P. M., & Ho, D. D. (1998). An African HIV-1 Sequence from 1959 and Implications for the Origin of the Epidemic. *Nature*, *391*(6667), 594-597.

Zhu, Z., Schuster, D. I., & Tuckerman, M. E. (2003). Molecular Dynamics Study of the Connection between Flap Closing and Binding of Fullerene-based Inhibitors of the HIV-1 Protease. *Biochemistry*, *42*(5), 1326-33.

Zyl, G. U. Van, Rabie, H., Nuttall, J. J., & Cotton, M. F. (2011). It is Time to Consider Third-Line options in Antiretroviral-Experienced Paediatric Patients? *Journal of the International AIDS Society*, *14*(1), 55.