STRUCTURAL ANALYSIS OF EFFECTS OF MUTATIONS ON HIV-1 SUBTYPE C PROTEASE ACTIVE SITE

A thesis submitted in fulfillment of the requirement for the Degree

of

MASTER OF SCIENCE

IN

BIOINFORMATICS

AND

COMPUTATIONAL MOLECULAR BIOLOGY

AT

RHODES UNIVERSITY

By

Alexander Muchugia Nganga Mathu

March 2012

ABSTRACT

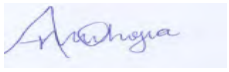
HIV/AIDS is a global pandemic that poses a great threat especially in Sub-Saharan Africa where the highest population of those infected with the virus is found. It has far reaching medical, socio-economic and scientific implications. The HIV-1 protease enzyme is a prime therapeutic target that has been exploited in an effort to reduce morbidity and mortality. However problems arise from drug toxicity and drug-resistant mutations of the protease which is a motivation for research for new, safer and effective therapies. Evidence exists to show that there are significant genomic differences in Subtype B and C that have a negative effect on the intrinsic binding of inhibitors. It is imperative to look at all perspectives from epidemiological, molecular to the pharmacological ones so as to achieve rational design of therapeutic agents.

This study involved the use of *in silico* structural analysis of the effects of mutations in the active site. The data was provided by the National Institute of Communicable Diseases consisting of HIV-1 Subtype C protease sequences of 29 infants exhibiting drug-resistance to ritonavir and lopinavir. The major active site mutations causing drug resistance identified in this study were M46I, I54V and V82A using the Stanford HIV database tool. Homology modeling without extra restraints produced models with improved quality in comparison to those with restraints. MetaMQAPII results differed when models were visualized as dimers giving erroneous modeled regions in comparison to monomers.

A broader study with a larger dataset of HIV-1 subtype C protease sequences is required to increase statistical confidence and in order to identify the pattern of drug resistant mutations. Homology modeling without extra restraints is preferred for calculating homology models for the HIV-1 subtype C. Further investigations needs to be done to ascertain the accuracy of validation results for dimers from MetaMQAPII as it is designed for evaluation of monomers.

DECLARATION

I, Alexander Muchugia Nganga Mathu, declare that the work presented here is my own and that it has not been submitted for examination in any other University.

Signature……  ……

Dated……23rd March 2012……

TABLE OF CONTENTS

ACKNOWLEGEMENTS

I would like to thank the Almighty God for this opportunity to pursue a MSc. degree and giving me the courage, energy and drive to see me through. Secondly I thank my parents Ruth and the Late Q.N. Mathu and for helping me get to where I am for the encouragement, prayers and support. I would also like to thank Dr. Nelly Mungai for directing me to this opportunity. Thirdly, My Supervisors Dr. Kevin Lobb and Özlem Tastan Bishop for accepting to supervise me and for the constant guidance and positive criticism which has enabled me to explore the world of science. My gratitude goes out to Professor Perry Kaye and collaborators from the National Institute of Communicable Diseases who were the initiators of the project. I would like to appreciate my colleagues Joyce, Dustin, James and Sarah who have immensely contributed towards this work. I would like to especially thank Rowan Hatherley for his assistance in python script writing and Matthys Kroon for use of python scripts. Last but not least I am indebted to Rhodes University, the Biochemistry, Microbiology and Biotechnology and Chemistry Departments and the Medical Research Council (MRC) for creating an enabling environment and financial support so as to enable me to complete this project.

LIST OF FIGURES AND ILLUSTRATIONS

LIST OF TABLES

1 CHAPTER 1

1 INTRODUCTION

An estimated 33.3 million people were reportedly infected with the Human Immunodeficiency Virus (HIV) by 2009. Thus, HIV infection is considered a pandemic according to the United Nations AIDS (UNAIDS) global report of 2010 (UNAIDS, 2010). The HIV virus is a retrovirus (RNA virus) belonging to the *retroviridae* and *orthoretrovirinae* family and sub-family respectively. It is further classified into the genus of Lentiviruses (a genus of several viruses of the *retroviridae* family characterized by a long incubation period and a unique ability among retroviruses of being able to replicate in non-dividing cells). There are two sub-types of HIV, HIV-1 and HIV-2. While HIV-1 was identified from a sample originating from the Democratic Republic of Congo in 1959 (Zhu *et al.*, 1998), HIV-2 was isolated in 1986 from patient samples from West Africa. HIV-2 exhibits 40-60% identity with HIV-1. The HIV-1 subtype is further divided into three groups M (Main group), O (Outlier group), and N (Non M, Non O). Group M constitutes most of the worlds infected population. Group M is further subdivided into subtypes A-D, F-H, J and K and circulating recombinant forms which consist of a mix of two subtypes for example, CRF02_AG (Stebbing and Moyle, 2003).

When HIV infection goes untreated it leads to decrease in the human immune cells and this leaves the body defenseless against easily treatable pathogens. These are referred to as opportunistic infections e.g. Tuberculosis, fungal infections (Curry *et al.*, 1991). The condition where these infections manifest is referred to Acquired Immuno-Deficiency Syndrome (AIDS) whose cause is linked to HIV (Araya *et al.*, 2011). The opportunistic infections are responsible for mortality and not the HIV infection itself.

1.1 THE HIV STRUCTURE, LIFE CYCLE & TREATMENT

While outside the human cell, HIV is known as a virion and is characterized by spherical, studded morphology with spicules. The virion is 0.1 microns in diameter, and is covered by an envelope. The spicules consist of a trimer of the surface glycoprotein (gp120) and trans-membrane glycoprotein (gp41). Just below the viral envelope is a layer called the matrix,

which is synthesized from the protein p17. The viral core, also known as the capsid, is synthesized from the protein p24. The viral core consists of two RNA strands together with three enzymes namely; reverse transcriptase, integrase and protease (see Figure 1). The HIV genome encodes nine genes, three of which are involved in coding for structural proteins for new virions namely; *gag*, *pol* and *env*. The other six genes namely *tat*, *rev*, *nef*, *vif*, *vpr* and *vpu* code for proteins important in pathogenesis (Nielsen *et al.*, 2005)

.

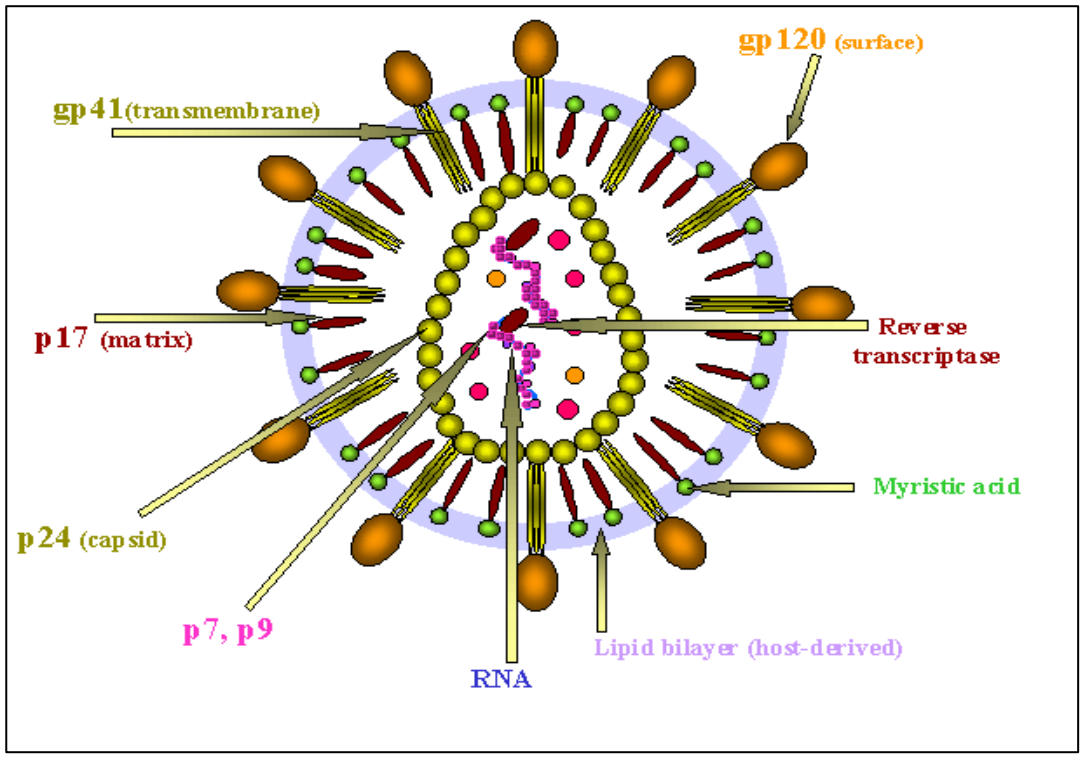


Figure 1.1: Structure of HIV virion showing the surface glycoprotein, trans-membrane proteins and enzymes (Obtained from the website http://pathmicro.med.sc.edu/lecture/hivstruct.gif and is reproduced here with permission from the author.)

The life cycle of the virus can be summarized by the following basic steps; transmission, infection, entry, reverse transcription, integration, replication, assembly budding and maturation (Nielsen *et al.*, 2005, Buckheit *et al.*, 2011). Infection with HIV begins when the virus transmitted by an infected person through contact with infected blood and/or body fluids. The virus attaches itself to the CD4 cells (CD4 cells are T-helper cells that are involved in cellular immunity and express the CD4 surface glycoprotein) using the surface glycoprotein

gp120. The chemokine co-receptors CCR5 and CXCR4 found on the CD4 cells consequently attach to the gp120 (Dragic *et al.*, 1996). A conformational change occurs on the gp120 exposing the trans-membrane gp41 allowing the virus to insert into CD4 cell membrane using its heptad repeat (HR)-1 domains' hydrophobic terminus. A zipping process involving heptad repeat (HR)-2 domain brings the HIV virus towards CD4 cell membrane and fusion occurs. Once inside the CD4 cell, the RNA molecules are reverse transcribed by reverse transcriptase into DNA which is incorporated into the host cell genome by the integrase enzyme. During the host cells' replication process, viral DNA is replicated and its polyproteins are synthesized. The polyproteins are then cleaved by the HIV protease into mature proteins. Assembly takes place and the new virus can infect other cells (see Figure 1.1).

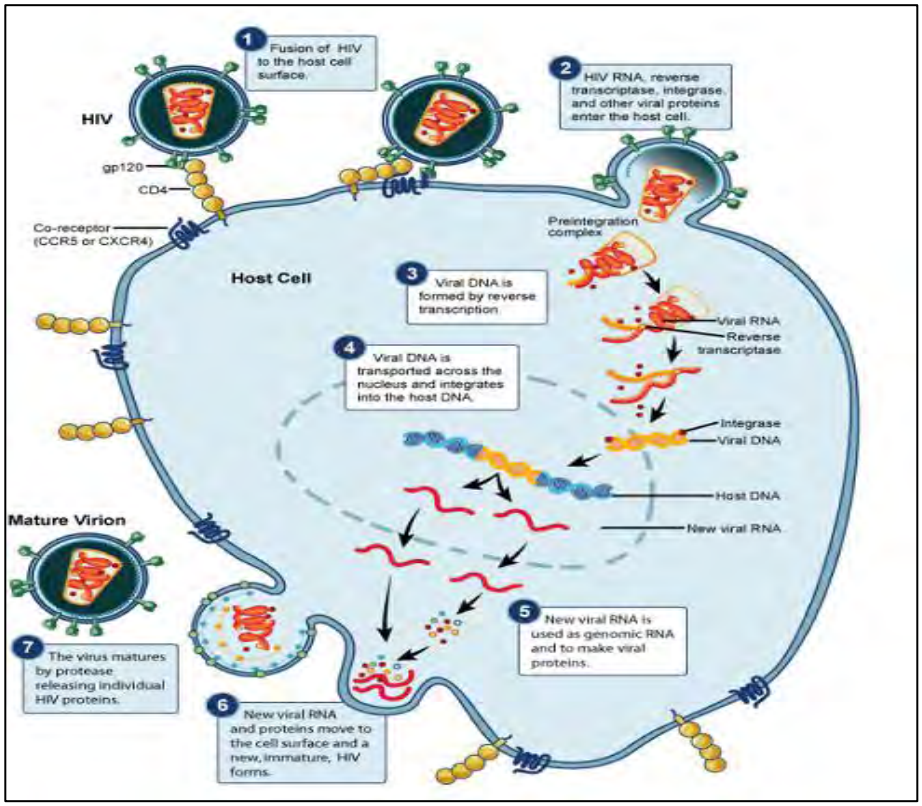


Figure 1.2: The Life cycle of HIV with the different stages; entry, reverse transcription, integration, replication, assembly budding and maturation. (Image courtesy of NIAID http://www.niaid.nih.gov)

The life cycle stages of the HIV infection have been used as drug targets, for example, fusion (fusion inhibitors), transcription (reverse transcriptase inhibitors), integration (integrase inhibitors) and maturation (protease inhibitors). HIV protease is very important for the final process of maturation as the virus can only function if the precursors gag and gag-pol polyproteins are cleaved. When cleaved they render a mature virus which can then continue to infect other cells. The gag polyprotein is the precursor of the matrix, capsid, nucleo-caspid and p6 proteins, whereas the gag-pol polyprotein gives rise to protease, integrase, and reverse transcriptase enzymes. The protease inhibitors therefore interrupt the process of cleavage and the polyproteins are rendered nonfunctional interrupting maturation into an infectious virus(Kandathil *et al.*, 2009).

Treatment of HIV positive patients with antiretroviral drug therapy begins once their CD4 cell count drops to levels below 350 cells/mm^3 from the normal which is usually 500-1200 cells/mm^3 (Bofill *et al.*, 1992). In addition the viral load (RNA copies per milliliter of blood plasma) is considered during initiation of treatment. In this regard, virological failure is defined as persistent plasma HIV RNA levels in the 200 to 1,000 copies/mL range and is more sensitive than the CD4 count as it is a direct measure of virus in plasma (Aleman *et al.*, 2002, Karlsson *et al.*, 2004)

Currently available therapeutic agents are classified into four major groups according to the particular stage they target in the lifecycle. As aforementioned, the enzyme inhibitors act on the reverse transcriptase, integrase and protease, while the fusion inhibitors that target the receptors for cell entry. The regimens consist of Highly Active Antiretroviral Therapy (HAART) which is made up of three classes of drugs; (i) Nucleoside Reverse Transcriptase Inhibitors (NRTI) for example, zidovudine and stavudine, (ii) Non-Nucleoside Reverse Transcriptase Inhibitors (NNRTI) for example, nevirapine and efavirenz and (iii) Protease Inhibitors (PI) for example lopinavir and ritonavir. Newer classes of drugs include fusion (cell entry) inhibitors such as maraviroc and integrase inhibitors such as raltegravir. They are all currently Food and Drug Administration (FDA) approved (Saste *et al.*, 2011).

The mode of action of current Food and Drug Administration (FDA) HIV protease inhibitors is by competitive inhibition of the binding site for natural substrates. The drugs are referred to as peptidomimetics as they share similar structure to the peptides hence can mimic their action but are designed not to be cleaved. Two important inhibitors that are currently in use in Southern Africa are ritonavir (Zeldin and Petruschke, 2004) and lopinavir (Sham et al., 1998). Ritonavir is a first generation protease inhibitor while lopinavir is a second generation inhibitor designed from ritonavir (see figure 1.3). Lopinavir is known to have low oral bioavailability and poor pharmacokinetic profile. The activity of Lopinavir is enhanced by giving a low dose of ritonavir which inhibits cytochrome p450 3A4 enzyme preventing metabolism of lopinavir. The drug is available as a fixed dose combination of lopinavir and ritonavir branded as Kaletra (Ali *et al.*, 2010, Venter *et al.*, 2006).

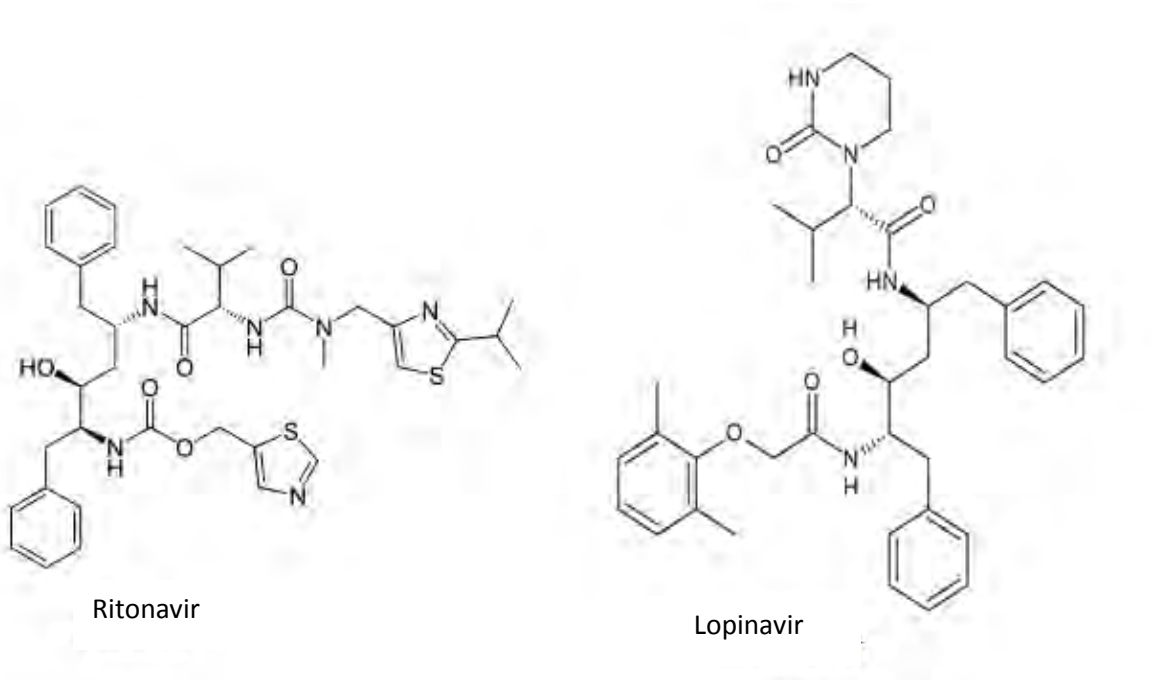


Figure 1.3: Chemical structures of protease inhibitors, ritonavir and lopinavir. Structures obtained from drug bank.

1.2 HIV -1 PROTEASE

The HIV-1 protease is classified as an aspartyl protease as its active site contains a conserved sequence motif of Asp-25, Thr-26, and Gly-27 (Castro *et al.*, 2011). It is a homodimer that is composed of two monomers each consisting of 99 amino acids. It has a binding pocket which comprises of two flexible glycine rich β-hairpins known as flaps. The protease is known to exist in three conformations open, semi-open and closed. The conformations are important for substrate and consequently, inhibitor access. The closed and semi-open forms have been characterized experimentally but no crystal structure exists for open conformation. These flaps are also important in binding of protease inhibitors and are part of the active site. The amino acids in this region interact with the protease inhibitors (Hornak, 2006). Substrates fill a common space in the protease enzyme referred to as the "substrate envelope". It is however observed that the inhibitors have protrusions outside the "substrate envelop". These protrusions are affected when mutations occur but do not affect the substrate as it is within the envelope. Protease mutations that are observed to affect the contact with these protrusions are I50V, V82I and V84A. Therefore inhibitors should be designed in a way not to protrude outside the "substrate envelop" so as to decrease susceptibility to resistance (Badley *et al.*, 2000, Doyon *et al.*, 2009).

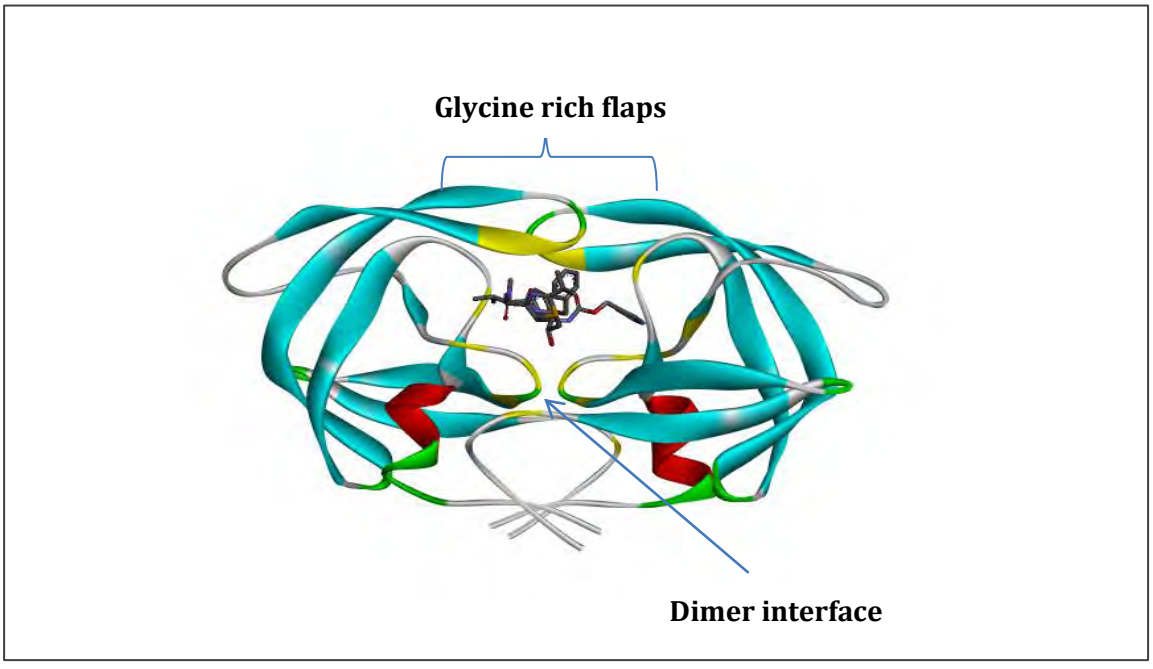


Figure 1.4: HIV protease with inhibitor ritonavir in its active site. The active site is highlighted in yellow consisting of residues 25-32, 47-53 and 80-84.

1.2.1 Mechanism of Action

The aspartate 25 residues of each HIV protease monomer and a molecule of water are proposed to take part in cleavage of amide bonds of the substrates. The cleavage takes place at the scissile bond and is positioned on two most hydrophobic residues (see Figure 1.5) (Tomasselli and Heinrikson, 2000). The monoprotonated form of the protease is responsible for the peptide bond hydrolysis of the substrates. The carboxyl group of one aspartate residue acts as an acid while the other acts as a nucleophile (base) in an acid base catalysis reaction (Palmeira *et al.*, 2006). The unprotonated carboxyl group deprotonates the water involved in the lytic action whereas the protonated carboxyl group of aspartic residue donates its proton to the scissile carboxyl oxygen and in this way the substrates are efficiently cleaved (Castro *et al.*, 2011). The HIV protease has broad specificity and cleaves a number of peptide bonds with the exception of lysine at position P2-P2' and β branched residue at P1 (see Figure 1.5). The minimum requirement for the protease to cleave substrates is 7 amino acid residues but commonly substrates have 8 residues. The enzyme consists of 8 subsites and has interactions with ideal substrates that are extended to subsites on the surface of the enzyme (Tomasselli and Heinrikson, 2000). The protease inhibitors mechanism of action is based on mimicking the transition states intermediates of the polypeptide substrates (see Figure 1.5). Protease inhibitors are designed using structure based rationale by replacing the scissile bond with either of the five (I-V) dipeptide isosteres (see Figure 1.5) (Wlodawer and Vondrasek, 1998). Ritonavir is based on the hydroxyethelene moiety (see Figure 1.5 (III)). Protease inhibitors have common features with the polypeptides and this include; a secondary hydroxyl group, a substitute for the P1 carbonyl group of the substrate that interacts with active site ASP25/25' for tight binding with protease . A water molecule is conserved and facilitates contact between P2/P1' carbonyl oxygen atoms and amide groups of Ile 50/50' of enzyme (Ali et al., 2010).

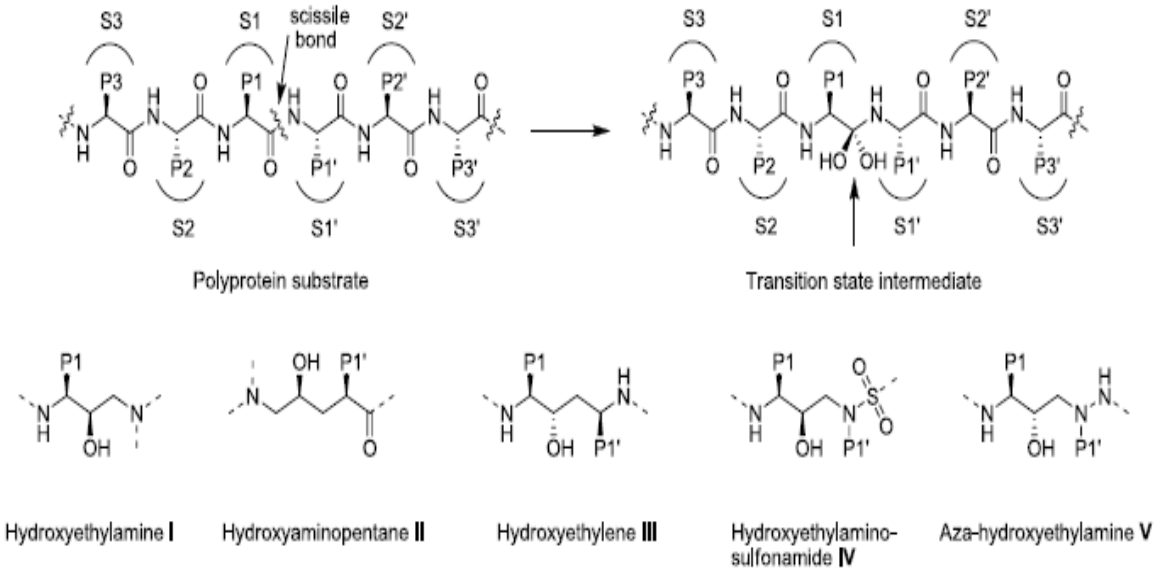


Figure 1.5: Illustration of polypeptide cleavage and transition state mimetics. S1…Sn, S1'...Sn' represent enzyme subsites while P1…Pn, P1'…Pn' represent the polypeptide residues closest to the scissile bond.

1.2.2 Mutations

There are various mutations that have been identified in the protease enzyme in patients observed to be failing treatment as defined by their increased viral load and decreased CD4 counts. The mutations observed include L10I, L24I, L33F, M46L, I54V, L63P, A71V, V82A, and I84V. These have been associated with resistance in patients on lopinavir. There is a lot of interest with mutations especially around the flap region such as I47A and I54V. Isoleucine has a longer side chain than both alanine and valine and it points towards the substrate binding site (Doyon *et al.*, 2009). The two mutations cause an increase in volume of the active site and also structural alteration of the flap. This affects the fitting of the inhibitor(s) to the active site. Some of these mutations affect catalytic activity. Nonetheless, the protease enzyme is still able to cleave the substrates with activity as low as 23% in vivo (Šašková *et al.*, 2008).

1.2.3 Variation in protease sequences

In previous studies there has been evidence suggesting that there exist some differences between the subtype B and C protease enzyme (Sanches *et al.*, 2007, Wainberg and Brenner, 2010). Protease inhibitors currently in use have been designed based on subtype B. There is an increase in the prevalence of the subtype C in various countries and this has been associated with either ease of spread in population or the "fitness" the virus (Gordon et al., 2003, Bessong, 2008b). HIV subtype C protease genes are more catalytically active than those of other subtypes. There are certain drug resistant mutations that are known to occur more frequently in subtype C protease than in subtype B (Arora *et al.*, 2008). Drug resistant mutations common for all protease inhibitors are M36I, I93L while K20R is mainly in patients on ritonavir (Shafer, 2006). The global subtype C consensus varies from the global subtype B consensus at the following eight positions T12S, I15V, L19I, M36I, R41K, H69K, L89M and I93L (Bessong, 2008b).

1.3 COMPUTATIONAL APPROACHES

There are two computational approaches that can be employed for drug development; (i) ligand based and (ii) structural based methods. The ligand based methods concentrate on physicochemical properties of active ligands necessary for biological affinity. They utilize ligand (drug in our case) modeling to maintain properties of either interacting with the target and causing pharmacological response or blocking it. Structural based methods on the other hand are dependent on three dimensional structures of the biomolecules, that is, the drug target(s). These are known to have an advantage of high predictive accuracy compared to their ligand based counterparts (Kirchmair *et al.*, 2011). Homology modeling approach has been employed where no existing structures have been characterized experimentally. Reliable structures are those obtained with a sequence identity of about 40% (Di Luccio and Koehl, 2011). It is now possible to produce high quality models as the gap between automation and expert produced models is closing.

1.3.1 Data analysis

The Stanford Mutation Database contains information about the prevalence of drug resistance mutations (Rhee *et al.*, 2003). This is useful for drug resistance surveillance, development of antiretrovirals and management of HIV infections. In the case of drug development it contains the frequent mutations occurring in the different HIV subtypes for patients before and after treatment. This enables researchers deduce which mutations are caused by selective drug pressure from the polymorphisms. The Stanford mutation database is also useful in identifying mutations such as those which differ from the subtype B consensus for the other subtypes (Shafer, 2006).

1.3.2 Homology modeling

Over the years, many structures have been characterized experimentally. The number of protein folds being discovered is decreasing which can be rationalized as there are only a given number of folds for the number of existing sequences. The premise that similar sequences give rise to similar structures and that sequence is less conserved than structure has enabled us to use computational techniques to model structures (Petsko, 2002). In essence this has reduced the time that one would require to experimentally characterize the structures and has even led to development of potentially lifesaving drugs as in the case of Severe Acute Respiratory Syndrome (SARS). Comparative modeling has enabled us to predict three dimension structures by relating similar sequences to those with elucidated structures (Eswar *et al.*, 2002).

1.3.3 De novo ligand design

Analysis of the binding site of the target protein is generally done with a ligand bound in the site. This facilitates identification of interactions that occur with the ligand around the active site (Ko *et al.*, 2010). Information of binding affinities which is obtained from the molecular docking process is assessed and can be used to either improve on existing ligand or carry out de novo synthesis of new compounds. By use of computational methods, it is now possible to extract information about the molecular geometry of the active site of the protein. The knowledge obtained can also be used for synthesis of novel compounds and can be used to improve on existing ligands. In addition, there exists libraries that contain structures of

existing ligands and are available to the public (Pini *et al.*, 2004). Ideally a pharmacophore (all the information necessary for interaction between ligand and active site) is obtained so as to be able to design the novel compounds (Klebe, 2000). To obtain feasible novel structures a synthetic organic chemist is involved in the design process. Once designed and synthesized the compounds must then be tested and validated. Information on their binding properties is determined using biochemical, crystallographic and spectroscopic methods. The data obtained is then further analyzed using three-dimensional geometry of crystallographically characterized complexes.

1.4 PROBLEM STATEMENT, JUSTIFICATION AND HYPOTHESIS

HIV infection leads to progression with time to Acquired Immuno-Deficiency Syndrome (AIDS) which in turn leads to the death of millions of people. The highest number reported of those newly infected is in Sub-Saharan Africa, 1.8 million [1.6 million–2.0 million] people as per 2009 (UNAIDS, 2010). Progression to AIDS is due to the fact that the HIV infects CD4 cells leading to their death through direct viral killing, apoptosis in infected cells as well as killing of CD4 cells by CD8 cytotoxic lymphocytes (Kandathil e*t al.*, 2009, Wonderlich *et al.*, 2011). As aforementioned, HAART has provided an avenue to mitigate this problem and has reduced the death of people living with HIV/AIDS. However, HIV is known to have a number of resistance mechanisms to antiretroviral drugs, some which may be natural due to the poor proof reading capacity of the reverse transcriptase and others due to drug pressure. Resistance mechanisms attributed to drug pressure are referred to as primary mutations as they lead to direct drug resistance. Ultimately, this causes treatment failure and due to cross-resistance across the group there are few treatment options. Therefore, there is a need to develop a new class of inhibitors (Rhee *et al.*, 2010). Most of the research done has been on the HIV-1 subtype B and as such it has not targeted other subtypes including C which is predominant in Southern Africa (Jakobsen *et al.*, 2010). There is a lesser amount of information on the drug resistance mutations occurring in non-subtype B. Indeed, there is only evidence of the ARV treatment working on non-subtype B (Dierynck *et al.*, 2010). However by targeting the subtype C, drugs that are tailor made for subtype C maybe developed given the genetic variation from subtype B (Paraschiv *et al.*, 2011). The HIV protease is a prime target as its mutations are limited due to the fact that substrates bind to

the active site. Many mutations do not occur in the protease as they would affect the activation of important proteins rendering the enzyme useless (Castro *et al.*, 2011, Swanstrom and Erona, 2000). Drugs that fit within the substrate envelop would be rarely affected by mutations as mutations occur outside the envelope. In this study, HIV subtype C protease sequences provided by the National Institute for Communicable Disease (NICD) from thirty South African infants presenting with treatment failure as a consequence of drug resistance will be used to shed more light into the mutations occurring and how they affect binding of the drugs to the protease (Morris, 2011).

1.5 RESEARCH AIMS AND OBJECTIVES

The overall aim of the study is to build homology models of the HIV-1 Subtype C protease for the wild type and drug resistant mutants that are suitable for docking and molecular dynamics simulation studies.

1.5.1 Specific objectives

1. To identify and critically analyze the mutations occurring in the HIV-1 subtype C protease sequences.
2. To create three dimensional structure of the wild type HIV-1 subtype C protease using homology modeling techniques.
3. To model the common mutations using the wild type and create three dimensional mutant structures.
4. To analyze data for purposes of testing new protease inhibitors that bind to the active site.

CHAPTER 2

2 INTRODUCTION

Protease sequence data analysis

The aim of this chapter is to examine and analyze the mutations that occurred in the data that was provided by the National Institute for Communicable Diseases (NICD) (see supplementary disk, in a directory sequences). The data comprises of twenty nine protease sequences obtained from plasma of South African infants before and after treatment for HIV with antiretroviral drugs. The protease inhibitors that were used for treatment include: ritonavir and a fixed dose combination lopinavir and ritonavir branded as kaletra. Several tools were employed on HIV protease data for the analysis. The methods used include HIV databases for sequence search and mutation analysis, phylogenetic reconstructions and python programming.

2.1.1 Types of mutations

Change is an inevitable constant that is best reflected in nature by mutation processes which could be a consequence of natural evolutionary mechanisms. These mechanisms in addition to selective processes are what have led to the great genetic diversity of HIV (Jakobsen *et al.*, 2010). Mutations in HIV have been classified as polymorphic or non-polymorphic. Polymorphic mutations can be defined as those mutations that occur frequently in absence of selective drug pressure. On the other hand non-polymorphic ones do not occur in absence of drug therapy. With respect to drug resistance, mutations can be classified into two types; (I) Primary resistance that occurs in treatment naïve patients as a result of drug resistant virus infecting them. These mutations are challenging in the sense that, it is difficult to distinguish between what is polymorphic and not as they occur in absence of drug therapy. (II) Acquired or secondary resistance which is due to selection in presence of drug therapy (Shafer *et al.*, 2007). Mutations are also classified based on their effective resistance. A classic example of this is major and minor mutations. Major mutations are those which are known to cause drug resistance on their own while the minor/accessory have little effect on their own.

Minor mutations occurring in association with major mutations decrease susceptibility to protease inhibitor (Rhee *et al.*, 2003).

There are a number of naturally occurring polymorphisms that distinguish the HIV subtype B from C. These occur in absence of drug therapy and are found at 8 positions T12S, 115V, L19I, M36I, R41K, H69K, L89M, I93L (Bessong, 2008a). The location of these polymorphisms is not within the binding site and biochemically they are known to have little significance to clinical drug resistance. However from experimental results it was observed that they could contribute to low grade resistance by decreasing binding constants and in presence of inhibitors lead to development of primary mutations in the active site (Coman *et al.,* 2008). The concept of naturally occurring polymorphisms together with mutations may lead to decreased drug susceptibility (de Medeiros *et al.*, 2011). D30N in association with M36I and A71V are mutations that are known to potentiate each other when they occur together leading to decreased drug susceptibility (Clemente *et al.*, 2003). A number of mutations have been observed to occur in clinical isolates after treatment with protease inhibitor in subtype C at positions 30, 46, 82, 90. Other mutations of interest in subtype C protease are at the following positions L10I/V, K20R/M/I, M36I/L/V, M46L, L63P, A71T and V77I (Shafer *et al.*, 2007). These are commonly observed in drug resistance surveillance but occur in absence of drug exposure.

A great proportion of mutations occur outside the active site which consists of residues in the region 25–32, 47–53 and 80–84 (Ohtaka *et al.*, 2004). Any mutations occurring outside of this region are referred to as non-active site mutations. The essence of selection is survival and thus mutations in active site are rare as they would affect catalytic activity in the active site. It is known that almost half of the residue positions in the protease undergo substitutions in different combinations leading to resistance to inhibitors. It is not clearly understood how they contribute to decreased inhibitor binding (Ali *et al.*, 2010, Wu *et al.*, 2003). Experimental results show that active site mutations in combination with non-active site mutations contribute to greater resistance than when they occur singularly. However non-active site mutations have shown a greater decrease on inhibitor binding rather than active site mutations alone (Muzammil *et al.*, 2003). This has led to two hypotheses on how the mutations affect the protease.

One school of thought is the residue substitutions affect the geometry of the active site and since the inhibitors are known to be rigid they cannot adapt to this change. Current inhibitors are designed under the old lock and key paradigm and this leads to a poor fit at the active site (Ode *et al*., 2005). The other mechanism by which the residue changes are thought to affect the protease is through conformational changes that occur when an inhibitor binds at the active site. The mutants are stabilized in the native state by the residue substitutions and this means that extra energy expenditure is needed in order for the changes to occur. However, the first hypothesis is greatly supported as the effect on the energetics of conformational change is not profound.

2.1.2 Phylogenetic analysis

Phylogenetics entails the study of evolutionary relatedness of organisms and further representing that information in form of trees. Phylogenetics has many applications such as taxonomic classification, epidemiological studies, paleoanthropology, studying of gene families among many others. In HIV studies it has been used to study the origin and evolution of the virus from other species and its relationship to date in addition to taxonomic classification (Sharp and Hahn, 2010). In bioinformatics the data that is used to classify organisms is molecular data mainly DNA and amino acid sequences. RNA data can also be used in place of DNA (Li *et al*., 2011a). Molecular data is advantageous due to the fact that it serves as a good record to the evolutionary process as mutations are recorded in sequence data. It is also easier to obtain as compared to fossil records which may not be well maintained and may have sampling bias. The process of creating phylogenetic trees takes places in the following steps: (1) Identifying molecular data (2) multiple sequence alignment (3) choosing the model of evolution (4) choosing tree building method (5) validation of the tree. Each step is critical in generating the correct tree as the wrong choice in identifying data, incorrect sequence alignment as well as unsuitability of the model of evolution may result in incorrect trees.

2.1.2.1 Identifying Molecular data

In the first step the molecular data consisting of nucleotide or protein sequences is obtained. The intent of the analysis of the data determines the sequences to be used. To elaborate this, nucleotide data is good for closely related organisms as it records every event whereas protein data is better for more distantly related organisms. The degeneracy of genetic code affects protein analysis for rapid changing sequences. This means that there is conservation of residues with change in nucleotides due to more than one codon coding for same residue. The end result is that there will be no change for amino acid residue but it is occurring at nucleotide level (Higgs and Attwood, 2005). The advantages and disadvantages of the using the different data have been summarized in the table below.

Table 2.1: Advantages and disadvantages of nucleotide and protein sequence for phylogenetic studies.

ADVANTAGES	DISAVANTAGES
NUCLEOTIDES	
1. **Good record for all mutations.**	1. Less sensitive to alignment due to four nucleotides versus twenty amino acids it
2. **It better for closely related organisms and for rapidly mutating organisms.**	2. Each organism has its own rate of use of codon leading to bias.
PROTEIN SEQUENCES	
1. **Good for slowly mutating data.**	1. Missing critical information as it shows only non-synonymous mutations.
2. **Has less noise compared to DNA sequences.**	2. Not useful for some gene studies as only the DNA sequence is altered.

2.1.2.2 Multiple sequence alignment

Multiple sequence alignment organizes sequence data on basis of nucleotide or residue position similarity. This is the first and critical step for construction of phylogenetic trees. Several programs can be used and alignments compared so as to select the best. Multiple sequence alignment can be done with several programs which include T-coffee (Notredame *et al.*, 2000), Clustal (Larkin *et al.*, 2007) , MAFFT (Katoh *et al.*, 2002) and Muscle (Edgar, 2004).

The errors in the alignments should be manually checked so as to ensure accuracy and precision of the alignment (Wang *et al.*, 2011).

2.1.2.3 Models of evolution

Several models of evolution that are used correct for homoplasy which is defined as those sequence similarities that arise other than from a homologous relationship (Wake *et al.*, 2011). There are substitution models that are optimized for nucleotide sequences such as Jukes cantor, Kimura and for amino acids, point accepted mutation (PAM) and Jones-Taylor-Thornton (JTT) (Wang *et al.*, 2008). The PAM is a model that was derived from substitution matrices constructed empirically from related protein sequences. The JTT matrices are an improvement on the PAM matrices.

2.1.2.4 Construction of the tree

Tree construction methods fall into two broad categories namely, simple and complex. The simple ones assume that there is little or no change in sequence sites while the more complex ones try to incorporate parameters that allow for the many possible changes in the sites. The simple methods are called clustering methods and are based on calculation of evolutionary distance from the alignments and use distance matrices. The methods in this group are neighbor joining and unweighted pair group method using arithmetic average (UPGMA) (Larkin *et al.*, 2007). The more complex methods also known as discrete methods use the alignments directly and avoid loss of information that occurs through conversion of sequence characters to distances. They include maximum parsimony, maximum likelihood and Bayesian methods (Zvelebil and Baum, 2008). Neighbor joining, UPGMA and maximum parsimony use minimum evolution principle. Maximum likelihood evaluates the tree topology from observed data that best fits a given model of evolution. Bayesian and maximum likelihood methods employ more statistical rigor and hence are considered to produce better results but are computationally intensive (Mar *et al.*, 2005).

Phylogenetic trees can be rooted or unrooted. A rooted tree shows the direction of evolution from the common ancestor and is more informative while an unrooted one only highlights the evolutionary relationship. There are two ways of rooting a tree; molecular clock

hypothesis and out-group criterion. Molecular clock hypothesis assumes a clock like behavior and roots trees from the assumption that all sequences evolve at the same rate. New techniques have been incorporated such as the relaxed molecular clock so as to account for variation in evolution time (Kimura, 1980, Maljkovic-Berry *et al.*, 2009). The out-group criterion on the other hand uses a sequence homologous to those of interest but separated at an earlier date during the evolution process (Lyons-Weiler *et al.*, 1998).

2.1.2.5 Validation of the tree

The final step involves the evaluating the reliability of the phylogenetic tree. Statistical methods are employed at this stage. Two main ones can be used there is bootstrapping and jackknifing. In bootstrapping new datasets are created by perturbing the original dataset. New trees are generated so as to investigate the phylogenetic relationship. If the phylogenetic relationship is strong then the slight perturbations should not affect the reproducibility of the tree and therefore the tree can be deemed to be precise. Two techniques can be used to introduce the perturbations. Non parametric bootstrapping entails the replacement of the sites and is a random process as opposed to that in parametric bootstrapping that follows a particular distribution (Makarenkov *et al.*, 2010). Jackknifing involves deleting one half of the sites from original dataset and creating new ones. The new data sets are then used for constructing the phylogenetic trees in the same way as the original. Bayesian inference on the other hand uses Markov Chain Monte Carlo method which is an iterative random sampling strategy. It is a procedure which seeks for a high likelihood score as it searches for the tree topologies. The optimal trees from this process are used to create a final consensus tree. Bayesian method has a higher probability of finding an optimum tree as it uses many optimal trees with a probability of being the true tree while maximum likelihood has only one tree (Huelsenbeck *et al.*, 2001, Mar *et al.*, 2005)

Commonly used methods in HIV phylogeny are neighbor joining and maximum likelihood. Neighbor joining is a quick method and gives a good idea or tree topology. Maximum likelihood however is a more rigorous method incorporating statistics and therefore greater confidence in the nodes for the tree branches (Lam *et al.*, 2010).

2.2 METHODOLOGY

The Figure 2.1 is a schema that shows the stepwise procedures that were followed and programs that were used to implement the methodology for this chapter.

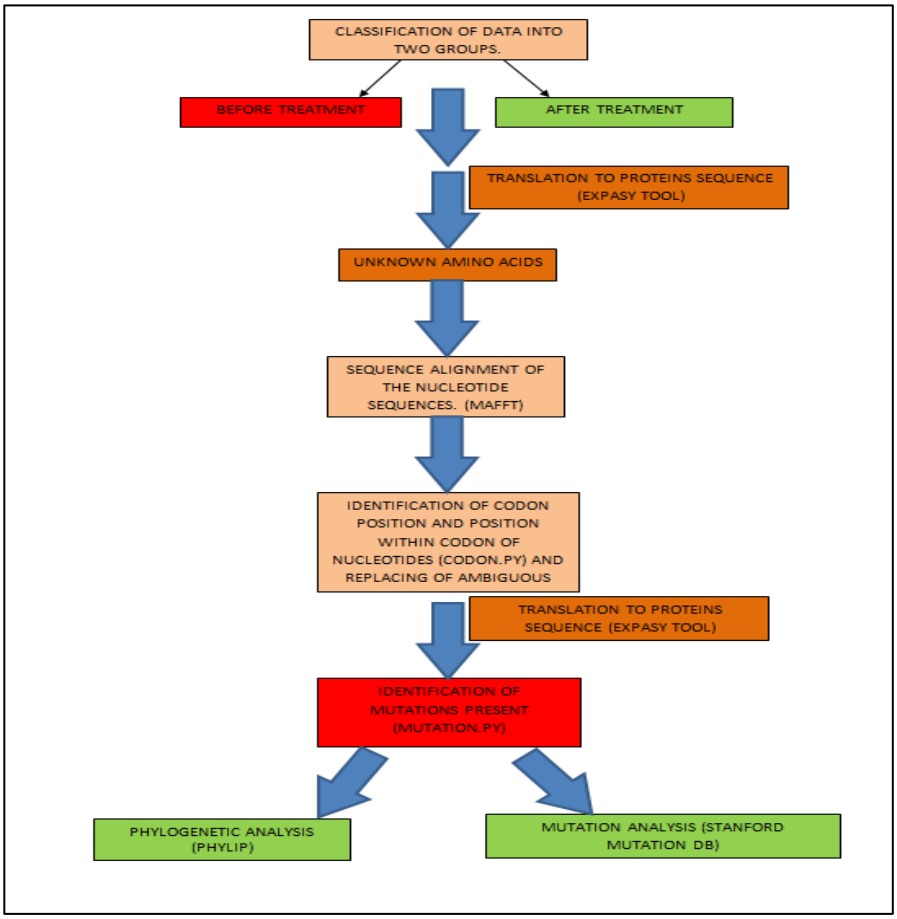


Figure 2.1: Flow diagram illustrating data analysis of twenty nine infant HIV sequences implemented using various programs

The online sites and tools used for Section 2.2.2 – 2.2.6 are presented in Table 2.3.

Table 2.2: Hyperlinks to online tools used in the methodology section.

WEBSERVER	URL
Expasy	http://web.expasy.org/translate/
IUPAC code	http://www.bioinformatics.org/sms/iupac.html
Los Alamos (Sequences)	http://www.hiv.lanl.gov/content/sequence/NEWALIGN/align.html.
Los Alamos (Gene Cutter)	http://www.hiv.lanl.gov/content/sequence/GENE_CUTTER/cutter.html
Stanford HIV database	http://sierra2.stanford.edu/sierra/servlet/JSierra?action=sequenceInput
Phylip (PHYML 3.0)	http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::phyml
MAFFT version 6	http://mafft.cbrc.jp/alignment/server/

2.2.1 Antiretroviral drug regimen for infants

The Table 2.3 shows information on the codes, treatment period and drug regimen for South African infants provided by the NICD. The infants were treated with the protease inhibitor ritonavir for a period of six months after which they were switched to Kaletra a fixed dose combination of lopinavir and ritonavir in addition to the NRTI's.

Table 2.3: The HIV protease sequences showing codes, treatment period, and drug regimen for South African infants as provided by the National Institute for Communicable Diseases (NICD).

SEQUENCE I.D.	TREATMENT PERIOD	DRUG REGIMEN
3018	12 months	Ritonavir and kaletra
3021	12 months	Ritonavir and kaletra
3043	other	-
3051	12 Months	Ritonavir and kaletra
3059	9 months	Ritonavir and kaletra
5014	24 weeks	-
5032	52 weeks	Ritonavir and kaletra
5045	9 months	Ritonavir and kaletra
5046	12 months	Ritonavir and kaletra
5074	36 weeks	Ritonavir and kaletra
5079	6 months	Ritonavir
5080	12 months	Ritonavir and kaletra
5086	52 weeks pre-random	Ritonavir and kaletra
5089	36 weeks pre-random	Ritonavir and kaletra
5094	24 weeks	Ritonavir
5114	12 months	Ritonavir and kaletra
5117	12 months	Ritonavir and kaletra
5144	12 months	Ritonavir and kaletra
5166	9 months	Ritonavir and kaletra
5178	24 weeks	Ritonavir
5198	9 months	Ritonavir and kaletra
5207	6 months	Ritonavir
5211	12 months	Ritonavir and kaletra
5228	12 weeks	Ritonavir
5242	3 months	Ritonavir
5245	4 weeks	Ritonavir
5252	other	-
5261	12 weeks	Ritonavir
5175	no sample	Ritonavir

N.B: Treatment period for some of the sequences was not specified such as 3043, 5252. Patient 5175 did not present for sample collection after the first visit. Pre-random refers to other periods of sample collection apart from the scheduled ones.

2.2.2 Data acquisition and pre-processing

The nucleotide protease sequences were obtained from plasma of twenty nine South African infants and had data before and after treatment with antiretroviral drugs. Treatment failure was defined as > 1000 copies per ml of virus (see supplementary disk in a directory chapter II, subdirectory sequences). The sequences were translated using online Expasy tool (Gasteiger et al., 2005). Initial analysis was done on the sequences to check if the data was in order to proceed to the next stage. The ambiguous nucleotides present were located by use of python programming language script (codon.py) (see supplementary disk, in a directory chapter II, subdirectory sequences in scripts.txt). The script located the position of the codon in the sequence and the position of ambiguous nucleotides within the codon. The ambiguous nucleotides were then identified from IUPAC code (see supplementary disk in a directory chapter II, subdirectory sequences in script outputs.docx). The nucleotides were observed for sequence conservation in their position. The nucleotide which was the most conserved at a specific position was then replaced for the ambiguous nucleotide codes. The processed nucleotides were translated into protein sequences using the Expasy tool.

2.2.3 Identification of the consensus sequences

Consensus B and C protease sequences were obtained from the Los Alamos website. The alignment type chosen was consensus/ancestral and the region of the genome was the *pol* gene where the protease is coded. Another tool that was used was the gene cutter found at which extracted the protease out of the *pol* gene.

2.2.4 Identification of mutations

A python script (mutation.py) was used to identify the mutations that were occurring in the sequences (see supplementary disk, directory chapter II, subdirectory sequences, scripts.txt). The data was subjected to further analysis using the Stanford HIV database tool (Rhee *et al.*, 2003). The tool identifies the mutations as those that are different from subtype B. The nucleotide data was uploaded to the database tool which gave an output (see supplementary disk, chapter II directory, stanford_data).

2.2.5 Phylogenetic analysis

Phylogenetic analysis was carried out using maximum likelihood algorithms. The online version of Phylip (PHYML 3.0) (Felsenstein, 2005) was used. The nucleotide datasets were analyzed with Generalized Time Reversible (GTR) as the substitution model for nucleotides, proportion of invariable sites was set as none while the number of substitution rate categories was set at one and the gamma shape parameter was set to the default blank. For tree searching, the starting tree was set at the program's default. Nearest Neighbor Interchange (NNI) was employed as the type of tree improvement. 1000 bootstrap resampling were performed on our dataset for branch support (Guindon and Gascuel, 2003).

The protein datasets were analyzed with John Thornton Taylor (JTT) as the substitution model for protein, proportion of invariable sites was set as none while the number of substitution rate categories was set at one and the gamma shape parameter was set to the default blank. For tree searching, the starting tree was set at the program's default. Nearest Neighbor Interchange (NNI) was employed as the type of tree improvement. 1000 bootstrap resampling were performed on our dataset for branch support.

2.2.6 Pairwise sequence alignment of all sequences

The pairwise sequence alignment between sequences before and after treatment was done for all the protein sequences using MAFFT version 6 (Katoh *et al.*, 2002) online webserver with a BLOSUM62 matrix. The alignments were then viewed using Jalview version 2.7 (Waterhouse *et al.*, 2009) for a conservation plot of amino acid residues for each of the 29 sequences.

2.4 RESULTS

In this section, the results from the python program mutation.py, Stanford HIV database tool, phylogenetic analysis of nucleotide, protein sequences and pairwise sequence alignment of the data provided by the NICD are presented.

2.4.1 Comparison of mutations identified using global consensus sequence for subtype C and B before antiretroviral therapy.

The Table 2.4 shows the position of mutions (using subtype C global consensus sequence as the reference sequence) that occurred in the twenty nine HIV sequences from the analysis that was done using the python program (mutation.py). The data in Table 2.4 is before treatment of infants with antiretroviral therapy. Most of the mutations were found to be out of the active site with the exception of position V82I. Majority of the mutations were physicochemically conserved with the exception of V11D, K14N, K20M, N37S/K/E, K41N, Q61D/E, L63T/S and C67Y. The position that exhibited highest polymorphism was E35D appearing in 38% of the sequences, followed by K20R with 30%. The positions which were known to have more than one type of residue substitution were 12, 14, 19, 20, 36, 37, 41, 61, 63 and 64. Position 63 had four different residues proline, valine which are hydrophobic like leucine and threonine and serine which are small non polar that are polymorphisms. Threonine and serine have a low frequency (3%).

Table 2.5 shows the data obtained using the Stanford HIV database tool which is based on the HIV subtype B. Mutations are identified on basis of variation in subtype B for the candidate subtype input. Data presented in Table 2.5 is before antiretroviral therapy. There are no major or minor mutations that are reported from the sequences except sequence 5207 L10IL. In comparison to the mutations subtyped with global subtype C reference (Table 2.4) there are certain differences that are observed such as the mutations at given positions; for the naturally polymorphic positions they are noted as mutations in the report generated from the Stanford mutation database tool as they are different from those in subtype C. This is noted at positions T12S, I15V, L19I, M36I, H69K, L89M, I93L. Rare mutations occurring in the Stanford data were found in two sequences 5045 which is K45KR while there other was in sequence 5261 at M36ILM, N37NT, and K70KR.

Table 2.4: Position and type of mutant residues identified using the global subtype C consensus sequence as a reference (before antiretroviral treatment of infants). Last column to the right shows frequency of mutant residues that occurred at each position.

POSITION OF AMINO ACID	CONSENSUS RESIDUE	MUTANT RESIDUE	FREQUENCY
11	V	D	0.03
12	S	T/A	0.21/0.03
13	I	V	0.10
14	K	N/R	0.03/0.03
15	V	I	0.07
16	G	E	0.10
19	I	L/V/T	0.07/0.07/0.21
20	K	R/M	0.31/0.03
35	E	D	0.38
36	I	M/L	0.07/0.10
37	N	S/K/E	0.03/0.10/0.03
41	K	R/N/I	0.14/0.03/0.03
57	R	K	0.03
60	D	E	0.14
61	Q	D/E	0.03/0.03
62	I	V	0.03
63	L	P/V/T/S	0.21/0.21/0.03/0.03
64	I	M/L	0.03/0.03
67	C	Y	0.07
70	K	R	0.03
71	A	T	0.03
74	T	S	0.17
77	V	I	0.07
82	V	I	0.10
89	M	I	0.07
93	L	I	0.03

KEY

Small nonpolar	G, A, S, T	**Orange**
Hydrophobic	C, V, I, L, P, F, Y, M, W	**Green**
Polar	N, Q, H	**Magenta**
Negatively charged	D, E	**Red**
Positively charged	K, R	**Blue**

Table 2.5: Results of the NICD sequences from analysis with the Stanford HIV database tool for infants before receiving antiretroviral therapy.

SEQUENCES	MAJOR MUTATIONS	MINOR MUTATIONS	OTHER MUTATIONS
1. **3018**	**NONE**	**NONE**	V11D, T12S, I15V, L19I, M36I, R41K, H69K, L89M,
2. **3021**	**NONE**	**NONE**	K14N, L19T, E35D, M36I, N37K, R41K, Q61D, H69K, L89M, I93L
3. **3043**	**NONE**	**NONE**	T12S, L19I, K20R, M36I, R41K, L63V, H69K, L89M,
4. **3051**	**NONE**	**NONE**	T12S, I13V, I15V, L19I, K20R, M36L, N37S, L63P, H69K, V77I, L89M, I93L
5. **3059**	**NONE**	**NONE**	T12S, I15V, L19T, K20R, E35D, M36I, R41K, L63V, H69K, L89M, I93L
6. **5014**	**NONE**	**NONE**	T12S, I15V, L19I, H69K, T74S, V77I, L89M, I93L
7. **5032**	**NONE**	**NONE**	T12S, K14R, I15V, L19T, K20R, M36I, R41N, D60E, H69K, L89M, I93L
8. **5045**	**NONE**	**NONE**	T12S, I15V, L19I, E35D, M36I, R41K, K45KR, L63P, H69K, L89M, I93L
9. **5046**	**NONE**	**NONE**	T12A, I15V, L19T, M36I, R41K, L63V, I64M, H69K, L89M, I93L
10. **5074**	**NONE**	**NONE**	T12S, I13V, I15V, L19I, M36I, R41K, D60E, L63P, C67Y, H69K, V82I, L89M, I93L
11. **5079**	**NONE**	**NONE**	L10*, T12S, I15V, L19I, N37K, R41I, L63T, H69K,
12. **5080**	**NONE**	**NONE**	T12S, I15V, L19I, M36I, R41K, H69K, V82I, L89M,
13. **5086**	**NONE**	**NONE**	T12S, I15V, L19I, K20R, E35D, M36I, R41K, I62V, H69K, T74S, L89M, I93L
14. **5089**	**NONE**	**NONE**	T12K, I15V, G16E, L19T, M36I, R41K, H69K, L89M,
15. **5094**	**NONE**	**NONE**	T12S, I15V, L19I, M36I, R41K, L63AV, H69K, T74S, L89M, I93L
16. **5114**	**NONE**	**NONE**	I15V, G16E, L19I, K20M, M36L, N37K, R41K, H69K, L89M, I93L
17. **5117**	**NONE**	**NONE**	T12S, I15V, L19I, E35D, M36I, R41K, L63P, H69K, I93L
18. **5144**	**NONE**	**NONE**	I15V, L19I, M36I, R41K, H69K, L89M, I93L
19. **5166**	**NONE**	**NONE**	I15V, M36I, R41K, H69K, L89M, I93L
20. **5175**	**NONE**	**NONE**	T12S, I15V, L19I, M36L, N37E, L63V, H69K, L89M,
21. **5178**	**NONE**	**NONE**	T12S, I15V, L19V, M36I, R41K, H69K, K70R, V82IV, L89M, I93L
22. **5198**	**NONE**	**NONE**	T12S, I15V, L19I, E35D, M36I, R41K, H69K, T74S, L89M, I93L
23. **5207**	**NONE**	**L10IL**	T12S, I15V, L19I, K20R, E35D, M36I, R41K, C67Y, H69K, L89M
24. **5211**	**NONE**	**NONE**	I15V, L19I, K20R, E35D, M36I, R41K, D60E, H69K, L89M, I93L
25. **5228**	**NONE**	**NONE**	T12S, I15V, L19T, M36I, N37S, R41K, L63P, H69K, V82I, L89M, I93L
26. **5242**	**NONE**	**NONE**	T12S, I15V, L19V, K20R, E35D, M36I, R41K, R57K, D60E, Q61E, L63P, H69K, L89M, I93L
27. **5245**	**NONE**	**NONE**	T12S, I15V, L19I, E35D, M36I, R41K, L63V, H69K, T74S, L89M, I93L
28. **5252**	**NONE**	**NONE**	T12S, I13V, I15V, G16E, K20R, E35D, M36I, L63S, I64L, H69K, L89M, I93L
29. **5261**	**NONE**	**NONE**	T12S, I15V, L19I, M36ILM, N37NT, R41K, H69K, K70KR, L89M, I93L

2.4.2 Comparison of mutations identified using global consensus sequence C and B after antiretroviral therapy.

Comparing Table 2.4 to 2.6 there are new positions with mutations such as 10, 23, 45, 46, 54 and 78. There was an increase in number of mutations after antiretroviral therapy. The new mutations had only one type of residue substitution with the exception of position 10 was highly polymorphic (L10F/M/I) with more than one residue at that position while the rest had single mutations. Position 23, 45 and 78 has a very low frequency of mutation in the 29 sequences with an occurrence of 3%.

The following observations were made for Table 2.6; there was an increase in frequency of mutations at positions 13, 20, 60, 63, 67, 70, 71, 77, 78, and 90. In comparison with Table 2.4 there was a new mutation that occurred at position V82A after initiation of drug therapy and this is in the active site. There was only one position that was noted to have a decrease in the mutation frequency which is position 74.

In the analysis using the Stanford HIV mutation database tool (Table 2.7) major mutations that appeared were M46I, I54V, V82A, and L90M. These mutations were also noted in Table 2.6 where the mutations were identified using the global consensus sequence for subtype C as a reference. Minor mutations present were L23I and L10F/I which were similar for both Table 2.6 and 2.7. Other minor mutations appearing in Table 2.7 were F53L and L33FL but were not present in the Table 2.6. There were mutations in the Stanford mutation database tool analysis (Table 2.7) that were referred to as other mutations. Most of these mutations that occurred were similar to those in Table 2.6. There were some mutations that were noticed that were different such as E21K, Q61H, and Q18P.

Table 2.6: Position and type of mutant residues identified using the global subtype C consensus sequence as a reference (after antiretroviral treatment of infants). Last column to the right shows frequency of mutant residues that occurred at each position.

POSITION OF AMINO ACID	CONSENSUS RESIDUE	MUTANT RESIDUE	FREQUENCY
10	L	F/M/I	0.07/0.03/0.03
12	T	A/K	0.17/0.03/0.03
13	I	V	0.14
14	K	R	0.03
15	V	I	0.07
16	G	E	0.10
19	L	T/V	0.07/0.17/0.07
20	K	R/M	0.34/0.03
23	L	I	0.03
35	E	D	0.38
36	I	L/M	0.10/0.07
37	N	S/K/E	0.14/0.10
41	K	R/N/I	0.14/0.03/0.03
45	K	R	0.03
46	M	I	0.10
54	I	V	0.21
57	R	K	0.10
60	D	E	0.17
61	Q	D/E	0.03/0.03
63	L	P/V/T/S	0.34/0.10/0.07/0.03
64	I	M/L	0.03/0.03
67	C	Y	0.07
70	K	R	0.07
71	A	T	0.07
74	T	S	0.10
77	V	I	0.10
78	G	R	0.03
82	V	A/I	0.28
89	L	M/I	0.10
90	L	M	0.03
93	I	L	0.03

KEY

Small nonpolar	G, A, S, T	**Orange**
Hydrophobic	C, V, I, L, P, F, Y, M, W	**Green**
Polar	N, Q, H	**Magenta**
Negatively charged	D, E	**Red**
Positively charged	K, R	**Blue**

Table 2.7: Results of the NICD sequences from analysis with the Stanford HIV database tool for infants after antiretroviral treatment with ritonavir and kaletra. Infants below six months were treated on ritonavir (501424, 50796, 509424, 517824, 52076, 52423, 5245w4, 526112).

SEQUENCES	MAJOR MUTATIONS	MINOR MUTATIONS	OTHER MUTATIONS
1. 301812	I54V, V82A	L23I	T12S, I15V, L19I, K20R, E35D, M36I, R41K, L63P, H69K, L89M, I93L
2. 3021212	NONE	NONE	K14N, L19T, E35D, M36I, N37K, R41K, Q61D, H69K, L89M, I93L
3. 3043pre	NONE	NONE	T12S, L19I, K20R, M36I, R41K, L63V, H69K, L89M, I93L
4. 305112	M46I, I54V, V82A	L10F	T12S, I13V, I15V, L19I, K20R, M36L, N37S, L63P, H69K, V77I, L89M, I93L
5. 305152	M46I, I54V, V82A	L10F	T12S, I13V, I15V, L19I, K20R, E35D, M36L, N37S, L63P, H69K, V77I, L89IM, I93L
6. 503252	M46I, I54V, V82A	F53L	T12S, K14R, I15V, L19T, K20R, E35DE, M36I, R41N, D60E, L63P, H69K, L89M, I93L
7. 30599	NONE	NONE	T12S, I15V, L19T, K20R, E35D, M36I, R41K, L63V, H69K, L89M, I93L
8. 501424	NONE	NONE	T12S, I15V, L19I, H69K, T74S, V77I, L89M, I93L
9. 50459	V82A, L90M	NONE	T12S, I15V, L19I, K20KR, E35D, M36I, R41K, K45R, L63P, H69K, L89M, I93L
10. 504612	NONE	NONE	T12A, I15V, L19T, M36I, R41K, L63V, I64M, H69K, L89M, I93L
11. 507436	NONE	NONE	T12S, I13V, I15V, L19I, M36I, R41K, D60E, L63P, C67Y, H69K, V82I, L89M, I93L
12. 50796	NONE	NONE	L10M, T12S, I15V, L19I, E21K, N37K, R41I, L63T, H69K, I72R, I93L
13. 508012	NONE	NONE	T12S, I15V, L19I, M36I, R41K, H69K, V82I, L89M, I93L
14. 508612	M46IM, I54V, V82A	L10FI	T12S, K14R, I15V, L19I, K20R, E35D, M36I, R41K, Q61H, I62V, L63P, H69K, T74S, L89M, I93L
15. 50899	NONE	NONE	T12K, I15V, G16E, L19T, M36I, R41K, H69K, L89M, I93L
16. 509424	V82AV	NONE	T12S, I15V, L19I, M36I, R41K, L63T, H69K, T74S, L89M, I93L
17. 511412	NONE	NONE	I15V, G16E, L19I, K20M, M36L, N37K, R41K, H69K, L89M, I93L
18. 511712	NONE	A71T	T12S, I15V, L19I, E35D, M36I, R41K, L63P, H69K, I93L
19. 514412	NONE	NONE	T12S, I15V, L19I, M36I, R41K, D60E, H69K, L89M, I93L
20. 51669	NONE	NONE	I15V, M36I, R41K, H69K, L89M, I93L
21. 517824	NONE	NONE	T12S, I15V, L19V, M36I, R41K, H69K, K70R, L89M, I93L
22. 51989	I54V, V82A	L33FL	T12S, I15V, L19I, K20R, E35D, M36I, R41K, H69K, T74S, L89I, I93L
23. 52076	V82A	L10I	I15V, L19I, K20R, E35D, M36I, R41K, C67Y, H69K, L89M
24. 521112	I54V	NONE	I15V, L19I, K20R, E35D, M36I, R41K, D60E, H69K, L89M, I93L
25. 522812	NONE	NONE	T12S, I15IV, L19IT, M36I, N37S, R41K, L63PS, H69K, V82I, L89M, I93L
26. 5242	V82AV	NONE	T12S, I15V, L19V, K20R, E35D, M36I, R41K, R57K, D60E, Q61E, L63P, H69K, L89M, I93L
27. 5245w4	NONE	NONE	T12S, I15V, L19I, E35D, M36I, R41K, L63V, H69K, T74S, L89M, I93L
28. 5252pre	NONE	NONE	T12S, I13V, I15V, G16E, Q18P, K20R, E35D, M36I, L63S, I64L, H69K, L89M, I93L
29. 526112	NONE	NONE	T12S, I15V, L19I, E35D, M36IL, R41K, H69K, K70R, V82IV, L89M, I93L

2.4.1 Pairwise alignment of amino acid sequences

Pairwise alignment was done between sequences before and after treatment. There were sixteen aligned sequences that exhibited mutations (see Figure 2.2) after treatment while the other 11 aligned sequences had no mutations (see supplementary disk chapter II, in a directory named phylogenetics).The number of mutations occurring in the sequences ranged from 1 to7. The sequences which were shown not to have any mutations after treatment are 5074, 5245, 3021, 5178, 5014, 5080, 5261, 5046, 5089, 5114, 5166, 5117, 3059, 3043, 5242, and 5252. These sequences are represented in the phylogenetic tree in Figure 2.3 and Figure 2.4.

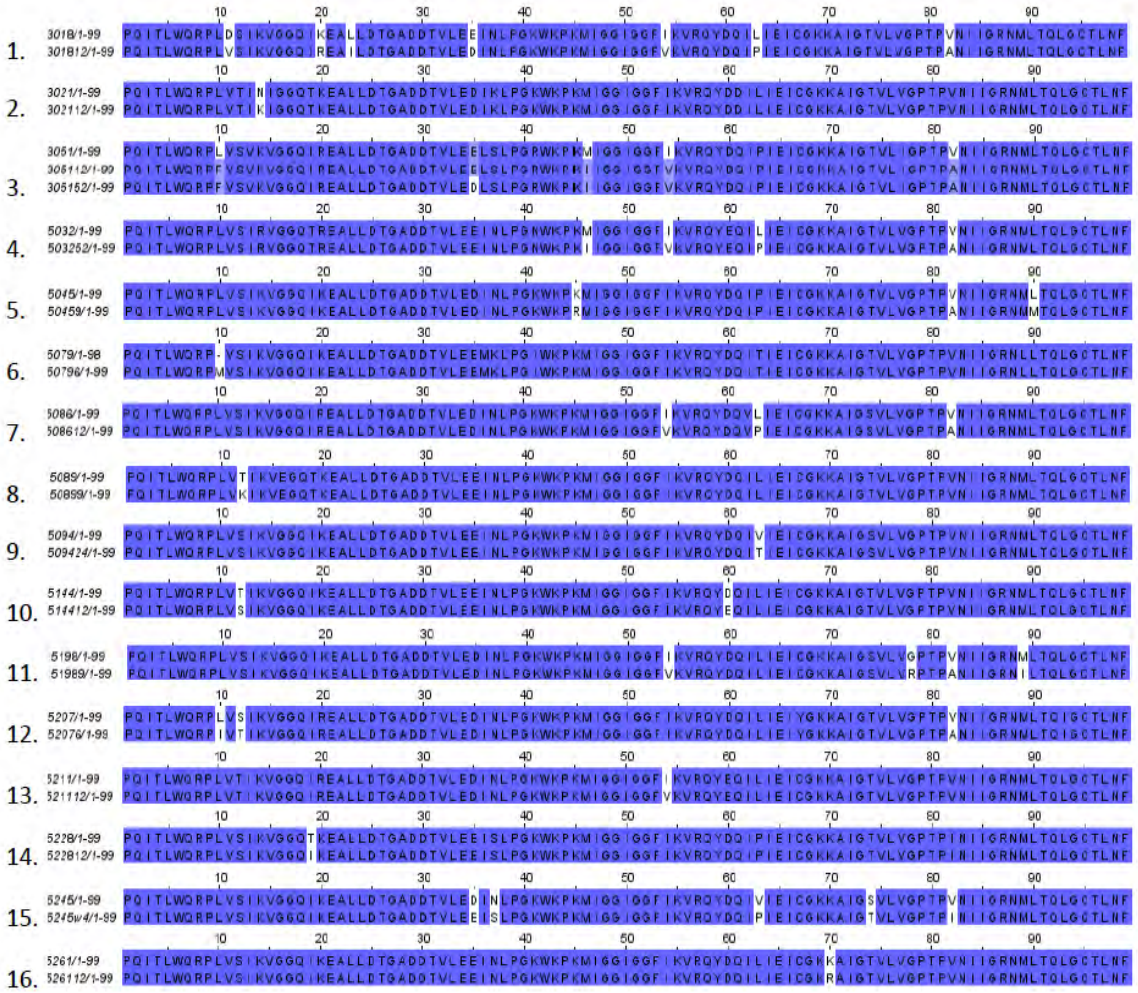


Figure 2.2: Pairwise sequence alignments showing the sequences that had mutations occurring after treatment.

2.4.2 Phylogenetic analysis

Phylogenetic analysis was done using both the nucleotide and protein sequence datasets and each of the datasets had the global consensus sequence for subtype C.

2.4.2.1 Nucleotide phylogenetic tree

The Figure 2.3 shows an unrooted phylogenetic tree that was constructed using a nucleotide dataset of infants before and after treatment with antiretroviral therapy. The sequences with four digit codes (e.g. 5074) are those before treatment while those after treatment have a five or six digit code (e.g. 50796, 507436). The two digits represent the period for which the infants were under antiretroviral therapy. In the Figure 2.3 some of the nucleotide sequences did not have any mutations after treatment with drugs. They include 5074, 5245, 3021, 5178, 5014, 5080, 5261, 5046, 5089, 5114, 5166, 5117, 3059, 3043, 5242 and 5252. A group of sequences form a cluster in which the subtype C consensus sequence is the origin and they all diverge from it. The sequence 5252 is distantly related to the subtype C consensus sequence.

2.4.2.2 Protein phylogenetic tree

The Figure 2.4 shows an unrooted phylogenetic tree that was constructed using an amino acid dataset of infants before and after treatment with antiretroviral therapy. The sequences with four digit codes (e.g. 5074) are those before treatment while those after treatment have a five or six digit code (e.g. 50796, 507436). The two digits represent the period for which the infants were under antiretroviral treatment. In the Figure 2.4 some of the amino acid sequences did not have any mutations after treatment with drugs. They include 5014, 5048, 5245, 5074, 5079, 3021, 5178, 5014, 5080, 5261, 5046, 5089, 5114, 5166, 5117, 3059, 3043, 5242 and 5252. A group of sequences form a cluster in which the subtype C consensus sequence is at the origin and they all diverge from it. Some sequences 5261, 5080, 514412 and 508012 are closely related to the subtype C consensus sequence as they are at the origin where the subtype C consensus sequence is found in the phylogenetic tree.

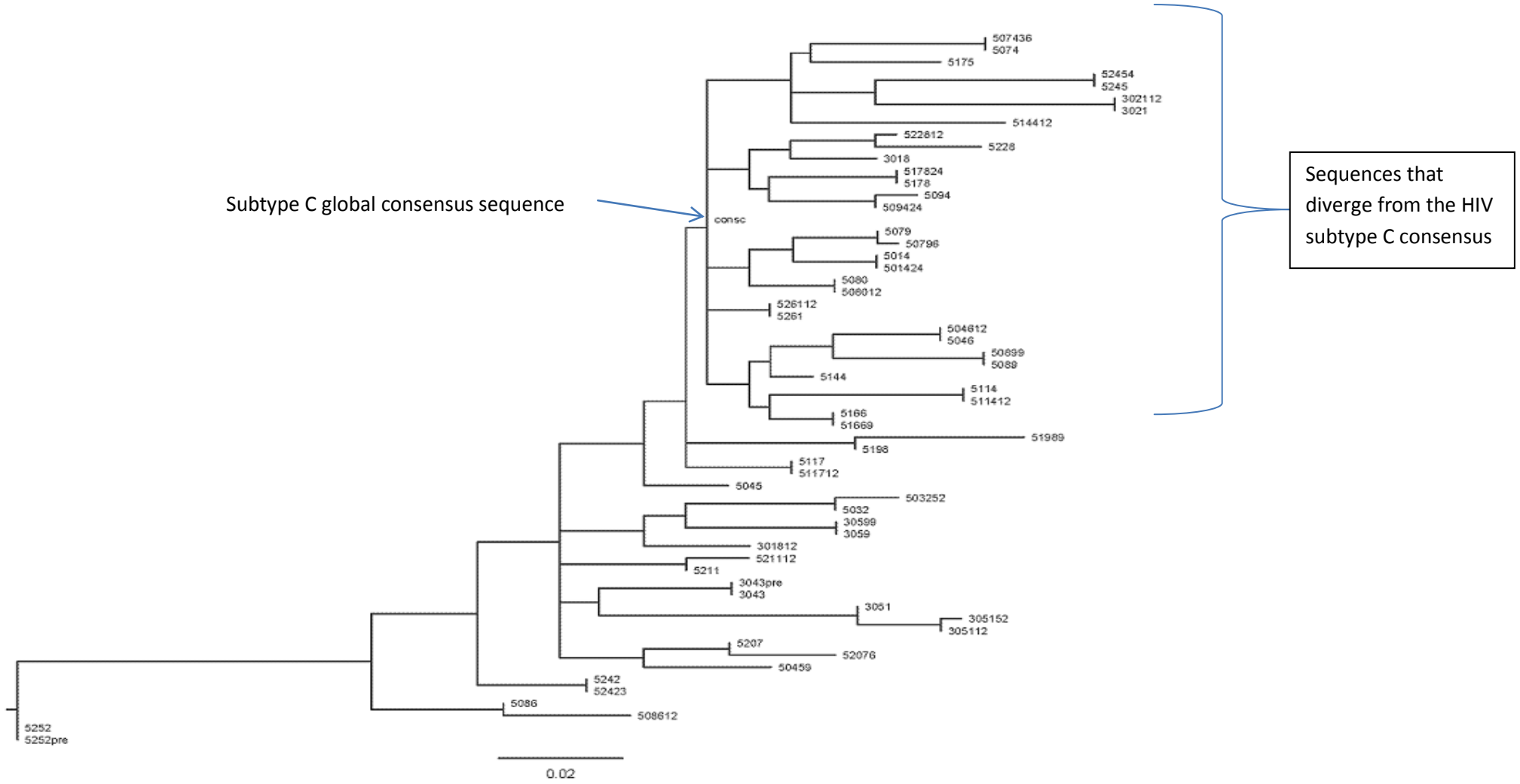


Figure 2.3: Phylogenetic tree constructed from nucleotide sequences which include consensus sequence for HIV subtype C, before and after antiretroviral treatment. Sequences before treatment are represented by 4 digit code while those after treatment by a 5 or 6 digit code. A group of sequences appear to be more closely related to the HIV-1 subtype C consensus sequence forming a cluster as highlighted in the diagram.

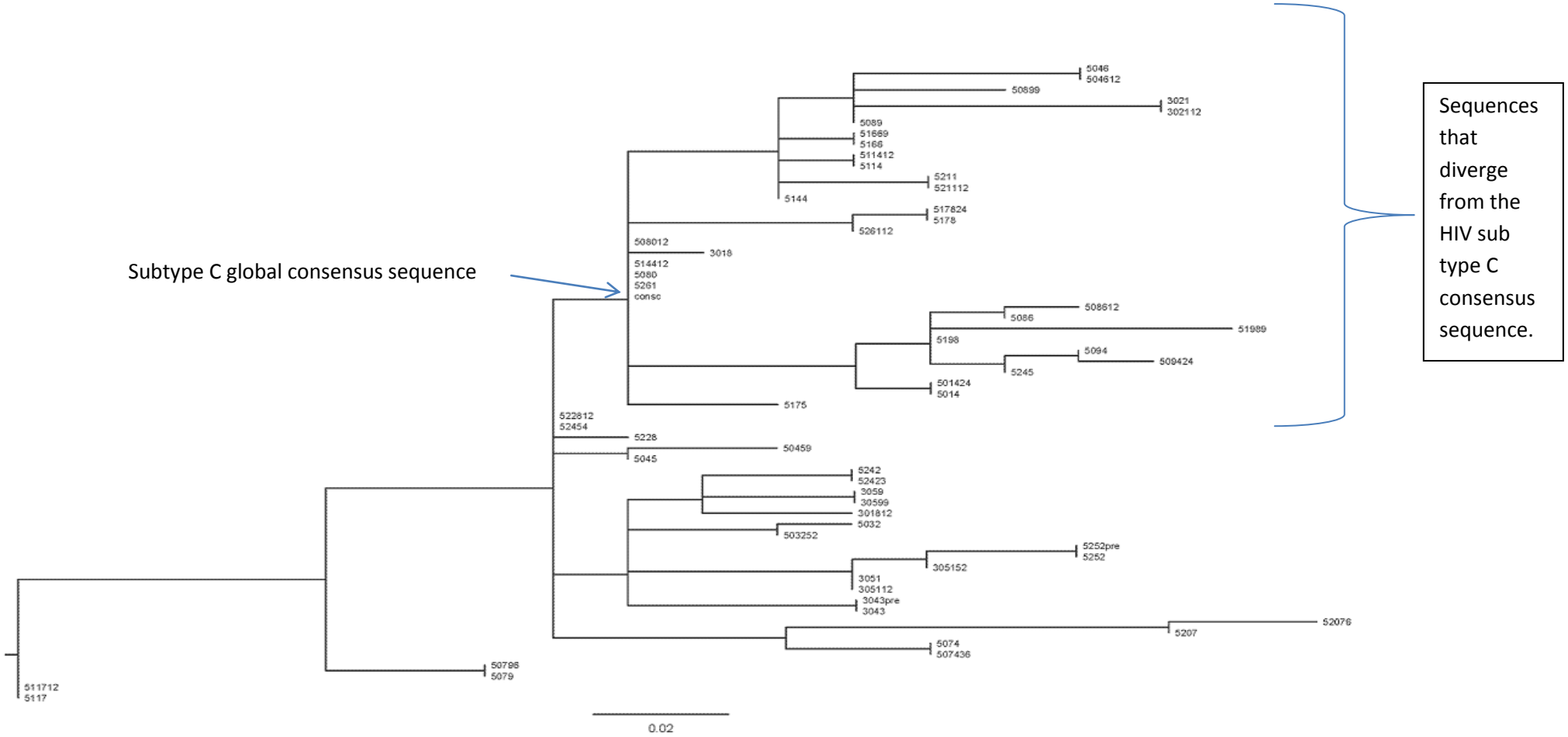


Figure 2.4: Phylogenetic tree constructed from protein sequences which include consensus sequence for HIV subtype C, before and after antiretroviral treatment. Sequences before treatment are represented by 4 digit code while those after treatment by a 5 or 6 digit code. A pattern where sequences related to consensus C form a cluster is represented by the curly bracket. There are some sequences which are very closely related to the subtype C consensus sequence namely: 5261, 5080, 514412, 508012.

2.5 DISCUSSION

2.5.1 Analysis of mutations identified using global consensus sequence C and B as references

The data provided was for South African infants who have experienced treatment failure. Mutations in the drug targets (in this case HIV protease) affect how the drugs interact with the active site and this leads to treatment failure. These mutations in the drug target(s) arise in an effort for the virus to survive (de Mulder *et al.*, 2011). It was therefore imperative to carry out analysis of the data so as to ascertain that there were mutations that occurred in the protease which was the drug target. In addition it was also important to identify which of these mutations contribute greatly to drug resistance especially in the subtype C and whether they vary from those in subtype B which is the mostly currently researched on (Kantor and Katzenstein, 2004).

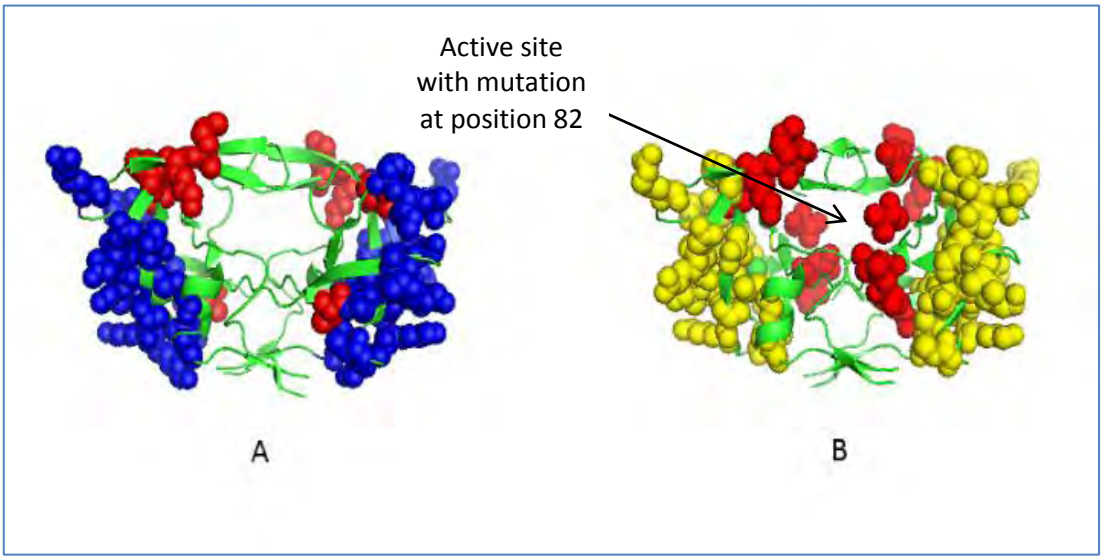


Figure 2.4: A mutation map with A showing the mutations before treatment while B shows those occurring after treatment using subtype C as the reference for the mutations. The areas sphere in red show mutations occurring in the flaps, active site and dimerization region. Blue spheres show mutations before treatment while yellow spheres are after treatment with antiretroviral.

A comparison of A and B in Figure 2.4 shows that there was an increase in the number of mutations occurring after treatment especially in the active site and in the interface between the two chains A and B. The mutations were mapped from the analysis done with subtype C as the reference for the mutations. The results revealed most mutations to be outside the active site with the exception of position 82. Mutations in the active site would

have detrimental effects to the catalytic activity of this enzyme. The protease can still cleave the substrates and maturation of the virus with very low catalytic activity (Šašková *et al.*, 2008).

The mutation E35D occurs in 38% of the sequences which is the highest from the analysis. In previous studies it has been shown to be an important position that forms a salt bridge interaction between aspartate 35 and arginine at position 57 (Kandathil *et al.*, 2009). The mutation at that position is known to cause a disruption in that glutamate has a γ methelene group while aspartate only has β methelene and this makes its side chain shorter. However it is disruptive to the inhibitors as it affects conformations in the drug binding pocket specifically valine at position 32 (Kandathil *et al.*, 2009). In the Figure 2.5 B, it can be observed that conformation of residues affects the formation of the salt bridge. In 2AQU the residues are in the right conformation to form a salt bridge whereas in 2HS1 the conformation of the residues GLU 35 and ARG 57 is different and the distance between the residues is 9.2 Å compared to 3.3 Å for 2AQU which interactions are too weak to form bonds (2AQU and 2HS1 are Protein Data Bank (PDB) codes).

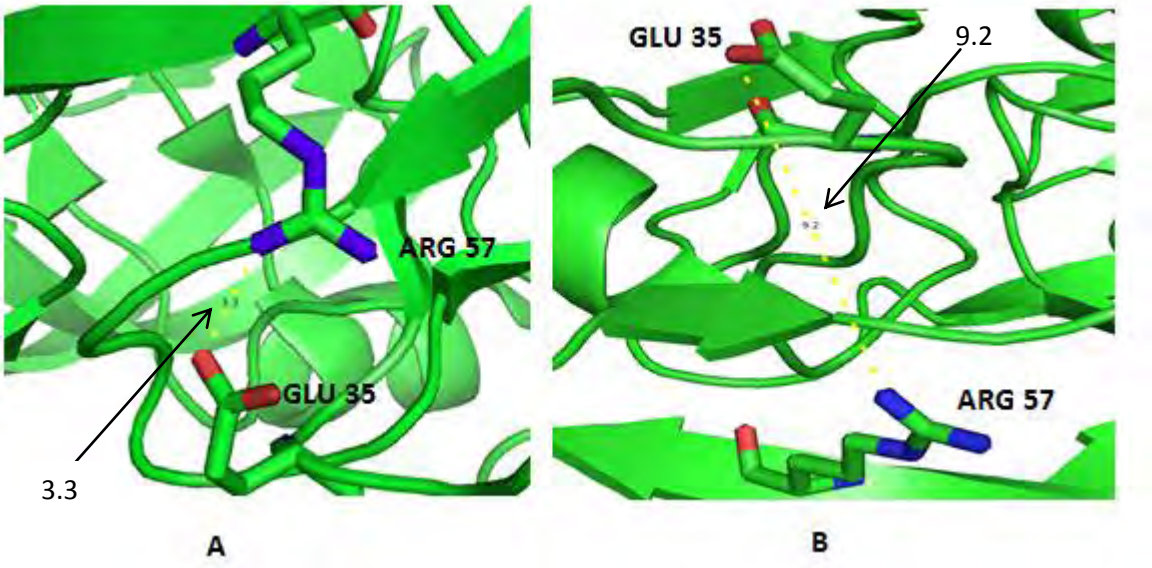


Figure 2.5: Illustration of salt brigde between Glu 35 and Arg 57 in 2AQU (A) and disruption of the salt bridge in 2HS1 (B).

A number of mutations are known to be of interest especially those that occur in the dimer interface region, active site and flap regions. Mutations at positions 10, 36, 37, 46, 57, 63, 71, 90 have been shown to affect binding of inhibitors experimentally by reducing the binding constants (Muzammil *et al.*, 2003). The above positions were found to have mutations in the sequences and hence it can be deduced that they could be causing the resistance in the subtype C. In the analysis carried out before treatment there were no major however there was one minor mutation L10I. This is expected as treatment naïve patients can only have mutations arising due to being infected with a drug resistant virus. Mutations that occur out of the active site are beneficial in the sense that they stabilize the mutants which are affected to a greater extent by environment conditions such as PH (Chang and Torbett, 2011). This would therefore make treatment options limited as there are already pre-existing mutations (Toor *et* al.).

Major mutations that occurred M46I, I54V and V82A are important as they are known to cause drug resistance. The mutations within the active site appeared after drug therapy was initiated. The mutations I54V and V82A lead to resistance especially to the drug ritonavir. Molecular dynamics simulations have shown that the active site cavity is reduced in size and this affects how ritonavir binds. This makes the conditions for binding less favourable by increasing the energy required for the inhibitor to bind. It is not clearly understood however how the mutation at position 54 affects inhibitor binding as it is far removed and hence no contact with inhibitor (Kumar and Jadhav, 2011).

There were some mutations that were noted not to be physicochemically conserved. The mutations however did not have high frequency of occurrence as this could lead to a change in the structure and consequently interactions that occur in the active site with both the substrate and the drug molecules. This would lead to the protease having a lower biochemical activity and 'fitness'. There were some mutations that were noted to be rare in occurrence as in sequence 5261 and 5045 and this is due to the fact that the mutations generated from the Stanford mutation database tool were from the raw nucleotide data and hence it may use a different algorithm to interpret the data. However in comparison most of the data was similar when compared from the two programs mutation.py and the Stanford mutation database tool.

2.5.2 Pairwise alignment and phylogenetic analysis

The pairwise alignment of amino acid sequences before and after treatment exhibited some sequences which had no mutations even after prolonged treatment and patients experiencing treatment failure. The phylogenetic analysis included sequences before and after treatment and revealed a pattern of similarity where sequences did not mutate even after drug treatment. These sequences are presented in Table 2.8. Since there were no mutations occurring in the sequences after treatment.

Table 2.8: HIV protease sequences that did not mutate even after showing treatment failure**.** The information is from pairwise alignment and phylogenetic analysis.

SEQUENCE BEFORE TREATMENT	SEQUENCE AFTER TREATMENT
1. **5166**	51669
2. **3043**	3043 PRE
3. **3059**	30599
4. **5046**	504612
5. **5117**	511712
6. **5178**	517824
7. **5074**	507436
8. **5079**	50796
9. **5080**	508012
10. **5014**	501424
11. **5245**	5245w4
12. **5252**	5252pre
13. **5242**	52423
14. **5114**	511412

The observation that there were no mutations occurring in the sequences after treatment could be a consequence of the sequences having primary drug resistance.

Phylogenetic analysis of both nucleotide and protein sequences showed that there were some sequences related to the HIV-1 subtype C consensus sequence. The trees however are not rooted hence we cannot tell the direction of evolution but we can understand how related the sequences are to each other. The phylogenetic trees show a branching pattern from HIV-1 subtype C consensus sequence that forms a cluster in both the nucleotide tree and the protein trees. This shows that the candidate sequences have a closer evolutionary relationship as opposed to that do not cluster to the HIV-1 subtype C consensus sequence. In the protein sequence phylogenetic tree (Figure 2.4) 508012 and 514412 are intriguing as they are very closely related to the HIV-1 subtype C consensus

sequence however the patients exhibit treatment failure and were on therapy with Kaletra. It is possible that the sequences could be drug resistant due to active site mutation. However, the case of non-active site mutation raises the question as to how the resistance arises. This also brings into to question how effective the therapy is on the HIV-1 subtype C but further investigation is needed so as to draw a conclusion. This would be in terms of carrying out a larger study so as to increase statistical confidence.

The data pre-processing has revealed interesting mutations that are not common in subtype B that may alter the geometry of the active site and this may lead to reduced binding constants. It is imperative that investigation is done so as to know what occurs due to these mutations as clinically they are known to have led to treatment failure. Laboratory experiments are expensive, tedious and time consuming and require a lot of expertise. Computational bioinformatics techniques are therefore very useful in giving clues as to the effects of these mutations. Homology modeling is one such technique that involves use of crystal structures to build models (de Beer *et al.*, 2009). Docking is useful in the sense that it gives the approximate binding constants and this enables screening of drugs that maybe active against the mutations or those that affected by resistance. (Dhaliwal and Chen, 2009).

CHAPTER 3

3 INTRODUCTION

3.1 HOMOLOGY MODELLING

The main objective of this chapter was to carry out homology modeling of six sequences. The consensus sequences B & C used were obtained from the Los Alamos website and four sequences 3018, 3051, 301812, 305152 from the National Institute of Communicable Diseases (NICD) .The criterion for selecting sequences was based on drug resistance. The sequences 301812 and 305152 had the major drug resistant mutations M46I, I54V and V82A. The HIV-1 subtype B, C consensus sequences, 3018 and 3051 were selected as they are considered to have no drug resistant mutations. The computational software used was Modeller 9v7. The search tools used were HHpred, PDB search and PSI-BLAST.

3.1.1 Application of Homology modeling in drug design

Homology modeling is a technique that has been gaining a lot of importance since the early 1960's due to the fact that experimental techniques are known to have several constraints (Kendrew *et al.*, 1958, Di Luccio and Koehl, 2011). Nonetheless, experimental techniques are still important as the structures in the Protein Data Bank (PDB) are derived from those techniques and are required for homology modeling (Kirchmair *et al.*, 2008). Homology modeling also known as comparative modeling involves use of known crystal or NMR structures to create the theoretical models (Chakravarty *et al.*, 2011). The structures are selected on basis of their sequence evolutionary relationship to the target sequence which we would like to build a model from (Petsko, 2002).

As technology has improved so has the quality of structures that have been derived from homology modeling. The structures are comparable to those obtained by experimental techniques (Mahdi *et al.*, 2011). This has therefore increased the confidence in this technique and hence it has found a number of applications. One of the major applications is in the field of drug design. Drugs function by affecting normal physiological processes in targets. These targets are proteins and require to be studied for developing molecules

that either inhibit or enhance normal functioning of the biological processes (Dhaliwal and Chen, 2009). Homology modeling is involved at the different stages of drug discovery; target identification, target validation, lead identification and lead optimization (Hillisch et al., 2004, Grant, 2009). Protein structures have been used in analysis of the active site, studying compensatory mutations (leading to drug resistance), design of pharmacological tools for determining biological function and refining of selectivity of lead compounds (Grant, 2009). Another application is in NMR studies where experimental information has gaps that are filled in by structures from homology modeling (Verdone *et al.*, 2009).

Homology modeling has four basic steps that include template search, target sequence to template alignment, modeling and validation. Loop refinement is only done when necessary during the validation step.

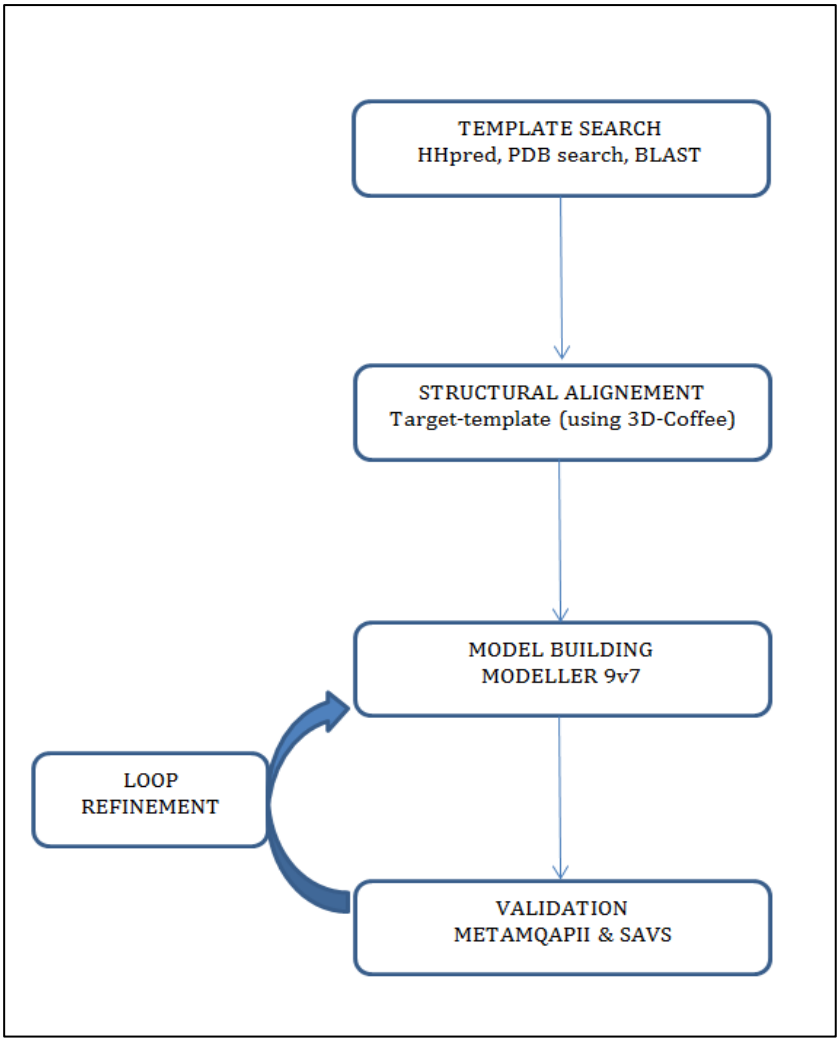


Figure 3.1: Diagram showing the steps required for calculation of homology models.

3.1.2 Steps in Homology modeling

Step 1: Template Search

The basis of 3D structure prediction is the premise that similar proteins sequences give rise to similar tertiary structures (Eswar *et al.*, 2007). In order to carry out homology modeling one uses a known sequence (target sequence) and searches for homologs with an experimentally derived 3D structure (template structure). The target sequence is input into the search tool such as BLAST (Altschul *et al.*, 1990), PDB search and HHpred (Soding *et al.*, 2005) . A protein database e.g. PDB is then searched for sequences with known structures using the search tool algorithms which do residue similarity searches (Hillisch *et al.*, 2004). Once templates are detected then using criteria such as high resolution, high sequence identity, secondary structure prediction and absence or presence of ligands are used to select suitable templates. The next step is to carry out alignment to optimize the correlation between template and target sequence.

Algorithms used in BLAST, PDB search and HHpred

Search tools such as BLAST and PDB search use a word method search which makes the search fast but is not as refined as other methods. It involves matching word pairs from the query sequence to those found in the database. Scoring matrices are then used to rate the matching pairs with a threshold T. The search for matching word pairs extends by pairwise alignment until the alignment score drops below a given threshold S due to gaps and mismatches. High scoring segment pairs consisting of continuous stretches of un-gapped aligned residues are then made from these pairwise alignments (Altschul *et al.*, 1998). Statistical significance is measured by an E-value (expectation value) for each alignment. E-value is calculated from the probability that the alignment occurs by chance (p-value), the size of the database and the length of sequence. The lower the E-value the greater the probability that the alignment is by random chance and thus can be deduced to have biological significance (Baxevanis and Ouellette, 2001).

PSI-BLAST (Altschul, 1999) follows the normal procedures for BLAST but includes building profiles consisting of Position Specific Scoring Matrices (PSSMs) from the first round of BLAST. These profiles undergo for more rounds of BLAST and thus the iterative process (Altschul, 1999). HHpred uses a far more refined algorithm as it uses Hidden Markov Models (HMM's). Initially a HMM profile is created from the target

alignment and the search for homologous templates is done from the weekly updated database PDB70. The ranking of the best database matches is done using the vertibi algorithm. A local HMM-HMM alignment is employed to further refine the search and makes use of the Maximum Accuracy (MAC) algorithm (Holmes and Durbin, 1998, Biegert and Söding, 2008). The algorithm optimizes the number of correctly aligned residues (Hildebrand *et al.*, 2009).

S**tep 2: Alignment**

The selected template(s) are then aligned with the target sequence (sequence we want to build our model for). This alignment should be optimal so as be able to carry out homology modeling in the most accurate way (Venclovas, 2001, Schonbrun *et al.*, 2002). Alignment can be done between the sequences or between a sequence and structure(s). There a range of programs that can be employed for sequence alignment and they include Clustal (Larkin *et al.*, 2007) , Muscle (Edgar, 2004) , and T-coffee (Notredame *et al.*, 2000) while for structural alignment tools such as Expresso (Armougom *et al.*, 2006) and Fugue (Shi *et al.*, 2001) are available.

Accuracy in this step is of ultimate importance as any error will result in incorrect structures. Misalignment of one residue results in an error of 4° in a model (Fiser, 2004). Sequence identity greater than 50 % results generally in correct alignment but as it drops to lower than 30% the alignments are likely to be error prone (Sunyaev *et al.*, 2004). Structural alignment is a better tool in cases of low sequence identity. It serves as a better guide as insertions and deletions are not placed in structurally unfavourable regions (Sauder *et al.*, 2000). When multiple templates are used it may lead to a more realistic picture as multiple sequence alignments provide a more true evolutionary relationship (Sokkar *et al.*, 2011). It is of great importance to carry out a visual inspection of alignments as the alignment programs are not always correct (Ginalski *et al.*, 2005).

Step 3: Model Building

A number of programs exist that can be used to calculate homology models such as SwissModel (Schwede *et al.*, 2003), SegMod (Levitt, 1992), Modeller (Sali and Blundell, 1993) amongst others. They use different methods such as rigid body (fragment assembly), segment matching (co-ordinate reconstruction) and satisfaction of spatial restraints (Al-Lazikani *et al.*, 2001, Grant, 2009). All these methods produce accurate models when the conditions are optimal and are therefore interchangeable. Modeller uses satisfaction of

spatial restraints as its model building technique (Sali and Blundell, 1993). The restraints which are expressed as probability density function include homology derived distance from Cα-Cα and N-O backbone and the dihedral angles (back bone and side chain). The modeller program tries to satisfy these restraints to produce a model with minimum violations (Wallner and Elofsson, 2003). Stereochemical restraints are satisfied by use of CHARMM force field (Sali and Blundell, 1993, Brooks *et al.*, 2009). Modeling of loops poses a fundamental problem especially where there is no alignment between target and template. This could be due to insertions or the fact that those regions in the target are different from the template. The prediction of the unaligned loops is not as accurate as those based on the alignment. There are two ways of loop building; use of databases and conformation search (Koehl and Delarue, 1995, Petrey *et al.*, 2003). Database search involves libraries with loops that fit that region. Conformation search involves generating many loops and energy is used as the criteria to select the best fitting loop. Modeller uses the second approach for modeling loops. However a combination of the two methods gives optimum loops with a range of 1.5-3.0 Å (Deane and Blundell, 2000, Deane and Blundell, 2001). Modeller is well established, fast and reliable as use of spatial restraints is less susceptible to alignment errors (Wallner and Elofsson, 2005, Qu *et al.*, 2009).

Step 4: Validation

This is the final step in homology modeling and it involves assessment of the quality of models that have been produced. The models are assessed on basis of similar characteristics with existing protein structures. There are two ways in which the evaluation can take place; looking at the physical properties such a steric and geometric criteria and conformational energy of the model (Sippl, 1995). Modeller provides a function for selecting the top ranking model using Discrete Optimized Protein Energy (DOPE) score. The DOPE score is based on enhanced reference state that relates to non-interacting atoms in a homogenous sphere. The radius of the sphere is dependent on the structure with the lowest energy and accounts for the spherical shape of the native structure (Shen and Sali, 2006). The DOPE score is arbitrary and is normalized by comparing it to different proteins and referred to as DOPE Z score.

There exists a number of webservers which check different parameters and give assessment on model quality. However there is no one program that assesses all parameters necessary for a model to be considered 'fit'. It is therefore important to use a combination of these different programs so as to give an evaluation of model accuracy

(Kihara *et al.*, 2009). Model Quality Assessment Programs (MQAPs) give an overall view of model accuracy. In order to give an overall prediction of model accuracy several "meta-predictors" have been developed. MetaMQAPII (see Table 3.1) is on such Meta –server that relies on assessment from other servers and gives a prediction of quality of computationally predicted models. It is designed based on a multivariate linear regression model where trivial parameters are controlled (Pawlowski *et al.*, 2008). Trivial features that are taken into account include global model quality, residue depth in structure, hydrophobicity and secondary structure assignment. MetaMQAPII calculates a composite of score from VERIFY3D (Luthy *et al.*, 1992), PROSA 2003 (Sippl, 1993), PROVE (Pontius *et al.*, 1996), ANOLEA (Melo and Feytmans, 1998), BALASNAPP (Krishnamoorthy and Tropsha, 2003), TUNE (Lin *et al.*, 2002), REFINER (Boniecki *et al.*, 2003), and PROQRES (Wallner and Elofsson, 2006). The overall model quality is presented by a score is referred to as Global Distance Total test score (GDT_TS). The score is an estimation of the greatest number of amino acid Cα atoms between the model and the true unknown structure that fall within defined cutoff distance of 1 Å, 2 Å, 4 Å and 8 Å. GDT_TS score is finally reported as an average of these scores (Cristobal *et al.*, 2001, Li *et al.*, 2011b). MetaMQAPII provides a PDB file which presents the erroneous regions in form of a color map on the protein with a spectrum of color from for red for bad regions and blue for good regions, a log file and a text file showing the raw evaluation data.

ProSA (see Table 3.1) is a protein structural analysis tool which calculates the pseudo energy profile of the model. It is based on statistical analysis of existing native structures in PDB. Two energies that are calculated for the model are distance based pair potentials and a potential based on residue solvent exposure. The two potentials are defined as a Z-score and a plot for deviation of residue energies known as knowledge based potentials (Sippl, 1993). The Z-score represents total energy deviation of the model from a particular energy distribution that is characteristic of native folds. Poorly folded structures deviate from this energy distribution and this is recognized by Z-score outside the range of the PDB structures (Sippl, 1995). The plot for deviation of residue energies is used to identify erroneous regions as those that have peaks greater than the base line zero as this are positive energies which are uncharacteristic of native protein folds. A color coded model with problematic areas is also shown as output from the ProSA webserver (Wiederstein and Sippl, 2007).

PROCHECK (see Table 3.1) assesses the stereochemical parameters of the protein structure and checks for deviation from the norm based on a comparison with crystal structures of high resolution. Some of the parameters include bond angles and bond lengths (Laskowski *et al.*, 1993). The results are presented as a number of plots; Ramachandran plot is one of them and is a graphical representation of psi against phi angles. The Ramachandran plot consists of four regions a most favorable, generously allowed, additionally and allowed and disallowed regions. A good quality model is one defined as having more than 90% of its residues in the most favored region. This is based on analysis of 118 structures with resolution of at least 2 Angstrom (Å), an R factor of no greater than 20% (Katiyar *et al.*, 2009).

Errors that occur during model building are due to misalignment of template and target, inserts and deletions, incorrect templates, side-chain packing and distortions and shifts in correctly aligned regions (Eswar *et al.*, 2007). During the validation step models which have unfavorable structures are taken for refinement as a whole or for loops. This involves correction of steric collisions and strains by use of energy minimization. It must be used with caution as it causes residues to shift from their correct positions when used excessively (Xiong, 2006). Refinement can be done on all atoms after building the models or on loops by specifying the residues which require the process.

Table 3.1: Model validation programs and servers.

PROGRAM/SERVER	URL
MetaMQAPII	https://genesilico.pl/toolkit/unimod?method=MetaMQAPII
ProSA	http://www.came.sbg.ac.at/prosa.php
PROCHECK	http://nihserver.mbi.ucla.edu/SAVES/

3.2 METHODOLOGY

3.2.1 Homology modeling

Step 1: Template search

The search for suitable templates was implemented using three database search tools namely HHpred, PSI-BLAST and PDB search. The sequences that were used as inputs for the template search include 3018, 3051, 301812, 305152, consensus sequences for HIV-1 subtype B and C. The pdb 70 was set as the HMM database and the alignment type was set at local for HHpred. BLAST had the following parameters were set; database (PDB), algorithm PSI-BLAST, and matrix BLOSUM 62. In PDB search the parameter for sorting templates was changed to resolution. The template 2HS1 was selected based on high resolution, sequence identity and presence of ligand.

Step 2: Target – Template alignment

Multiple sequence alignment for sequences 3018, 3051, 301812, 305152 and consensus sequence for HIV-1 subtype B and C was done using the web version of Clustalw. Structural alignment was carried out using the webserver Expresso for the six sequences in previous alignment to the suitable selected templates. The output for the alignment was in form of PIR files (see supplementary disk directory chapter III, pirfiles.txt). The files were manually adjusted to the modeller PIR format.

Step 3: Model building

Modeller9V7 package was used to build homology models for sequences 3018, 3051, 301812, 305152 and consensus sequences for HIV-1 subtype B and C. This step made use of different modeller scripts. Those used for model building were homodimer.py and model_m2.py (see supplementary disk, chapter III in a directory modeller scripts). The suitable templates that were selected had all heteroatoms removed by use of script model.write.py (see supplementary disk, chapter III in a directory preparation scripts).

(i.) Homodimer.py script

The modeller script (homodimer.py) was used which builds multichain models and introduces extra symmetry restraints such that chain A and B are constrained to have the same conformation. Each of the six sequences had a total of 500 models built. This was based on the rationale that with a greater number of models there is greater probability of DOPE Z scores being closer to the native state (-1). Slow refinement was performed on all the models.

(ii.) Model_m2.py script

The modeller script (model_m2.py) was used which builds models but has no extra symmetry restraints. Each of the six sequences had a total of 100 models built. This was based on rationale that with a greater number of models there is greater probability of DOPE Z scores closer to the native state. The number of models was reduced as building 500 models was computationally intensive and did not show much improvement in terms of DOPE Z scores. Slow refinement was performed on all the models. The model with the lowest DOPE Z score for each of the six sequences was then selected for loop refinement. The DOPE Z score were calculated using scripts calculate_dope.py, zdope_single.py, sort_zdope_scores.py (see supplementary disk, chapter III in a directory modeller scripts)

Step 4: **Validation**

MetaMQAPII, PROCHECK and ProSA were used to assess the quality of models for this study. MetaMQAPII and ProSA evaluate single chain structures. To rectify this, models PDB files were edited using a script chain_mod_renum.py (see supplementary disk chapter III, preparation directory). The script renumbers the chain B to A in the PDB files. Pymol (Schrodinger, 2010) software was used to visualize the results of the homology models.

Step 5: Loop refinement for Homodimer.py

Loop refinement was carried out on six models 3018, 3051, 301812, 305152 and HIV-1 subtype B and C. The criterion for selection of models for this step was based on the lowest DOPE Z scores for each of the six groups which had 500 models built. This was done for problematic loops and the script used was refine.py (see supplementary disk, chapter III in a directory modeller_scripts). The models with the lowest DOPE Z score after loop refinement were selected by use of a python script log_file_zscore_sorter.py.

Table 3.2: Template search and alignment programs and their websites

PROGRAM/SERVER	URL
HHpred	http: //toolkit.tuebingen.mpg.de/hhpred
PSI-BLAST	http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins
PDB-Search	http://www.pdb.org/pdb/home/home.do.
Clustalw	http://www.ebi.ac.uk/Tools/msa/clustalw2/
Expresso	http://tcoffee.crg.cat/apps/tcoffee/play?name=expresso

3.3 RESULTS

3.3.1 Multiple sequence alignment

Figure 3.2 shows the multiple sequence alignment of the six sequences 3018, 3051, 301812, 305152 and consensus sequences for HIV-1 subtype B and C against the templates 2hs1 and 1hxw. The residues are highly conserved in the whole alignment. Residue substitutions are physicochemically similar apart from position 7, 67, 69 and 95.

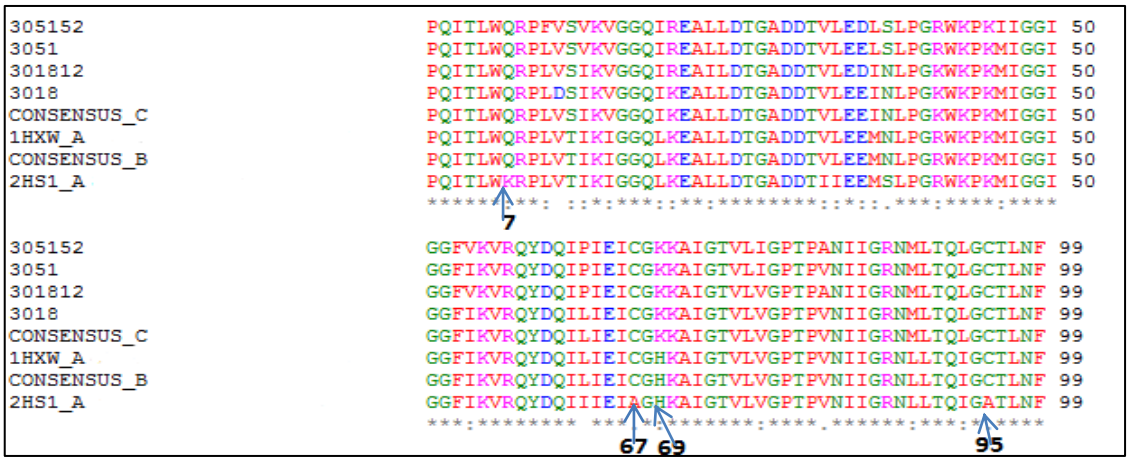


Figure 3.2: Multiple sequence alignment for consensus for HIV subtype B & C, patient sequences 3018, 3051,301812,305152, against templates 2hs1 and 1hxw done with Clustalw.

3.3.2 Percentage identity matrix

In Table 3.3 the percentage identity matrix shows very high percentage identities for the sequences relative to the templates. The highest percentage identity score is 100% for 1HXW to subtype B and 78% for 2HS1 to 305152.

Table 3.3: Percentage identity matrix showing the relative percentage identities of the six sequences for HIV-1 protease subtype B & C global consensus, 3018, 3051,301812,305152 and the two templates 2hs1 and 1hxw. All numbers in the matrix are percentages.

	1HXW	SUBTYPE B	SUBTYPE C	3018	3051	301812	305152
2HS1	92	92	84	83	83	79	78
1HXW		100	91	90	87	85	82
SUBTYPE B			91	90	83	85	82
SUBTYPE C				98	92	93	87
3018					91	92	86
3051						90	94
301812							90

3.3.3 Template search

Table 3.4 shows 2HS1 as the first result from HHpred and PDB search whereas PSI-BLAST returned different homologs for the sequences for HIV-1 Subtype B and C, 3018, 3051, 301812 and 305152. 2HS1 has a resolution of 0.84 Å and high GDT_TS score of 93.69. However for Subtype C, HHpred results yielded 3KA2 as the first result with a resolution 1.40 Å and a GDT_TS score 79.79 which are lower values than the 2HS1. The candidate templates had high percentage identities to the sequences with the highest being 100% for 5HVP for the subtype B sequence whereas the lowest was 79% for 305152 to 2HS1. All the candidates for templates had ligands found in their active site. The lowest E-value was 5.41E-68 while the highest was 6.2E-24.

Table 3.4: Results showing top hits for template search for HIV-1 protease sequences B & C consensus sequence, 3018, 3051,301812,305152. (ID represents identity).

	DATABASE	PDB ID	Percentage Identity	Resolution	E-value	GDT_TS	Ligand
Subtype B	HHpred, PDB search	2HS1_A	93%	0.84 Å	1.4E-33	93.69	Darunavir
	PSI-BLAST	5HVP_A	100%	2.0 Å	5.41E-68	93.31	Acetyl pep statin
Subtype C	HHpred	3KA2_A	83%	1.40 Å	2e-24	79.79	MVT-101
	HHpred, PDB search	2HS1_A	85%	0.84 Å	6.2E-24	93.69	Darunavir
	PSI-BLAST	2R5P_A	95%	2.30 Å	9.35e-65	92.93	Indinavir
3018	HHpred, PDB search	2HS1_A	84%	0.84 Å	5.9E -33	93.69	Darunavir
	PSI-BLAST	3D3T_ A	93%	2.80 Å	1.15E-63	90.91	Substrate p1-p6
3051	HHpred, PDB search	2HS1_A	84%	0.84 Å	1.8E-35	93.69	Darunavir
	PSI-BLAST	2R5P_A	92%	2.30 Å	6.99E-64	92.93	Indinavir
301812	HHpred, PDB search	2HS1_A	80%	0.84 Å	3.7E-38	93.69	Darunavir
	PSI-BLAST	3LZS_A	89%	1.95 Å	1.21E-62	92.92	Darunavir
305152	HHpred, PDB search	2HS1_A	79%	0.84 Å	1.5e-33	93.69	Darunavir
	PSI-BLAST	2WL0_A	84%	1.90 Å	1.14E-61	65.32	5AH

3.3.4 Homology modeling, loop refinement, model validation using homodimer.py

The following section includes the results of the six models for 3018, 3051, 301812, 305152, HIV-1 subtype B and C sequences.

3.3.5 MetaMQAPII

3.3.5.1 3018

The model that had the lowest DOPE Z score -1.49 was selected and the result from MetaMQAPII was a GDT_TS score of 90.40. The loop regions which were visualized to be problematic from results below were refined. The loops refined consisted of residues which included: 37-41, 50-51, 81-84,103-108,127-129,136-141 and 177-181. After refining the GDT_TS scores decreased to 87.88 and the DOPE Z score increased to -1.32.

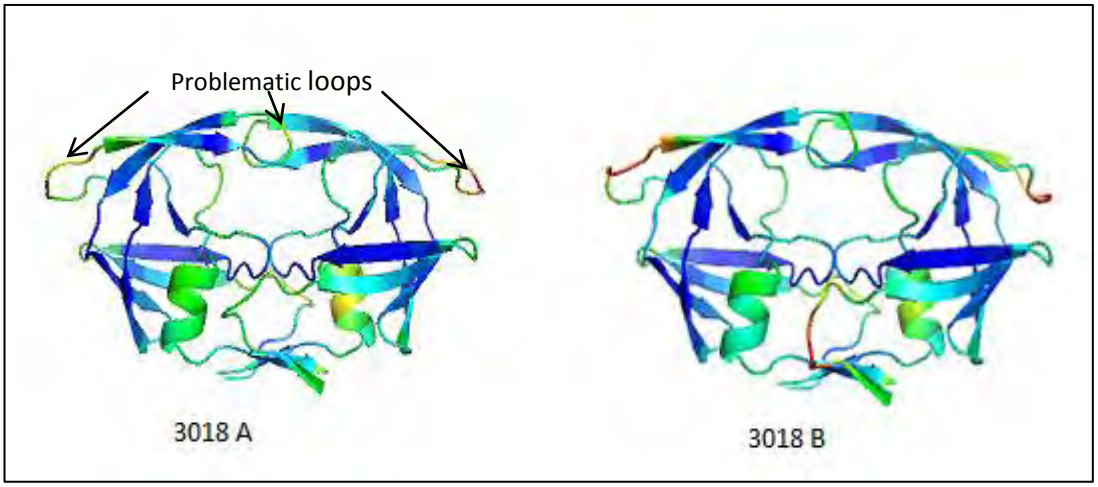


Figure 3.3: Homology models of sequence 3018. 3018A and 3018 B show the models before and after loop refinement respectively. Some of the poorly modeled loops highlighted in the diagram.

3.3.5.2 3051

MetaMQAPII returned a result of a GDT_TS score of 86.74 for the selected model which had a DOPE Z score of -1.14. The loop regions that were refined were those visualized from MetaMQAPII results and seen to be problematic. The loops refined consisted of residues which included: 37-41, 50-51, 81-84, 103-108, 127-129, 136-141 and 177-181. After refining the GDT_TS scores increased marginally to 86.99 and the normalized score improved to -1.31.

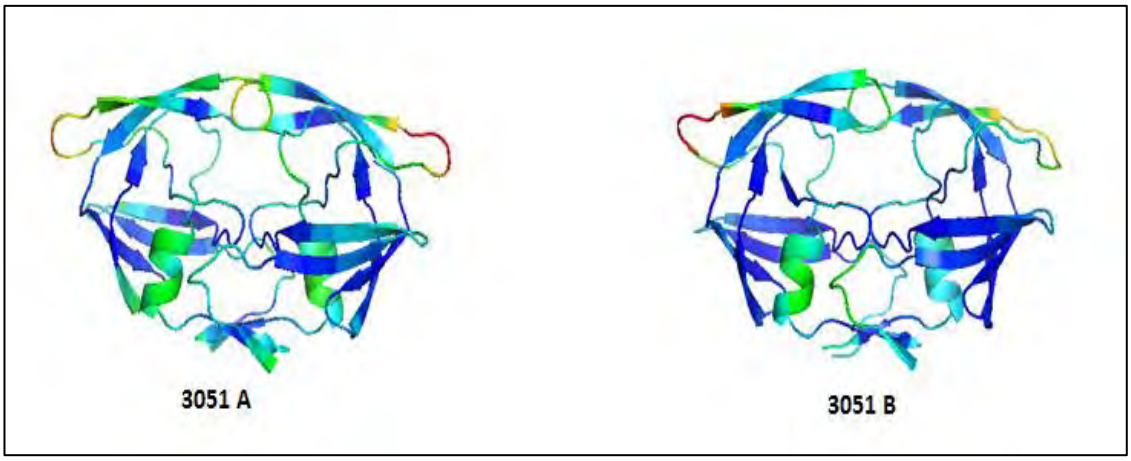


Figure 3.4: Homology models of sequence 3051. 3051A and 3051B show the models before and after loop refinement respectively.

3.3.5.3 301812

The structure with the lowest DOPE Z score of -1.39 was selected and MetaMQAPII returned the result as GDT_TS score of 88.76. The loop regions that were refined were those seen to be problematic in the results. The loops refined consisted of residues which included: 37-41, 50-51, 81-84,103-108,127-129,136-141 and 177-181. After refinement the GDT_TS scores decreased to 83.71 and the DOPE Z score improved to -1.51.

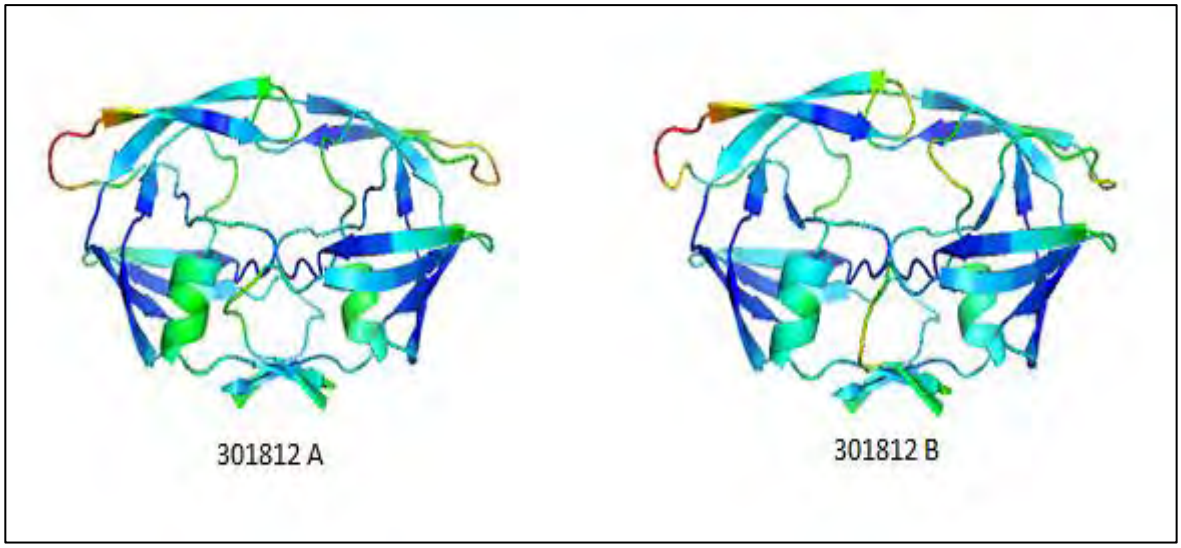


Figure 3.5: Homology models of sequence 301812. 301812A and 301812 B show the models before and after loop refinement respectively.

3.3.5.4 305152

The model with the lowest DOPE Z score of -1.21 returned a result from MetaMQAPII with a GDT_TS score of 88.76. The loop regions that were refined were those seen to be problematic in the results from MetaMQAPII. The loops refined consisted of residues which included: 37-41, 50-51, 81-84,103-108,127-129,136-141 and 177-181. Results obtained from refining show the GDT_TS scores remained the same at 88.76 and the DOPE Z score was lower at -1.39.

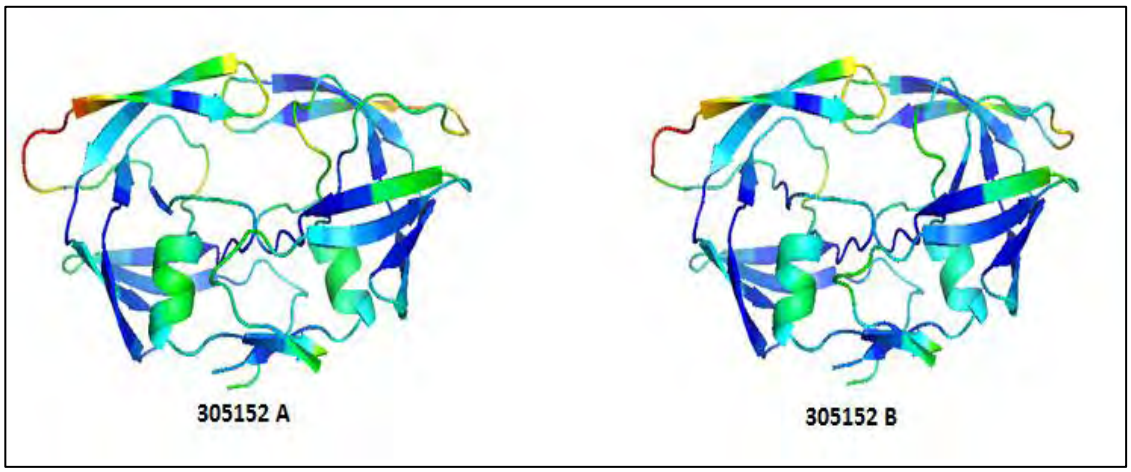


Figure 3.6: Homology models of sequence 305152. 305152A and 305152 B show the models before and after loop refinement respectively. The problematic regions are highlighted in red.

3.3.5.5 Subtype B

The model with lowest DOPE Z score of -1.24 and returned a MetaMQAPII result with a GDT_TS score of 93.31. The loop regions that were refined were those seen to be problematic in the results from MetaMQAPII. The loops refined consisted of residues which included: 37-41, 50-51, 81-84,103-108,127-129,136-141 and 177-181. The results from the loop refinement showed a decreased GDT_TS scores to 91.41 but had a lower DOPE Z score of -1.38.

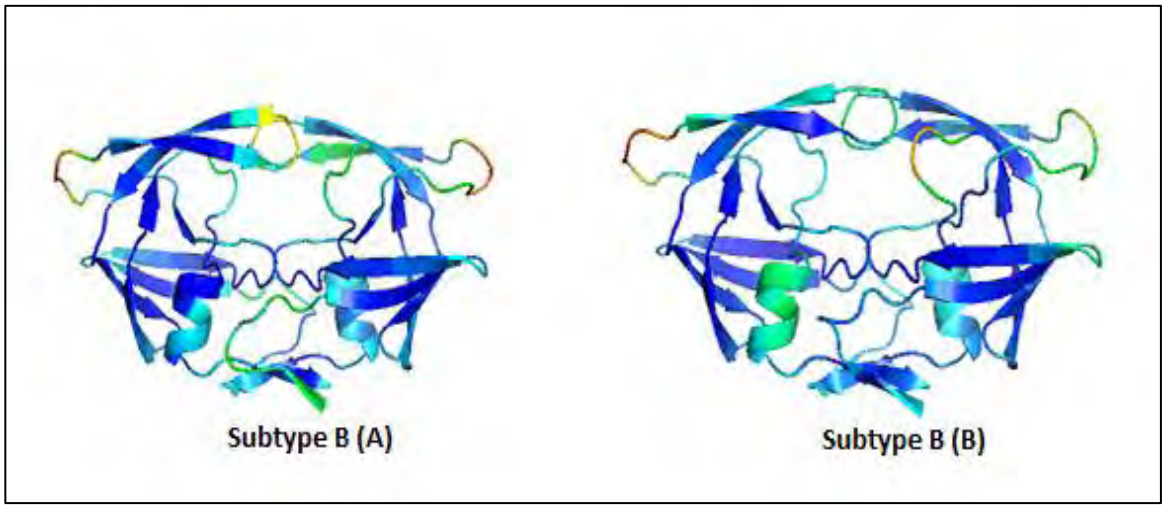


Figure 3.7: Homology models of sequence for Subtype B global consensus. Subtype B (A) and Subtype B (B) show the models before and after loop refinement respectively.

3.3.5.6 Subtype C

The structure selected had DOPE Z score of -1.18 and MetaMQAPII returned a result with a GDT_TS score of 91.67. The loop regions refined were those seen to be problematic from MetaMQAPII results. The residues which were refined were on the loops which included: 37-41, 50-51, 81-84, 103-108, 127-129, 136-141 and 177-181. The results from the loop refinement had a GDT_TS score which increased to 92.05 and had an improved DOPE Z score of -1.30.

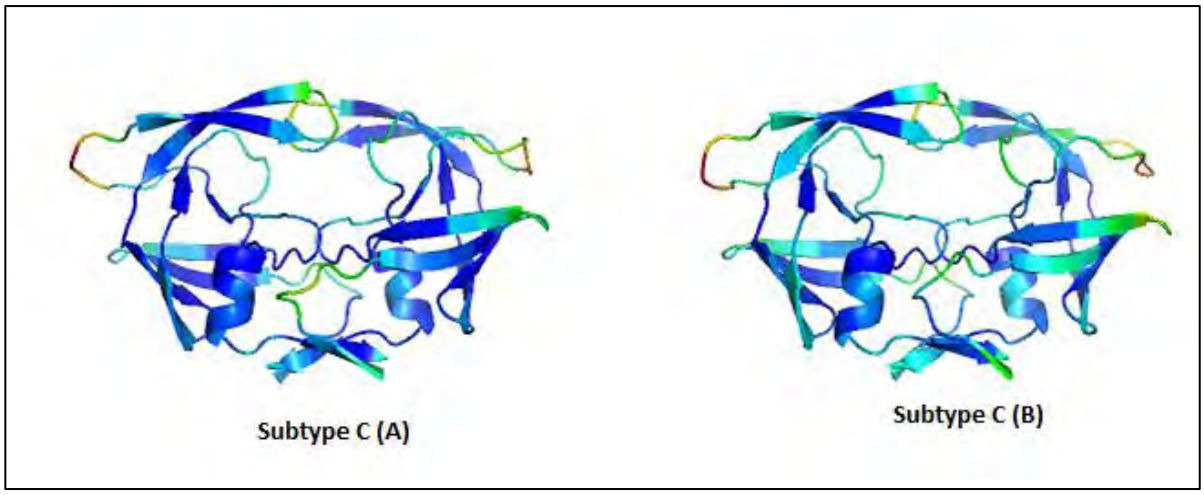


Figure 3.8: Homology models of sequence for Subtype C global consensus. Subtype C (A) and Subtype C (B) show the models before and after loop refinement respectively.

3.3.5.7 Comparison of DOPE Z and GDT_TS scores for Homology models built from homodimer.py and model_m2.py

Table 3.5 shows the DOPE Z scores for all homology models built from the scripts homodimer.py and model_m2.py. The values for b) are for homology models after loop refinement. The values for a) and b) have been described earlier in section 3.3.5.1 -3.3.5.6 and the models were built using the homodimer.py modeller script. The values in (c) were for the homology models built using the modeller script model-m2.py which has no extra symmetry restraints. The models which were built with homodimer.py had higher GDT_TS scores for the models 3018, 3051, 301812, 305152, consensus C and consensus B.

Table 3.5: DOPE Z Scores & GDT_ TS scores for homology models 3018, 3051, 301812, 305152, consensus B &C. (a) Models built using homodimer.py script (b) Models in (a) after loop refinement (c) Models built using model-m2.py

ID	DOPE Z scores	GDT_TS Scores
a) 3018	-1.49	90.40
b) 3018	-1.32	87.88
c) 3018	-1.02	83.84
a) 3051	-1.14	86.74
b) 3051	-1.31	86.99
c) 3051	-1.21	83.84
a) 301812	-1.39	88.76
b) 301812	-1.54	83.71
c) 301812	-1.33	79.79
a) 305152	-1.21	88.76
b) 305152	-1.39	88.76
c) 305152	-1.18	77.27
a) Consensus B	-1.24	93.31
b) Consensus B	-1.38	91.41
c) Consensus B	-1.23	92.93
a) Consensus C	-1.18	91.67
b) Consensus C	-1.30	92.05
c) Consensus C	-1.13	84.85

3.3.5.8 Comparison of models built using homodimer.py and model_m2.py Figure 3.9 & 3.10

In this section the spectrum from red to blue indicates the problematic regions with red for erroneous regions and blue for correct regions. The general trend for the models that were built with homodimer.py in comparison to model_m2.py is the models had improved visually in terms of the spectrum from green to blue. The loop regions are problematic for residues in the range of 40 – 44 for chain A and 140 -144 for chain B.

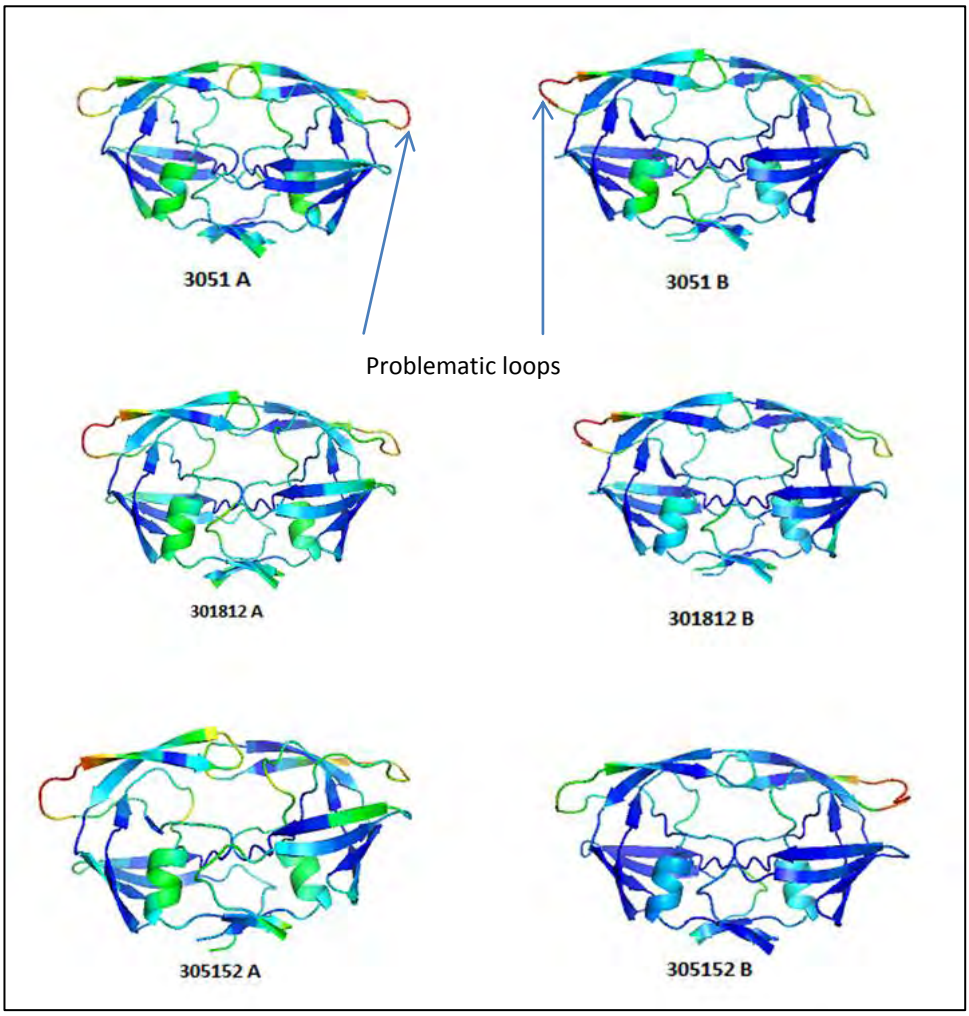


Figure 3.9: Comparison of Models 3051, 301812 and 305152 built using homodimer.py script (A) and model_m2.py (B). The model quality improves from A to B as shown with the spectrum from red (erroneous regions) and blue (correctly modeled regions). As indicated the loop regions were the incorrectly modeled.

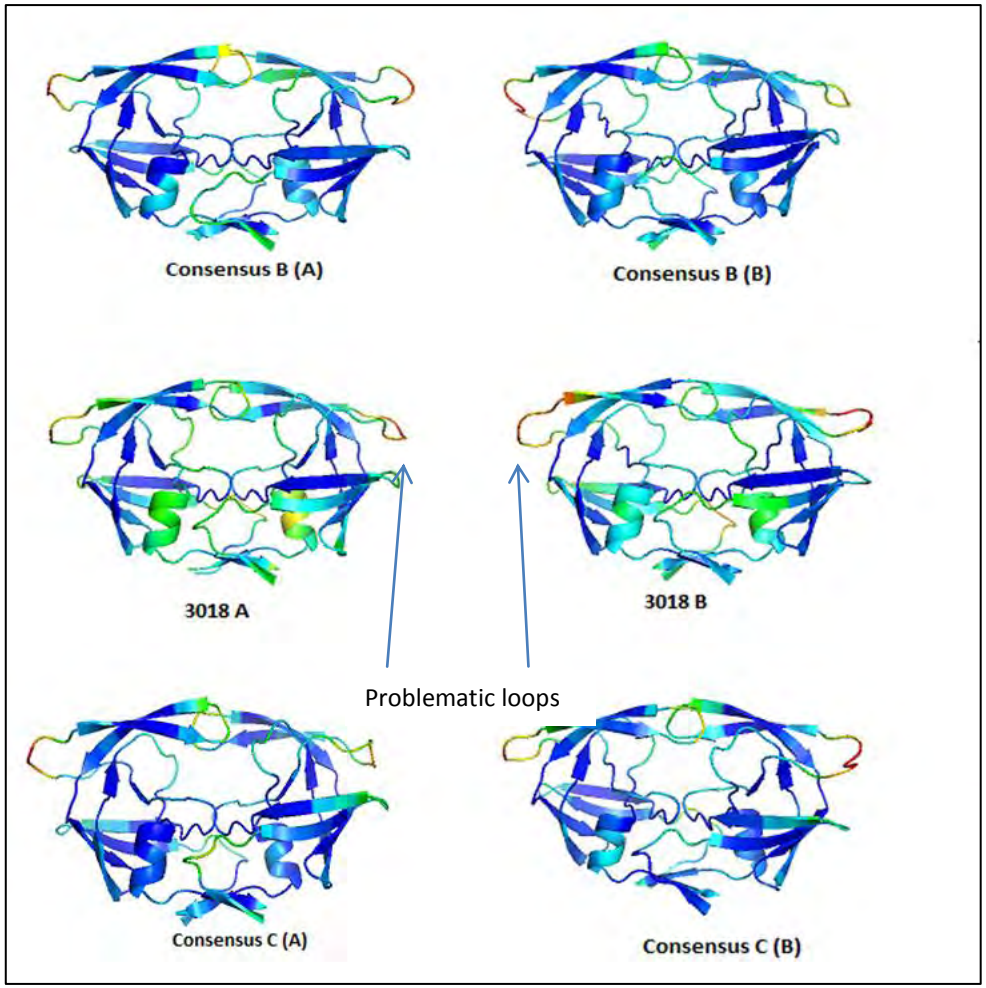


Figure 3.10 Comparison of Models 3018, Consensus B and Consensus C built using homodimer.py script (A) and model_m2.py (B). The model quality improves from A to B as shown with the spectrum from red (erroneous regions) to blue (correctly modeled regions). As indicated the loop regions were the incorrectly modeled.

3.3.6 Ramachandran plots

Table 3.6: Results from PROCHECK: Ramachandran plots showing the percentages of residues in the most favorable regions, additionally allowed regions, generously allowed regions and disallowed regions. The areas shaded in grey show the models before loop refinement (Darker grey for models built with script homodimer.py and lighter grey model-m2.py) while white is after refinement.

PROCHECK	3018	3051	301812	305152	Consensus B	Consensus C
Most favorable regions	100	98.0	96.8	94.9	95.6	96.8
	94.3	93.6	92.3	94.9	90.5	93.7
	96.2	96.2	97.4	96.2	96.8	97.5
Additionally allowed regions	0.0	1.9	3.2	5.1	4.4	3.2
	4.4	5.8	7.1	5.1	9.5	6.3
	3.8	3.8	2.6	3.8	3.8	2.5
Generously allowed regions	0.0	0.0	0.0	0.0	0.0	0.0
	0.0	0.0	0.0	0.0	0.0	0.0
	0.0	0.0	0.0	0.0	0.0	0.0
Disallowed regions	0.0	0.0	0.0	0.0	0.0	0.0
	1.3	0.6	0.6	0.0	0.0	0.0
	0.0	0.0	0.0	0.0	0.0	0.0
	Total					100

The results from PROCHECK showed that all the models passed as they had more than 90% of their residues in the most favourable regions. However loop refinement led to residues being shifted into the disallowed regions for sequences 3018, 3051 and 301812. The Ramachandran plots show residues in most sterically favored regions as red, additionally allowed regions as dark yellow, generously allowed regions as light yellow and disallowed regions as white.

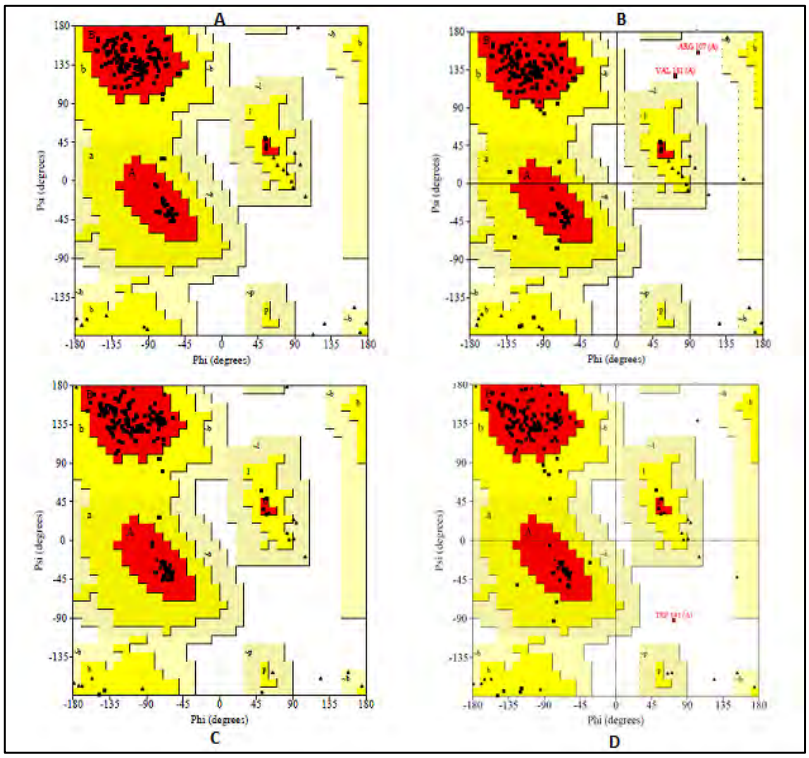


Figure 3.11: Ramachandran plots for models for 3018 and 3051 sequences. A & C are the models before loop refinement while B and D are after refinement. 3018 represented by A & B while 3051 is C & D.

In figure 3.11 B shows the Ramachandran plot for the model 3018 after refinement. It has two residues arginine and valine in its disallowed regions while the other models had all their residues in most favorable, additionally allowed regions or generously allowed regions.

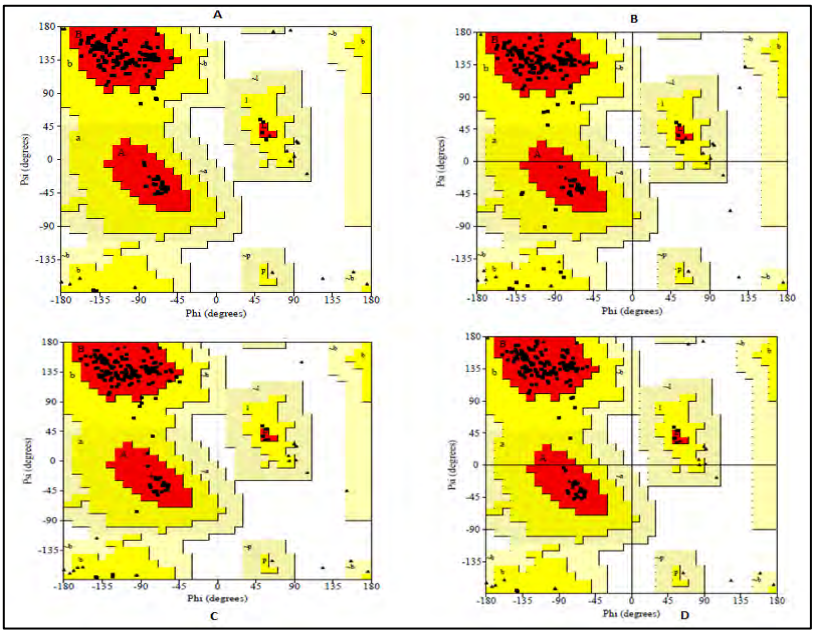


Figure 3.12: Ramachandran plots for models for subtype B and C consensus sequences. A & C are before loop refinement while B and D are after refinement. Subtype B is represented by A & B while subtype C is C & D

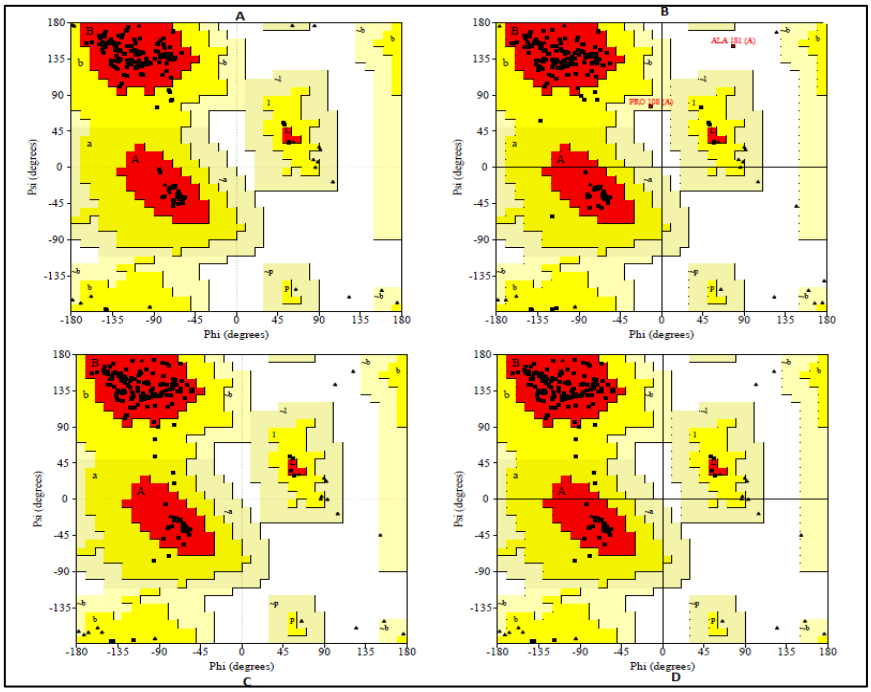


Figure 3.13: Ramachandran plots for models for 301812 and 305152 sequences. A & C are the models before loop refinement while B and D are after refinement. 301812 represented by A & B while 305152 is C & D. 301812 B after loop refinement shows a plot with residues in stearically unfavorable positions.

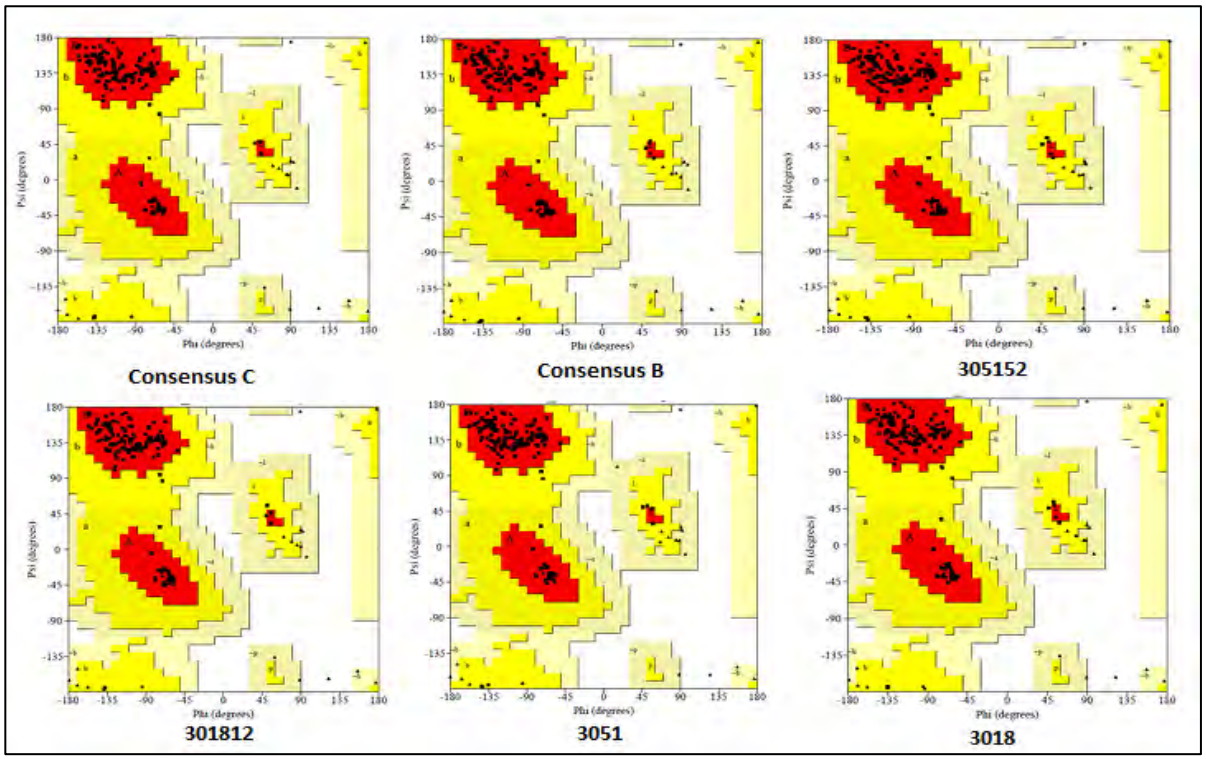


Figure 3.14: Ramachandran plots for models for 3018, 3051, 301812, 305152, consensus B and C sequences. The models shown have their residues in sterically favorable positions. These are models built with the script model_m2.py.

3.3.7 ProSA

The Z-scores for all the models built with modeller scripts homodimer.py and model_m2.py are within the range of those of native proteins from PDB as represented by their plots in Figure 3.15 - 3.23. The Z-scores were comparable to those of structures resolved by NMR. The plots showing knowledge based energy were below zero for the 40 residue window which is less sensitive than the 10 residue window. There two sequence positions residues 1-10 and residue 100-110 where there are peaks above the zero which represent positive energy indication errors.

Table 3.7: Z-Scores for Homology models built using script Homodimer.py and model_m2.py.

Model	Z-score before refinement	Z-score after refinement	Z-scores of Models from model_m2.py
3018	**-6.80**	**-6.50**	**-6.47**
3051	**-7.41**	**-7.37**	**-6.79**
301812	**-7.71**	**-7.26**	**-7.22**
305152	**- 7.12**	**-7.12**	**-6.5**
Consensus B	**- 6.82**	**-6.76**	**-6.3**
Consensus C	**-6.89**	**-6.89**	**-6.71**

3.3.7.1 Before loop refinement

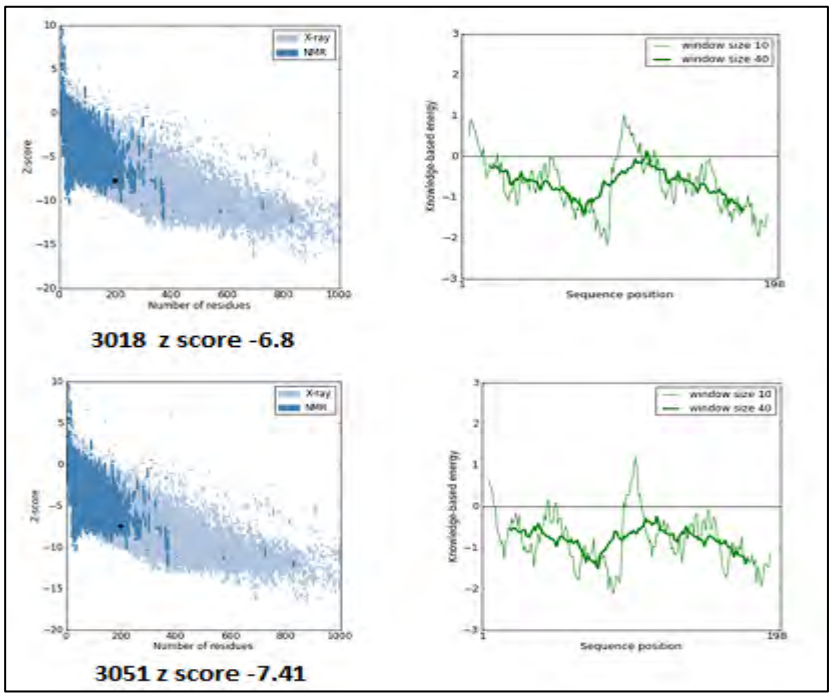


Figure 3.15: ProSA results for 3018 & 3051 showing plots for Z score vs. number of residues and knowledge based energy vs. sequence position.

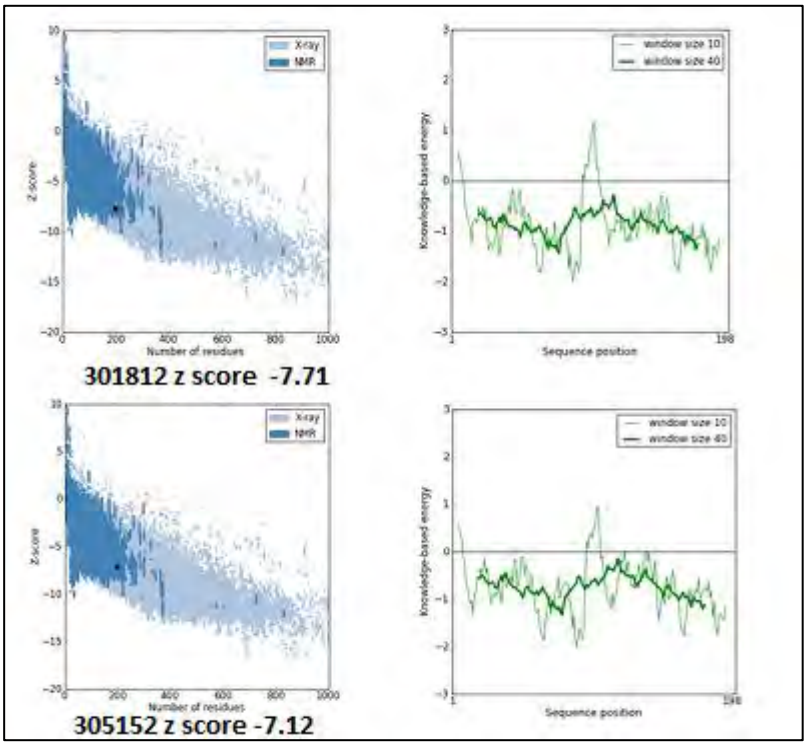


Figure 3.16: ProSA results for 301812 & 305152 showing plots for Z score vs. number of residues and knowledge based energy vs. sequence position.

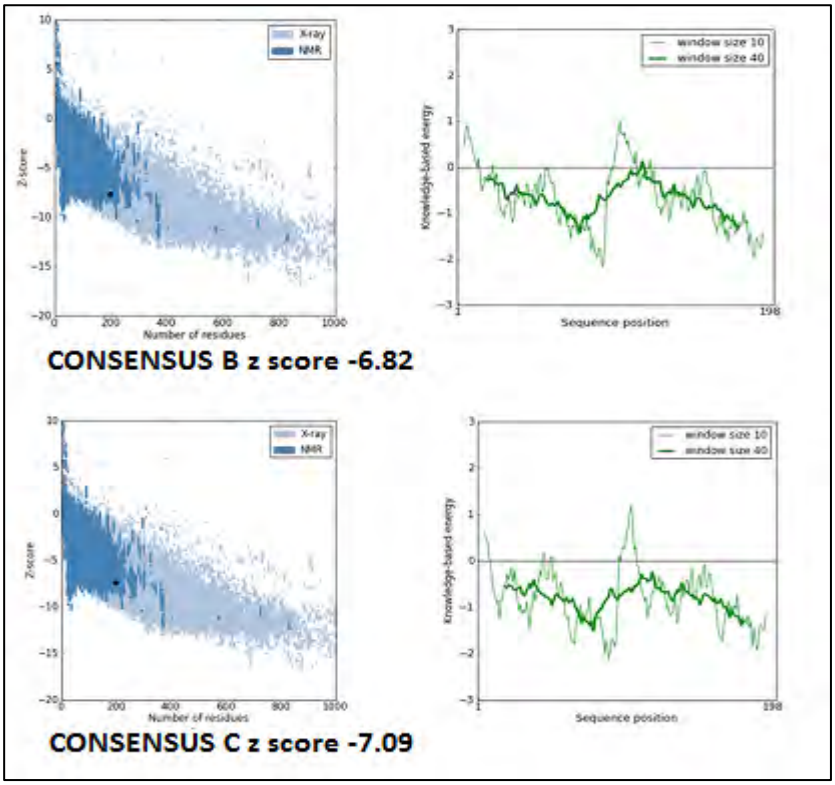


Figure 3.17: ProSA results for Consensus B & C showing plots for Z score vs. number of residues and knowledge based energy vs. sequence position.

3.3.7.2 After loop refinement

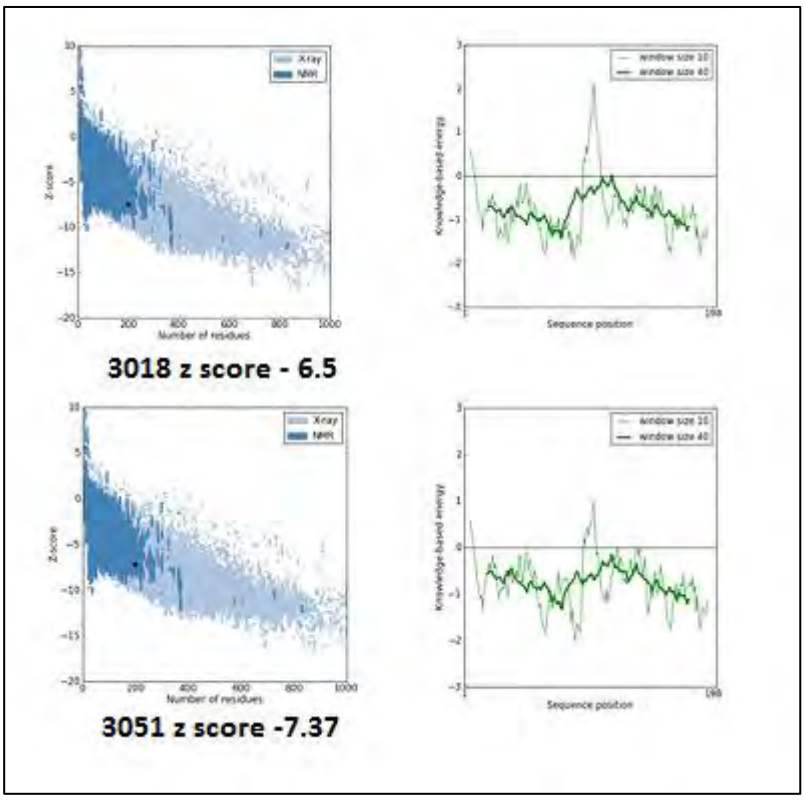


Figure 3.18: ProSA results for 3018 & 3051 showing plots for Z score vs. number of residues and knowledge based energy vs. sequence position.

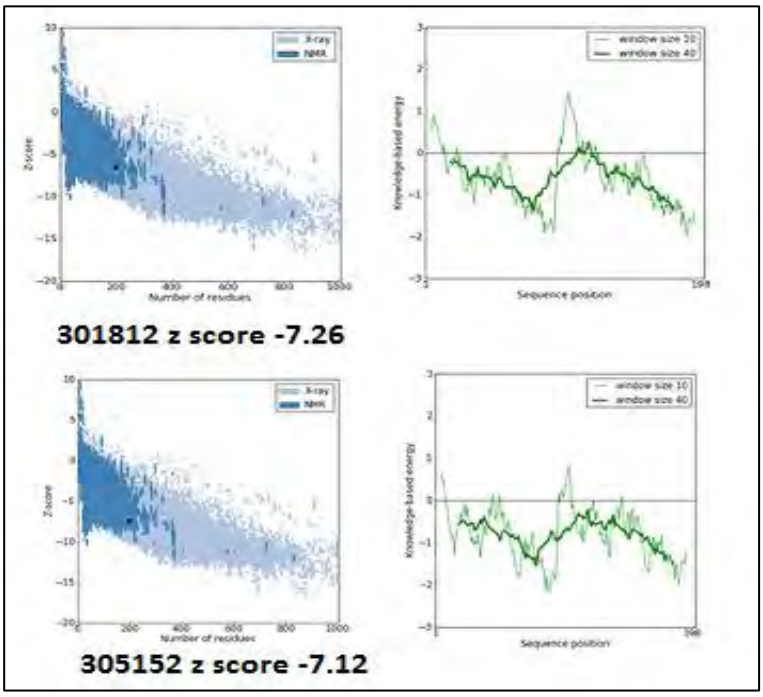


Figure 3.19: ProSA results for 301812 & 305152 showing plots for Z score vs. number of residues and knowledge based energy vs. sequence position.

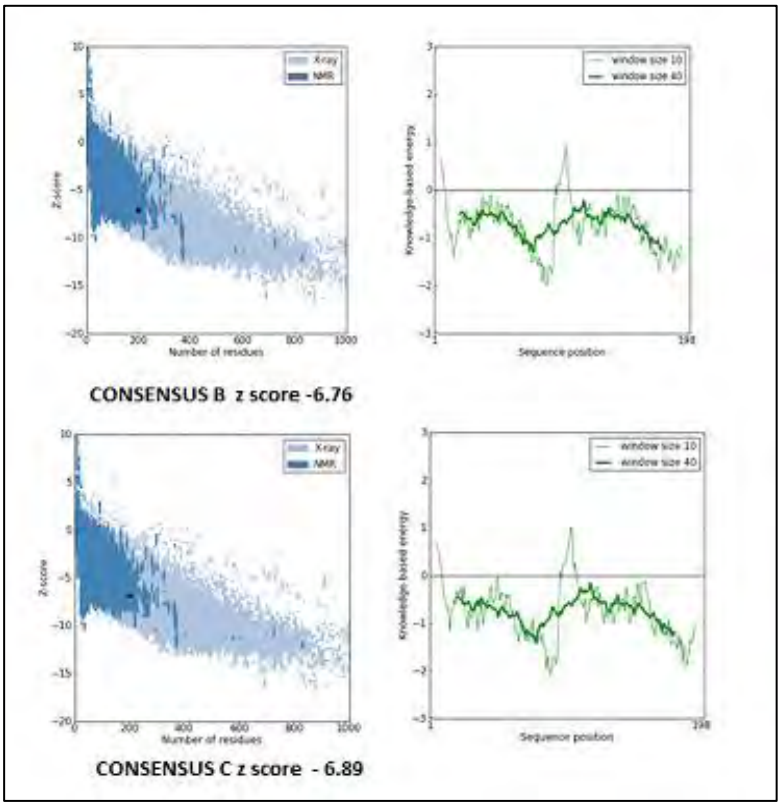


Figure 3.20: ProSA results for consensus B & C showing plots for Z score vs. number of residues and knowledge based energy vs. sequence position.

3.3.7.3 Models built with model_m2.py

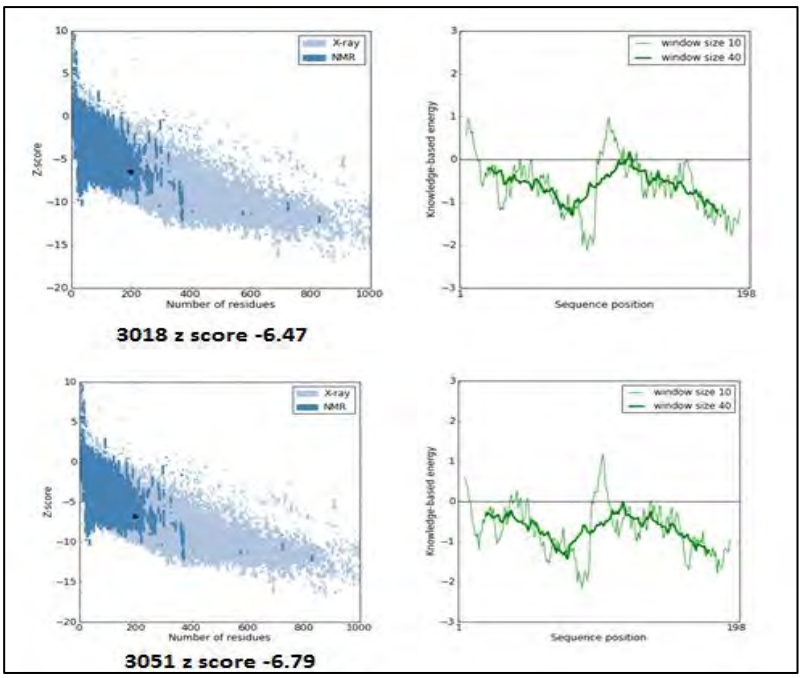


Figure 3.21: ProSA results for 3018 & 3051 showing plots for Z score vs. number of residues and knowledge based energy vs. sequence position

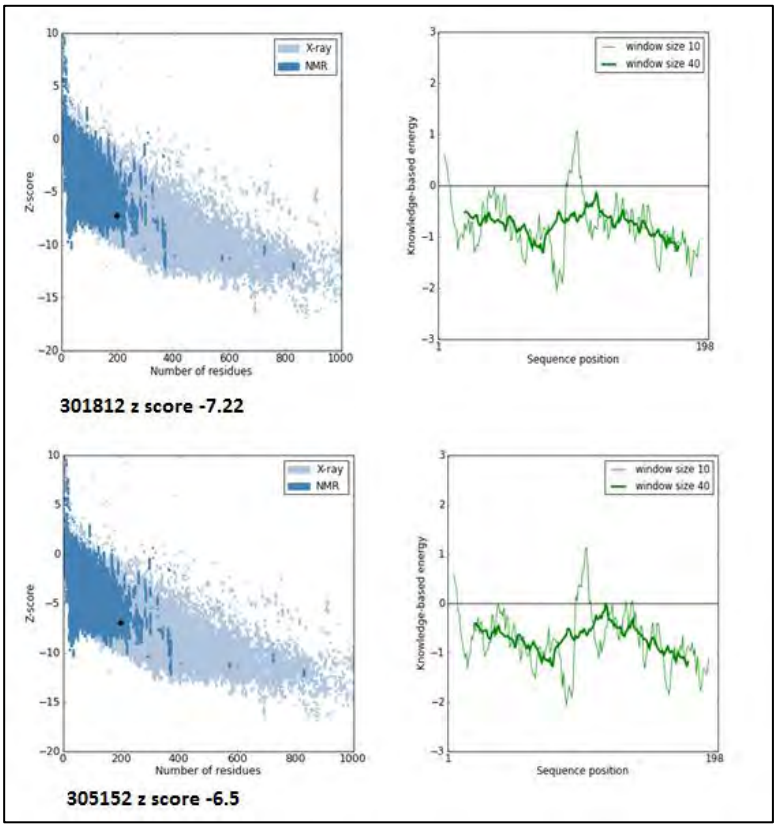


Figure 3.22: ProSA results for 301812 & 305152 showing plots for Z score vs. number of residues and knowledge based energy vs. sequence position.

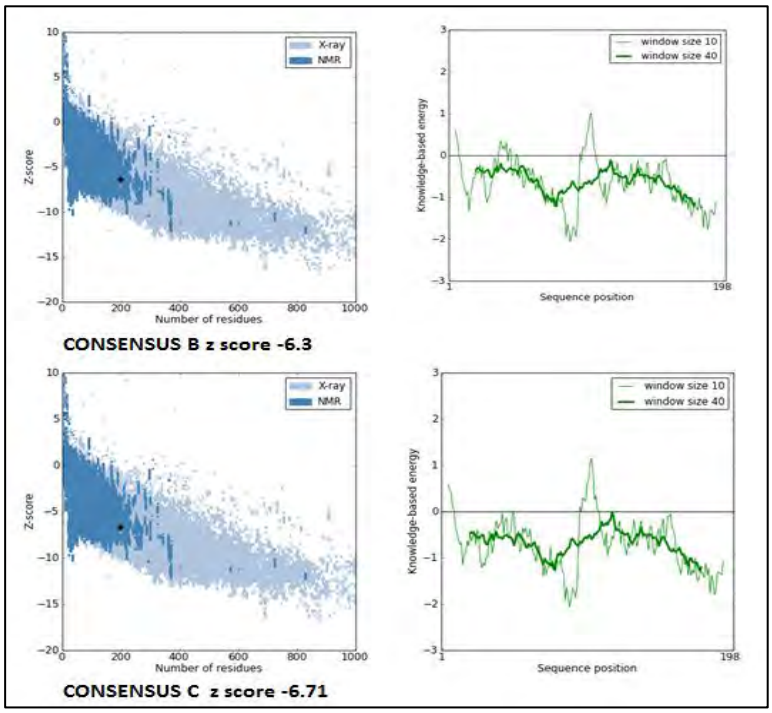


Figure 3.23: ProSA results for consensus B & C showing plots for Z score vs. number of residues and knowledge based energy vs. sequence position.

3.4 DISCUSSION

Homology modeling as a technique of resolving protein structures has been gaining appeal in the scientific world and has led to great developments in solving structures that have not been elucidated using other techniques. It is a fast and affordable bioinformatics technique as opposed to the traditional trial and error methods of drug discovery. The sequences that were chosen for purposes of homology modeling were chosen on the basis of drug resistance. The sequences 301812 and 305152 had the major drug resistant mutations I54V, V82A and M46I is found in only the former sequence. M46I has been reported in literature however has not been conclusively investigated as to how it leads and contributes to drug resistance (Clemente *et al.*, 2004, Alcaro *et al.*, 2009). Another criterion to select 305152 was the fact that the patient was the longest exposed to treatment for a period of 52 weeks. The HIV-1 subtype B and C consensus sequences, were selected as they are considered to have no mutations while 3018 and 3051 had no drug resistant mutations.

The results of the multiple sequence alignment and the percentage identity matrix showed high sequence conservation for all the sequences and their candidate templates from HHpred, PDB search. There was also high physicochemical similarity between aligned residues. The template search results from HHpred and PDB search returned 2HS1 as the first result for five sequences 3018, 3051, 301812, 305152 and subtype B. Subtype C had 3KA2 as its first result for HHpred. 2HS1 (Figure 3.24) was selected as a suitable template as it had a very high resolution of 0.84 Å which made it very suitable for modeling. It also had a closed conformation which made it suitable for docking studies as this conformation in which inhibitors bind (Durdagi *et al.*, 2008). The percentage identity for 2HS1 against the six sequences to be modeled was also high. The lowest percentage identity is 79% which meets the cutoff suggested by literature of > 40% as the lower limit appropriate for building reliable models for structure based design and >50% for docking studies comparable to those of crystal structures (Nayeem *et al.*, 2006). Therefore from the preliminary data the template was suitable to proceed to the next step in the modeling process.

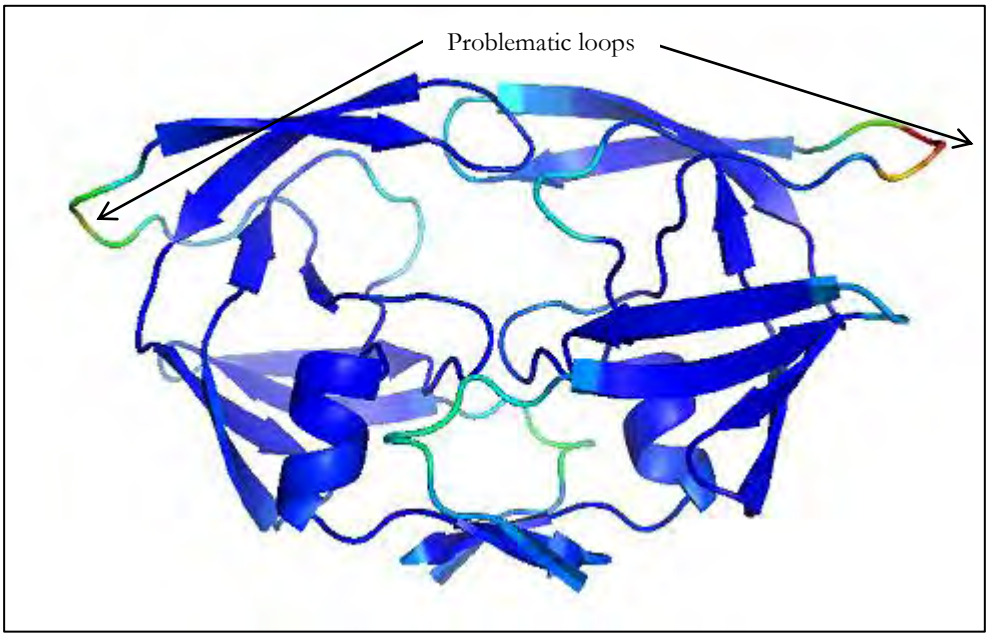


Figure 3.24: MetaMQAPII result showing 2HS1 template with a resolution of 0.84 Å and GDT_TS score of 92.17. The loop regions with problems are highlighted in the diagram.

Modeller package was used to build models and validation using MetaMQAPII, PROCHECK and PROSA. The MetaMQAPII results show very high GDT_TS scores ranging between 92.93 -77.27. The DOPE Z scores are also in range of -1.02 - -1.49. The GDT_TS scores assess the overall model quality and it is therefore important to look at PDB files provided as part of the MetaMQAPII results so as to determine the regions that may have been inaccurately modeled. A modeller script (homodimer.py) was used to build one a set of 500 structures for each of the six sequences selected 3018, 3051, 301812, 305152 and subtype B and C. The script is designed for multichain models and introduces extra symmetry restraints to constrain chain A and B to have the same conformation. The modeller script (model_m2.py) was used to build one a set of 100 structures for each of the six sequences selected 3018, 3051, 301812, 305152 and subtype B and C but had no extra symmetry restraints.

3.4.1 Evaluation of models with METAMQAPII built using homodimer.py before and after loop refinement.

The models which were evaluated for the HIV-1 subtype B, C, 3018, 3051, 301812 and 305152 after visualization of the PDB files provided by MetaMQAPII had some problematic regions. These regions were mainly in the loops away from the active site

residues. The model 305152 had a residue alanine 82 in both chains A and B which is in a loop in the active site which was highlighted as being incorrectly built by Modeller.

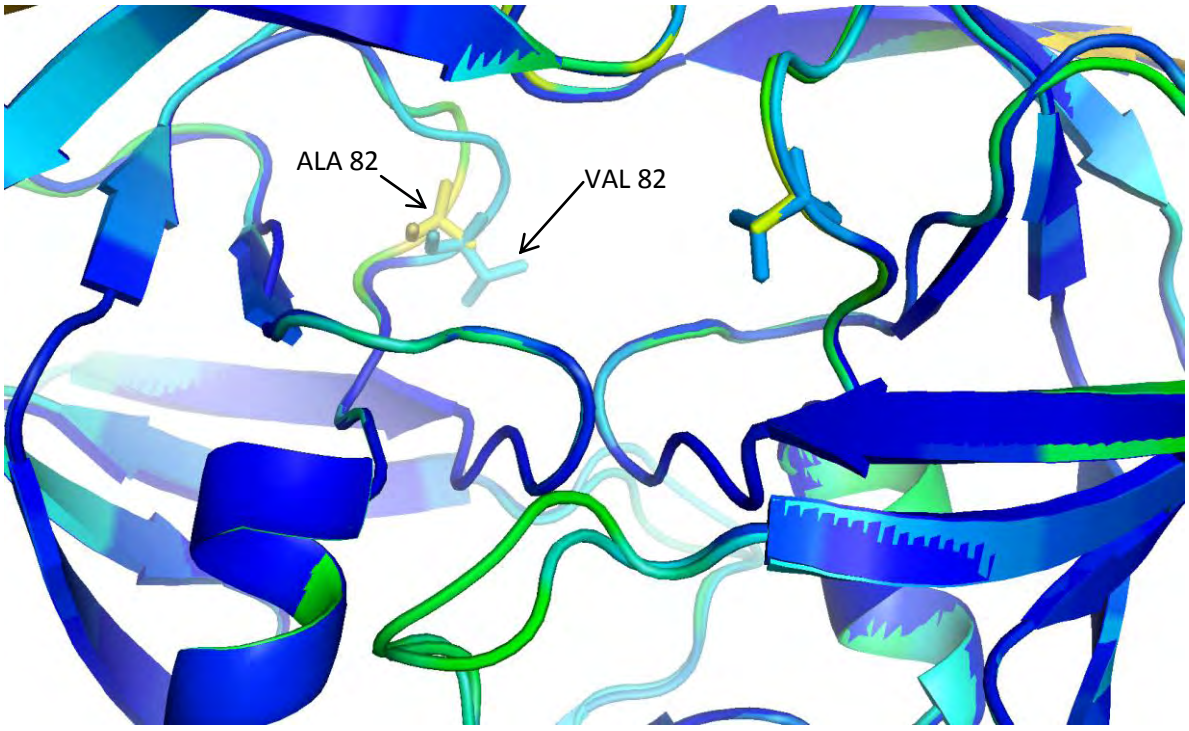


Figure 3.25: Model for sequence 305152 superimposed on the template 2HS1. Residue at position 82 is highlighted to be incorrectly modeled by MetaMQAPII.

Figure 3.25 shows model 305152 with residue 82 viewed at residue level. The residue differences are a result of conformational isomerism and hence rotamers. This is further confirmed when a single chain of the model is submitted to MetaMQAPII; the problematic regions found are the loops in Figure 3.26.

Four of the six models after loop refinement had their GDT_TS score decrease with the exception of 3051 and 305152. The MetaMQAPII PDB files after loop refinement showed the models to have new problematic areas introduced which was corroborated by the decreased GDT_TS scores that showed decreased model quality. Modeller builds loops by conformation search; as such the loop with the lowest energy is built. The lowest energy conformation may be built but is not necessarily the correct loop.

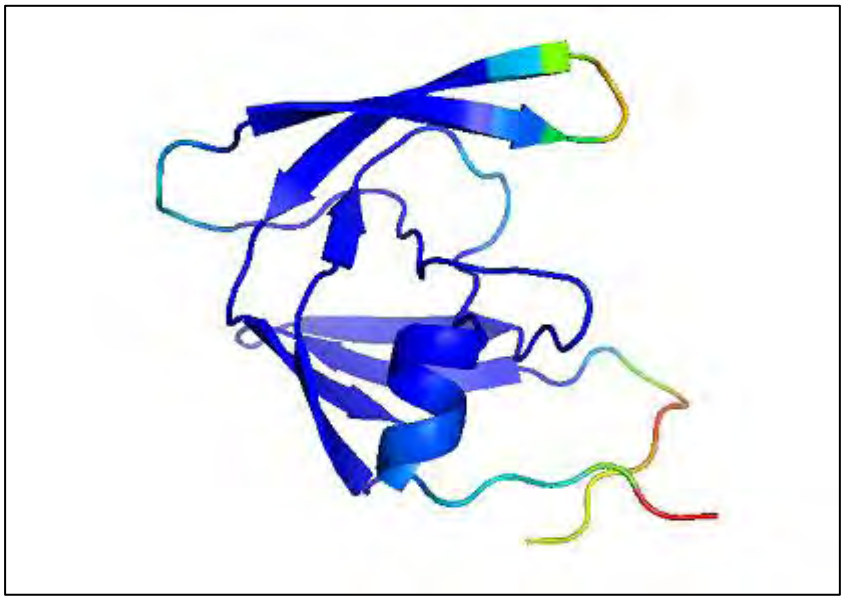


Figure 3.26: Chain A of model 305152 showing problematic regions as loops. Result from METAMQAPII

3.4.2 Comparison of MetaMQAPII results of models built with homodimer.py and model_m2.py

The models built using script model_m2.py had lower GDT_TS score in comparison with those of script homodimer.py. This indicates that in addition to looking at the GDT_TS scores it important to look at the PDB files provided from MetaMQAPII so as to localize the areas that are modeled poorly and determine whether loop refinement should be done. The models built using model_m2.py improved in terms of the spectrum from red to blue in comparison to those from homodimer.py and are more suitable for docking studies.

3.4.3 Evaluation using PROCHECK and ProSA for homodimer.py and model_m2.py

The Ramachandran plots before and after loop refinement showed that the models passed the quality test as the all had 90% of their residues in the most favorable regions. This notwithstanding after loop refinement there were some residues that were in sterically unfavorable positions and this therefore lowered the quality of the models 3018, 3051 and 301812. The models built using script model_m2.py had their residues in the most favorable region and additionally allowed regions and were over the 90% cutoff hence they are stereochemically 'fit'.

The ProSA results show the Z scores were within the required range for native structures in the PDB and hence all models evaluated using MetaMQAPII passed this quality test. They were mainly found in the NMR region.

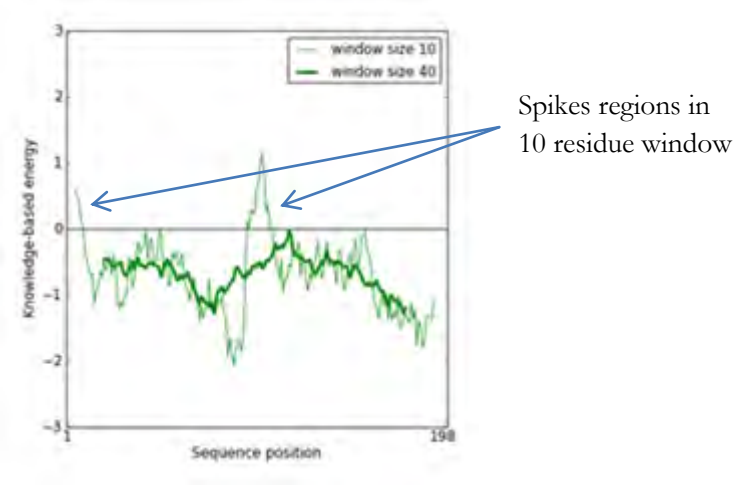


Figure 3.27: ProSA Knowledge based energy plot showing incorrectly modeled regions.

The plots which represent knowledge based energy (an example in Figure 3.27) showed that for all models using homodimer.py and model_m2 had their energy in the negative range for the 40 residue window. This is the range which is preferred as models with incorrect regions are those which have residues with positive values. In the case of the 10 residue window large fluctuations were observed for residues at the N-terminal for residues 1-10 and also at residue 100-110 in all graphs. These are similar regions in both graphs as these are the N-terminal regions for chain A and B. Residues in this regions form loops and are not in the active site hence do not affect docking. A reason for the fluctuation is that the 10 residue window shows a plot averaged for 10 residue fragment and is more sensitive than the 40 residue window where the averaging is for a 40 residue fragment.

3.4.4 Analysis of mutations M46I, I54V and V82A

The mutation V82A is found in the active site while M46I and I54V are found just near the active site which consists of residues 25-32,47-53 and 80-84 (Ali *et al.*, 2010). M46I and I54V are found at the flap region. The residue substitutions are chemically conservative with M46I, I54V and V82A being hydrophobic in nature. Valine for the mutation V82A has a longer side chain than alanine and thus the interaction is affected between it and the inhibitors. The location of the mutations M46I and I54V is indicative that they affect the

geometry of the active site as they occur just at the beginning and end of the active site respectively. Inhibitors are designed to fit the active site using a lock and key paradigm and any change in residues which affect the how the inhibitor 'sits' in the active site (Ohtaka and Freire, 2005).

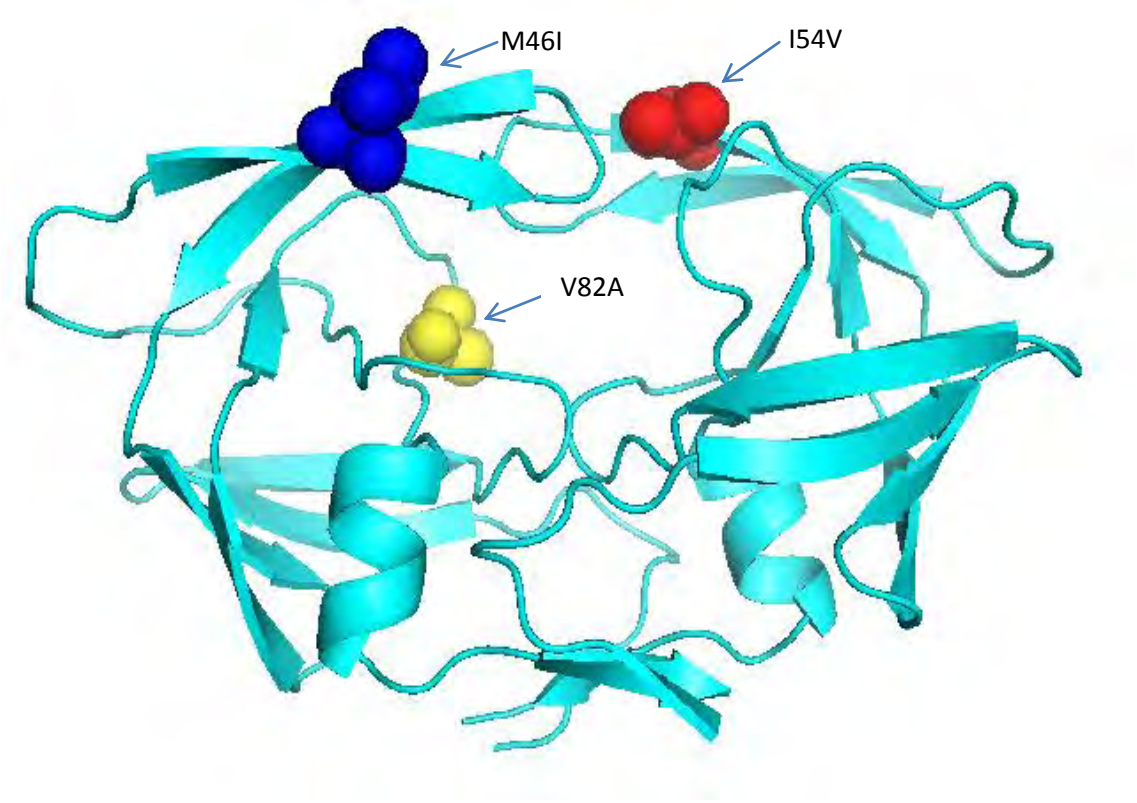


Figure 3.28: M46I, I54V located in the flap region and V82A mutation in the active site from homology model of sequence 305152.

M46I and I54V occur in the flap region which is flexible and opens to allow substrates into the active site. Molecular dynamics simulations which allow for the movement of the flaps would give more information as to how the mutations affect the inhibitor binding.

3.5 CONCLUSION AND FUTURE WORK

All organisms adapt through evolutionary mechanisms in an effort to survive the environment they live in and HIV is no exception to this biological phenomenon (Iweriebor *et al.*, 2011). HIV evolves at a rapid rate due to its error prone reverse transcriptase enzyme and also in an effort to evade the drugs. This leads to HIV mutant proteases which are more selected during the phase where there is a drug challenge. This has therefore led to widespread resistance hence the need to search for newer inhibitors from existing ones that escape resistance mechanisms or design adaptive inhibitors (Kuritzkes, 2011). Since the protease inhibitor targets the protease active site, any change in the shape affects how these inhibitors fit. Designing drugs that adapt to this change in the active site for the drug resistant protease will evade resistance (Ohtaka and Freire, 2005).

By use of Stanford HIV database tool, the mutations in the twenty nine HIV-1 subtype C sequences were identified and compared to those from the Mutation.py tool. The Stanford mutation database tool in addition to identifying the mutations goes a step further and classifies them as major and minor mutations. This information is useful in isolating which mutations are likely to cause high level resistance. A comparison of the two tools shows that the Stanford mutation database is superior to our tool however as subtype C is the most widely spread it is important to identify the resistance patterns in the protease and map them to the active site so as to design specific inhibitors.

The phylogenetic analysis identified the proteases as being related to subtype C but also showed that mutations accumulate rapidly as the sequences appeared to diverge from the global consensus sequence. A broader study of protein protease sequences would be able to identify and characterize the mutation patterns by clustering similar mutations together. These mutations could be mapped to know the commonly occurring resistance patterns and this would assist in drug design for inhibitors that are specific for subtype C in the era of rational structure based design of drugs.

Homology modeling involved building of models of six sequences: HIV-1 subtype B and C global consensus sequences, 3018, 3051, 301812 and 305152. The sequences 305152 and 301812 were chosen on basis of their major resistance mutations M46I, I54V, V82A while subtype B, C, 3018, 3051 had no mutations. Models built from homodimer.py had higher GDT_TS scores however after visualization of the PDB files the models made

with model_m2 seemed to have less poorly modeled regions as the spectrum intensity was blue as analyzed by MetaMQAPII. Due to the extra restraints introduced by homodimer.py the models most suitable for docking are those of the script model_m2.py as both had high GDT_TS scores. It is therefore important to look at both the GDT_TS score and PDB file provided by MetaMQAPII. MetaMQAPII can only view single chain models and more investigation needs to be done into how renumbering a multichain model affects the results. In this case when viewed as a single chain the results showed the models to be of good quality opposed to a renumbered multichain models built with for the homodimer.py.

The major mutations M46I and I54V were found to occur around the active site in the flap region whereas the V82A is located in the active site. Only V82A interacts with inhibitors and the mechanisms with which the mutations cause resistance need to be further investigated with molecular dynamics simulations in order to see how they affect the inhibitor interaction with the active site. It would also be important to carry out docking studies with inhibitors so as to investigate the effect of the mutations in terms of decreasing the binding constants and also in the orientation of binding at the active site.

REFERENCES

Al-Lazikani, B., Jung, J., Xiang, Z. & Honig, B. 2001. Protein Structure Prediction. *Current Opinion In Chemical Biology,* 5**,** 51-56.

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. *Journal Of Chemical Information And Modeling,* 49**,** 1751-1761.

Aleman, S., Soderbarg, K., Visco-Comandini, U., Sitbon, G. & Sonnerborg, A. 2002. Drug Resistance At Low Viraemia In HIV-1-Infected Patients With Antiretroviral Combination Therapy. *AIDS,* 16**,** 1039-1044.

Ali, A., Bandaranayake, R. M., Cai, Y., King, N. M., Kolli, M., Mittal, S., Murzycki, J. F., Nalam, M. N. L., Nalivaika, E. A., Özen, A., Prabu-Jeyabalan, M. M., Thayer, K. & Schiffer, C. A. 2010. Molecular Basis For Drug Resistance In HIV-1 Protease. *Viruses,* 2**,** 2509-2535.

Altschul, S. 1999. Hot Papers - Bioinformatics - Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs By S.F. Altschul, T.L. Madden, A.A. Schaffer, J.H. Zhang, Z. Zhang, W. Miller, D.J. Lipman - Comments. *Scientist,* 13**,** 15-15.

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. 1998. Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Faseb Journal,* 12.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. 1990. Basic Local Alignment Search Tool. *Journal Of Molecular Biology,* 215**,** 403-410.

Araya, T., Tensou, B., Davey, G. & Berhane, Y. 2011. Burial surveillance detected significant reduction in hiv-related deaths in addis ababa, ethiopia. *Tropical medicine and International health,* 16**,** 1483-1489.

Armougom, F., Moretti, S., Poirot, O., Audic, S., Dumas, P., Schaeli, B., Keduas, V. & Notredame, C. 2006. Expresso: Automatic Incorporation Of Structural Information In Multiple Sequence Alignments Using 3D-Coffee. *Nucleic Acids Research,* 34**,** W604-W608.

Arora, S. K., Gupta, S., Toor, J. S. & Singla, A. 2008. Drug Resistance-Associated Genotypic Alterations In The *Pol* Gene Of HIV Type 1 Isolates In ART-Naive Individuals in North India. *AIDS Research And Human Retroviruses,* 24**,** 125-130.

Badley, A. D., Pilon, A. A., Landay, A. & Lynch, D. H. 2000. Mechanisms of HIV-associated lymphocyte apoptosis. *Blood,* 96**,** 2951-2964.

Baxevanis, A. & Ouellette, F. 2001. *Bioinformatics: a practical guide to the analysis of genes and proteins, second edition*, wiley-interscience.

Bessong, P. O. 2008a. Polymorphisms in HIV-1 subtype c proteases and the potential impact on protease inhibitors. *Tropical medicine and international health,* 13**,** 144-151.

Bessong, P. O. 2008b. Polymorphisms in HIV-1 subtype c proteases and the potential impact on protease inhibitors. *Trop med int health,* 13**,** 144-51.

Biegert, A. & Söding, J. 2008. De Novo Identification Of Highly Diverged Protein Repeats By Probabilistic Consistency. *Bioinformatics,* 24**,** 807-814.

Bofill, M., Janossy, G., Lee, C. A., Macdonaldburns, D., Phillips, A. N., Sabin, C., Timms, A., Johnson, M. A. & Kernoff, P. B. A. 1992. Laboratory Control Values For CD4 and CD8 Lymphocytes-T - Implications For Hiv-1 Diagnosis. *Clinical And Experimental Immunology,* 88**,** 243-252.

Boniecki, M., Rotkiewicz, P., Skolnick, J. & Kolinski, A. 2003. Protein Fragment Reconstruction Using Various Modeling Techniques. *Journal Of Computer-Aided Molecular Design,* 17**,** 725-738.

Brooks, B. R., Brooks, C. L., Mackerell, A. D., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A. R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R. W., Post, C. B., Pu, J. Z., Schaefer, M., Tidor, B., Venable, R. M., Woodcock, H. L., Wu, X., Yang, W., York, D. M. & Karplus, M. 2009. Charmm: The Biomolecular Simulation Program. *Journal Of Computational Chemistry,* 30**,** 1545-1614.

Buckheit, K. W., Yang, L. & Buckheit Jr, R. W. 2011. Development Of Dual-Acting Pyrimidinediones As Novel And Highly Potent Topical Anti-HIV Microbicides. *Antimicrob Agents Chemother,* 55**,** 5243-5254.

Castro, H. C., Abreu, P. A., Geraldo, R. B., Martins, R. C., Dos Santos, R., Loureiro, N. I., Cabral, L. M. & Rodrigues, C. R. 2011. Looking At The Proteases from A Simple Perspective. *J Mol Recognit,* 24**,** 165-81.

Chakravarty, S., Ghersi, D. & Sanchez, R. 2011. Systematic Assessment Of Accuracy of Comparative Model of Proteins Belonging To Different Structural Fold Classes. *Journal Of Molecular Modeling,* 17**,** 2831-2837.

Chang, M. W. & Torbett, B. E. 2011. Accessory Mutations Maintain Stability In Drug-Resistant HIV-1 Protease. *Journal Of Molecular Biology,* 410**,** 756-760.

Clemente, J. C., Hemrajani, R., Blum, L. E., Goodenow, M. M. & Dunn, B. M. 2003. Secondary Mutations M36I AND A71V In The Human Immunodeficiency Virus Type 1 Protease Can Provide An Advantage For The Emergence Of The Primary Mutation D30n†. *Biochemistry,* 42**,** 15029-15035.

Clemente, J. C., Moose, R. E., Hemrajani, R., Whitford, L. R. S., Govindasamy, L., Reutzel, R., Mckenna, R., Agbandje-Mckenna, M., Goodenow, M. M. & Dunn, B. M. 2004. Comparing The Accumulation of Active- And Nonactive-Site Mutations In The HIV-1 Protease. *Biochemistry,* 43**,** 12141-12151.

Coman, R. M., Robbins, A. H., Fernandez, M. A., Gilliland, C. T., Sochet, A. A., Goodenow, M. M., Mckenna, R. & Dunn, B. M. 2008. The Contribution Of Naturally Occurring Polymorphisms In Altering The Biochemical And Structural Characteristics Of HIV-1 Subtype C Protease. *Biochemistry,* 47**,** 731-743.

Cristobal, S., Zemla, A., Fischer, D., Rychlewski, L. & Elofsson, A. 2001. A Study Of Quality Measures for Protein Threading Models. *BMC Bioinformatics,* 2.

Curry, A., Turner, A. J. & Lucas, S. 1991. Opportunistic Protozoan Infections In Human-Immunodeficiency-Virus Disease - Review Highlighting Diagnostic and Therapeutic Aspects. *Journal Of Clinical Pathology,* 44**,** 182-193.

De Beer, T. A. P., Wells, G. A., Burger, P. B., Joubert, F., Marechal, E., Birkholtz, L. & Louw, A. I. 2009. Antimalarial Drug Discovery: In Silico Structural Biology And Rational Drug Design. *Infectious Disorders - Drug Targets,* 9**,** 304-318.

De Medeiros, R. M., Junqueira, D. M., Matte, M. C. C., Barcellos, N. T., Chies, J. A. B. & Matos Almeida, S. E. 2011. Co-Circulation HIV-1 Subtypes B, C, And CRF31-Bc In A Drug-Naïve Population From Southernmost Brazil: Analysis Of Primary Resistance Mutations. *Journal Of Medical Virology,* 83**,** 1682-1688.

De Mulder, M., Yebra, G., Martín, L., Prieto, L., Mellado, M. J., Rojo, P., Muñoz-Fernández, M. A., De Ory, S. J., Ramos, J. T., Holguín, Á., De Jose, M. I., Gonzalez-Tome, M. I., Gurbindo, M. D., Navarro, M. L., Saavedra-Lozano, J., Delgado, R., Martin-Fontelos, P., Guillen, S., Martinez, J., Roa, M. A., Beceiro, J., Navas, A., Gonzalez-Granados, I. & Blazquez, D. 2011. Drug Resistance Prevalence And Hiv-1 Variant Characterization In The Naive and Pretreated Hiv-1-Infected Paediatric Population In Madrid, Spain. *Journal Of Antimicrobial Chemotherapy,* 66**,** 2362-2371.

Deane, C. M. & Blundell, T. L. 2000. A Novel Exhaustive Search Algorithm for Predicting The Conformation Of Polypeptide Segments In Proteins. *Proteins: Structure, Function And Genetics,* 40**,** 135-144.

Deane, C. M. & Blundell, T. L. 2001. Coda: A Combined Algorithm For Predicting The Structurally Variable Regions of Protein Models. *Protein Science,* 10**,** 599-612.

Dhaliwal, B. & Chen, Y. W. 2009. Computational Resources For Protein Modelling And Drug Discovery Applications. *Infectious Disorders - Drug Targets,* 9**,** 557-562.

Di Luccio, E. & Koehl, P. 2011. A Quality Metric For Homology Modeling: The H-Factor. *BMC Bioinformatics,* 12.

Dierynck, I., De Meyer, S., Lathouwers, E., Abeele, C. V., Van De Casteele, T., Spinosa-Guzman, S., De Béthune, M. P. & Picchio, G. 2010. In Vitro Susceptibility And Virological Outcome To Darunavir And Lopinavir Are Independent of HIV Type-1 Subtype In Treatment-Naive Patients. *Antiviral Therapy,* 15**,** 1161-1169.

Doyon, L., Elston, R. & Bonneau, P. R. 2009. Resistance To HIV-1 Protease Inhibitors. 477-492.

Dragic, T., Litwin, V., Allaway, G. P., Martin, S. R., Huang, Y. X., Nagashima, K. A., Cayanan, C., Maddon, P. J., Koup, R. A., Moore, J. P. & Paxton, W. A. 1996. HIV-1 Entry Into CD4(+) Cells Is Mediated By The Chemokine Receptor Cc-Ckr-5. *Nature,* 381**,** 667-673.

Durdagi, S., Mavromoustakos, T., Chronakis, N. & Papadopoulos, M. G. 2008. Computational Design Of Novel Fullerene Analogues As Potential Hiv-1 Pr Inhibitors: Analysis Of The Binding Interactions Between Fullerene Inhibitors And HIV-1 PR Residues Using 3D QSAR, Molecular Docking and Molecular Dynamics Simulations. *Bioorganic And Medicinal Chemistry,* 16**,** 9957-9974.

Edgar, R. C. 2004. Muscle: A Multiple Sequence Alignment Method With Reduced Time And Space Complexity. *BMC Bioinformatics,* 5**,** 1-19.

Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M.-Y., Pieper, U. & Sali, A. 2002. Comparative Protein Structure Modeling Using Modeller. *Current Protocols In Bioinformatics*. John Wiley & Sons, Inc.

Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M. Y., Pieper, U. & Sali, A. 2007. Comparative Protein Structure Modeling Using Modeller. *Current Protocols In Protein Science / Editorial Board, John E. Coligan ... [Et Al.],* Chapter 2.

Felsenstein, J. 2005. *Phylip (Phylogeny Inference Package) Version 3.0.Distributed By The Author. Department Of Genome Sciences, University Of Washington, Seattle.* [Online].

Fiser, A. 2004. Protein Structure Modeling In The Proteomics Era. *Expert Review Of Proteomics,* 1**,** 97-110.

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D. & Bairoch, A. 2005. Protein Identification and Analysis Tools On The Expasy Server. *In:* Walker, J. M. (Ed.) *The Proteomics Protocols Handbook*. Humana Press.

Ginalski, K., Grishin, N. V., Godzik, A. & Rychlewski, L. 2005. Practical Lessons from Protein Structure Prediction. *Nucleic Acids Research,* 33**,** 1874-1891.

Gordon, M., De Oliveira, T., Bishop, K., Coovadia, H. M., Madurai, L., Engelbrecht, S., Van Rensburg, E. J., Mosam, A., Smith, A. & Cassol, S. 2003. Molecular Characteristics of Human Immunodeficiency Virus Type 1 Subtype C Viruses From Kwazulu-Natal, South Africa: Implications For Vaccine And Antiretroviral Control Strategies. *Journal Of Virology,* 77**,** 2587-2599.

Grant, M. A. 2009. Protein Structure Prediction In Structure-Based Ligand Design and Virtual Screening. *Combinatorial Chemistry And High Throughput Screening,* 12**,** 940-960.

Guindon, S. & Gascuel, O. 2003. A Simple, Fast, And Accurate Algorithm To Estimate Large Phylogenies By Maximum Likelihood. *Systematic Biology,* 52**,** 696-704.

Higgs, P. & Attwood, T. 2005. *Bioinformatics And Molecular Evolution*, Wiley-Blackwell.

Hildebrand, A., Remmert, M., Biegert, A. & Söding, J. 2009. Fast and Accurate Automatic Structure Prediction with HHpred. *Proteins: Structure, Function and Bioformatics,* 77**,** 128-132.

Hillisch, A., Pineda, L. F. & Hilgenfeld, R. 2004. Utility of Homology Models in The Drug Discovery Process. *Drug Discovery Today,* 9**,** 659-669.

Holmes, I. & Durbin, R. 1998. Dynamic Programming Alignment Accuracy. *Journal of Computational Biology,* 5**,** 493-504.

Hornak, V. 2006. Hiv-1 Protease Flaps Spontaneously Open and Reclose In Molecular Dynamics Simulations. *Proceedings Of The National Academy Of Sciences,* 103**,** 915-920.

Huelsenbeck, J. P., Ronquist, F., Nielsen, R. & Bollback, J. P. 2001. Bayesian Inference of Phylogeny and Its Impact On Evolutionary Biology. *Science,* 294**,** 2310-2314.

Iweriebor, B. C., Masebe, T., Nwobegahay, J., Mphahlele, J. M. & Bessong, P. O. 2011. Impact Of Nucleotide Polymorphisms At Drug Resistance Sites On Genetic Barrier In Human Immunodeficiency Virus Type 1 Subtype C Resistance Evolution. *African Journal Of Biotechnology,* 10**,** 15320-15326.

Jakobsen, M. R., Ellett, A., Churchill, M. J. & Gorry, P. R. 2010. Viral Tropism, Fitness And Pathogenicity Of HIV-1 Subtype C. *Future Virology,* 5**,** 219-231.

Kandathil, A. J., 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**,** 511-519.

Kantor, R. & Katzenstein, D. 2004. Drug Resistance In Non-Subtype B HIV-1. *Journal Of Clinical Virology,* 29**,** 152-159.

Karlsson, A. C., Younger, S. R., Martin, J. N., Grossman, Z., Sinclair, E., Hunt, P. W., Hagos, E., Nixon, D. F. & Deeks, S. G. 2004. Immunologic And Virologic Evolution During Periods Of Intermittent And Persistent Low-Level Viremia. *Aids,* 18**,** 981-989.

Katiyar, A., Lenka, S. K., Lakshmi, K., Chinnusamy, V. & Bansal, K. C. 2009. In Silico Characterization And Homology Modeling Of Thylakoid-Bound Ascorbate Peroxidase From A Drought Tolerant Wheat Cultivar. *Genomics, Proteomics And Bioinformatics,* 7**,** 185-193.

Katoh, K., Misawa, K., Kuma, K. I. & Miyata, T. 2002. Mafft: A Novel Method For Rapid Multiple Sequence Alignment Based On Fast Fourier Transform. *Nucleic Acids Research,* 30**,** 3059-3066.

Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H. & Phillips, D. C. 1958. A Three-Dimensional Model of The Myoglobin Molecule Obtained By X-Ray Analysis. *Nature,* 181**,** 662-666.

Kihara, D., Chen, H. & Yang, Y. D. 2009. Quality Assessment of Protein Structure Models. *Current Protein And Peptide Science,* 10**,** 216-228.

Kimura, M. 1980. A Simple Method For Estimating Evolutionary Rates Of Base Substitutions Through Comparative Studies of Nucleotide Sequences. *Journal Of Molecular Evolution,* 16**,** 111-120.

Kirchmair, J., Distinto, S., Liedl, K. R., Markt, P., Rollinger, J. M., Schuster, D., Spitzer, G. M. & Wolber, G. 2011. Development of Anti-Viral Agents Using Molecular Modeling and Virtual Screening Techniques. *Infectious Disorders - Drug Targets,* 11**,** 64-93.

Kirchmair, J., Distinto, S., Schuster, D., Spitzer, G., Langer, T. & Wolber, G. 2008. Enhancing Drug Discovery Through In Silico Screening: Strategies To Increase True Positives Retrieval Rates. *Current Medicinal Chemistry,* 15**,** 2040-2053.

Klebe, G. 2000. Recent Developments In Structure-Based Drug Design. *Journal Of Molecular Medicine,* 78**,** 269-281.

Ko, G. M., Reddy, A. S., Kumar, S., Bailey, B. A. & Garg, R. 2010. Computational Analysis of HIV-1 Protease Protein Binding Pockets. *Journal Of Chemical Information And Modeling,* 50**,** 1759-1771.

Koehl, P. & Delarue, M. 1995. A Self Consistent Mean Field Approach To Simultaneous Gap Closure and Side-Chain Positioning In Homology Modelling. *Nature Structural Biology,* 2**,** 163-170.

Krishnamoorthy, B. & Tropsha, A. 2003. Development of A Four-Body Statistical Pseudo-Potential To Discriminate Native from Non-Native Protein Conformations. *Bioinformatics,* 19**,** 1540-1548.

Kumar, A. & Jadhav, C. 2011. Genotypic Prediction of Resistant Mutation In HIV-1 Pol Gene Towards The Antiretroviral Drugs. *International Journal Of Bioinformatics Research And Applications,* 7**,** 15-23.

Kuritzkes, D. R. 2011. Drug Resistance In HIV-1. *Current Opinion In Virology*.

Lam, T. T. Y., Hon, C. C. & Tang, J. W. 2010. Use Of Phylogenetics In The Molecular Epidemiology And Evolutionary Studies of Viral Infections. *Critical Reviews In Clinical Laboratory Sciences,* 47**,** 5-49.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., Mcgettigan, P. A., Mcwilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J. & Higgins, D. G. 2007. Clustal W And Clustal X Version 2.0. *Bioinformatics,* 23**,** 2947-2948.

Laskowski, R. A., Moss, D. S. & Thornton, J. M. 1993. Main-Chain Bond Lengths and Bond Angles in Protein Structures. *Journal Of Molecular Biology,* 231**,** 1049-1067.

Levitt, M. 1992. Accurate Modeling Of Protein Conformation By Automatic Segment Matching. *Journal Of Molecular Biology,* 226**,** 507-533.

Li, C. P., De Grave, S., Chan, T. Y., Lei, H. C. & Chu, K. H. 2011a. Molecular Systematics of Caridean Shrimps Based On Five Nuclear Genes: Implications For Superfamily Classification. *Zoologischer Anzeiger,* 250**,** 270-279.

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

Lin, K., May, A. C. W. & Taylor, W. R. 2002. Threading Using Neural Network (Tune): The Measure of Protein Sequence-Structure Compatibility. *Bioinformatics,* 18**,** 1350-1357.

Luthy, R., Bowie, J. U. & Eisenberg, D. 1992. Assesment of Protein Models with Three-Dimensional Profiles. *Nature,* 356**,** 83-85.

Lyons-Weiler, J., Hoelzer, G. A. & Tausch, R. J. 1998. Optimal Outgroup Analysis. *Biological Journal Of The Linnean Society,* 64**,** 493-511.

Mahdi, E. S., Khairudin, N. B. A. & Wahab, H. A. 2011. The Influence Of Sequence Alignment Length of Template On The Accuracy And Quality Of Homology Modelling: Human Cytochrome P450 2d6 (Cyp2d6). *International Journal Of Drug Delivery,* 3**,** 386-396.

Makarenkov, V., Boc, A., Xie, J., Peres-Neto, P., Lapointe, F. J. & Legendre, P. 2010. Weighted Bootstrapping: A Correction Method for Assessing The Robustness of Phylogenetic Trees. *BMC Evolutionary Biology,* 10.

Maljkovic-Berry, I., Athreya, G., Kothari, M., Daniels, M., Bruno, W. J., Korber, B., Kuiken, C., Ribeiro, R. M. & Leitner, T. 2009. The Evolutionary Rate Dynamically Tracks Changes In Hiv-1 Epidemics: Application Of A Simple Method for Optimizing The Evolutionary Rate In Phylogenetic Trees With Longitudinal Data. *Epidemics,* 1**,** 230-239.

Mar, J. C., Harlow, T. J. & Ragan, M. A. 2005. Bayesian And Maximum Likelihood Phylogenetic Analyses Of Protein Sequence Data Under Relative Branch-Length Differences And Model Violation. *BMC Evolutionary Biology,* 5.

Melo, F. & Feytmans, E. 1998. Assessing Protein Structures With A Non-Local Atomic Interaction Energy. *Journal Of Molecular Biology,* 277**,** 1141-1152.

Morris, L. 2011. *Re: Email Correspondence with National Institute Of Communicable Diseases*

Muzammil, S., Ross, P. & Freire, E. 2003. A Major Role For A Set of Non-Active Site Mutations In The Development Of HIV-1 Protease Drug Resistance. *Biochemistry,* 42**,** 631-638.

Nayeem, A., Sitkoff, D. & Krystek Jr, S. 2006. A Comparative Study Of Available Software For High-Accuracy Homology Modeling: From Sequence Alignments To Structural Models. *Protein Science,* 15**,** 808-824.

Nielsen, M. H., Pedersen, F. S. & Kjems, J. 2005. Molecular Strategies To Inhibit HIV-1 Replication. *Retrovirology,* 2.

Notredame, C., Higgins, D. G. & Heringa, J. 2000. T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment. *Journal Of Molecular Biology,* 302**,** 205-217.

Ode, H., Ota, M., Neya, S., Hata, M., Sugiura, W. & Hoshino, T. 2005. Resistant Mechanism Against Nelfinavir of Human Immunodeficiency Virus Type Proteases. *Journal Of Physical Chemistry B,* 109**,** 565-574.

Ohtaka, H. & Freire, E. 2005. Adaptive Inhibitors Of The Hiv-1 Protease. *Progress In Biophysics and Molecular Biology,* 88**,** 193-208.

Ohtaka, H., Muzammil, S., Schön, A., Velazquez-Campoy, A., Vega, S. & Freire, E. 2004. Thermodynamic Rules For The Design Of High Affinity HIV-1 Protease Inhibitors with Adaptability To Mutations and High Selectivity Towards Unwanted Targets. *International Journal Of Biochemistry And Cell Biology,* 36**,** 1787-1799.

Palmeira, V. F., Kneipp, L. F., Alviano, C. S. & Santos, A. L. S. D. 2006. The Major Chromoblastomycosis Fungal Pathogen, Fonsecaea Pedrosoi, Extracellularly Releases Proteolytic Enzymes Whose Expression Is Modulated By Culture Medium Composition: Implications On The Fungal Development And Cleavage Of Key's Host Structures. *Fems Immunology And Medical Microbiology,* 46**,** 21-29.

Paraschiv, S., Foley, B. & Otelea, D. 2011. Diversity Of HIV-1 Subtype C Strains Isolated In Romania. *Infection, Genetics And Evolution,* 11**,** 270-275.

Pawlowski, M., Gajda, M. J., Matlak, R. & Bujnicki, J. M. 2008. MetaMQAP: A Meta-Server For The Quality Assessment Of Protein Models. *Bmc Bioinformatics,* 9.

Petrey, D., Xiang, Z., Tang, C. L., Xie, L., Gimpelev, M., Mitros, T., Soto, C. S., Goldsmith-Fischman, S., Kernytsky, A., Schlessinger, A., Koh, I. Y. Y., Alexov, E. & Honig, B. 2003. Using Multiple Structure Alignments, Fast Model Building, and Energetic Analysis In Fold Recognition and Homology Modeling. *Proteins: Structure, Function And Genetics,* 53**,** 430-435.

Petsko, G. A. 2002. An Introduction To Modeling Structure From Sequence. *Current Protocols In Bioinformatics.* John Wiley & Sons, Inc.

Pini, A., Giuliani, A., Ricci, C., Runci, Y. & Bracci, L. 2004. Strategies for The Construction And Use Of Peptide And Antibody Libraries Displayed On Phages. *Current Protein And Peptide Science,* 5**,** 487-496.

Pontius, J., Richelle, J. & Wodak, S. J. 1996. Deviations from Standard Atomic Volumes As A Quality Measure For Protein Crystal Structures. *Journal Of Molecular Biology,* 264**,** 121-136.

Qu, X., Swanson, R., Day, R. & Tsai, J. 2009. A Guide To Template Based Structure Prediction. *Current Protein and Peptide Science,* 10**,** 270-285.

Rhee, S. Y., Gonzales, M. J., Kantor, R., Betts, B. J., Ravela, J. & Shafer, R. W. 2003. Human Immunodeficiency Virus Reverse Transcriptase and Protease Sequence Database. *Nucleic Acids Research,* 31**,** 298-303.

Rhee, S. Y., Taylor, J., Fessel, W. J., Kaufman, D., Towner, W., Troia, P., Ruane, P., Hellinger, J., Shirvani, V., Zolopa, A. & Shafer, R. W. 2010. HIV-1 Protease Mutations and Protease Inhibitor Cross-Resistance. *Antimicrob Agents Chemother,* 54**,** 4253-4261.

Sali, A. & Blundell, T. L. 1993. Comparative Protein Modeling By Satisfaction of Spatial Restraints. *Journal Of Molecular Biology,* 234**,** 779-815.

Sanches, M., Krauchenco, S., Martins, N. H., Gustchina, A., Wlodawer, A. & Polikarpov, I. 2007. Structural Characterization Of B And Non-B Subtypes Of Hiv-Protease: Insights Into The Natural Susceptibility To Drug Resistance Development. *Journal Of Molecular Biology,* 369**,** 1029-1040.

Šašková, K. G., Kožíšek, M., Lepšík, M., Brynda, J., Řezáčová, P., Václavíková, J., Kagan, R. M., Machala, L. & Konvalinka, J. 2008. Enzymatic And Structural Analysis Of The I47a Mutation Contributing To The Reduced Susceptibility To HIV Protease Inhibitor Lopinavir. *Protein Science,* 17**,** 1555-1564.

Saste, V. S., Kale, S. S., Sapate, M. K. & Baviskar, D. T. 2011. Modern Aspects for Antiretroviral Treatment. *International Journal Of Pharmaceutical Sciences Review And Research,* 9**,** 18-24.

Sauder, J. M., Arthur, J. W. & Dunbrack Jr, R. L. 2000. Large-Scale Comparison of Protein Sequence Alignment Algorithms With Structure Alignments. *Proteins: Structure, Function And Genetics,* 40**,** 6-22.

Schonbrun, J., Wedemeyer, W. J. & Baker, D. 2002. Protein Structure Prediction In 2002. *Current Opinion In Structural Biology,* 12**,** 348-354.

Schrodinger, Llc 2010. The Pymol Molecular Graphics System, Version 1.3r1.

Schwede, T., Kopp, J., Guex, N. & Peitsch, M. C. 2003. Swiss-Model: An Automated Protein Homology-Modeling Server. *Nucleic Acids Research,* 31**,** 3381-3385.

Shafer, R. W. 2006. Rationale And Uses of A Public HIV Drug-Resistance Database. *Journal Of Infectious Diseases,* 194**,** S51-S58.

Shafer, R. W., Rhee, S. Y., Pillay, D., Miller, V., Sandstrom, P., Schapiro, J. M., Kuritzkes, D. R. & Bennett, D. 2007. HIV-1 Protease and Reverse Transcriptase Mutations for Drug Resistance Surveillance. *AIDS,* 21**,** 215-223.

Sham, H. L., Kempf, D. J., Molla, A., Marsh, K. C., Kumar, G. N., Chen, C. M., Kati, W., Stewart, K., Lal, R., Hsu, A., Betebenner, D., Korneyeva, M., Vasavanonda, S., Mcdonald, E., Saldivar, A., Wideburg, N., Chen, X. Q., Niu, P., Park, C., Jayanti, V., Grabowski, B., Granneman, G. R., Sun, E., Japour, A. J., Leonard, J. M., Plattner, J. J. & Norbeck, D. W. 1998. ABT-378, A Highly Potent Inhibitor of the Human Immunodeficiency Virus Protease. *Antimicrob Agents Chemother,* 42**,** 3218-3224.

Sharp, P. M. & Hahn, B. H. 2010. The Evolution Of Hiv-1 and The Origin of Aids. *Philosophical Transactions Of The Royal Society B: Biological Sciences,* 365**,** 2487-2494.

Shen, M. Y. & Sali, A. 2006. Statistical Potential for Assessment and Prediction Of Protein Structures. *Protein Science,* 15**,** 2507-2524.

Shi, J. Y., Blundell, T. L. & Mizuguchi, K. 2001. Fugue: Sequence-Structure Homology Recognition Using Environment-Specific Substitution Tables and Structure-Dependent Gap Penalties. *Journal Of Molecular Biology,* 310**,** 243-257.

Sippl, M. J. 1993. Recognition of Errors in Three-Dimensional Structures of Proteins. *Proteins: Structure, Function and Genetics,* 17**,** 355-362.

Sippl, M. J. 1995. Knowledge-Based Potentials for Proteins. *Current Opinion In Structural Biology,* 5**,** 229-235.

Soding, J., Biegert, A. & Lupas, A. N. 2005. The HHpred Interactive Server For Protein Homology Detection and Structure Prediction. *Nucleic Acids Research,* 33**,** W244-W248.

Sokkar, P., Mohandass, S. & Ramachandran, M. 2011. Multiple Templates-Based Homology Modeling Enhances Structure Quality Of AT1 Receptor: Validation By Molecular Dynamics and Antagonist Docking. *Journal Of Molecular Modeling,* 17**,** 1565-1577.

Stebbing, J. & Moyle, G. 2003. The Clades Of Hiv: Their Origins And Clinical Significance. *Aids Reviews,* 5**,** 205-213.

Sunyaev, S. R., Bogopolsky, G. A., Oleynikova, N. V., Vlasov, P. K., Finkelstein, A. V. & Roytberg, M. A. 2004. From Analysis Of Protein Structural Alignments Toward A Novel Approach To Align Protein Sequences. *Proteins: Structure, Function And Genetics,* 54**,** 569-582.

Swanstrom, R. & Erona, J. 2000. Human Immunodeficiency Virus Type-1 Protease Inhibitors therapeutic successes and failures, suppression and resistance. *Pharmacology And Therapeutics,* 86**,** 145-170.

Tomasselli, A. G. & Heinrikson, R. L. 2000. Targeting The Hiv-Protease In Aids Therapy: A Current Clinical Perspective. *Biochimica Et Biophysica Acta - Protein Structure And Molecular Enzymology,* 1477**,** 189-214.

Toor, J. S., Sharma, A., Kumar, R., Gupta, P., Garg, P. & Arora, S. K. 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*.

UNAIDS 2010. Global Report: UNAIDS Report On The Global Aids Epidemic 2010: UNAIDS.

Venclovas, C. 2001. Comparative modeling of casp4 target proteins: combining results of sequence search with three-dimensional structure assessment. *Proteins: structure, function and genetics,* 45**,** 47-54.

Venter, W. D. F., Wilson, D., Conradie, F. & Variava, E. 2006. Antiretroviral Drug Resistance: A guide for the Southern African clinician. *Southern African Journal Of Hiv Medicine***,** 30-36.

Verdone, G., Corazza, A., Colebrooke, S. A., Cicero, D., Eliseo, T., Boyd, J., Doliana, R., Fogolari, F., Viglino, P., Colombatti, A., Campbell, I. D. & Esposito, G. 2009. NMR-based homology model for the solution structure of the C-terminal globular domain of emilin1. *Journal of Biomolecular NMR,* 43**,** 79-96.

Wainberg, M. A. & Brenner, B. G. 2010. Role of hiv subtype diversity in the development of resistance to antiviral drugs. *Viruses,* 2**,** 2493-2508.

Wake, D. B., Wake, M. H. & Specht, C. D. 2011. Homoplasy: From detecting pattern to determining process and mechanism of evolution. *Science,* 331**,** 1032-1035.

Wallner, B. & Elofsson, A. 2003. Can correct protein models be identified? *Protein science,* 12**,** 1073-1086.

Wallner, B. & Elofsson, A. 2005. All are not equal: A benchmark of different homology modeling programs. *Protein science,* 14**,** 1315-1327.

Wallner, B. & Elofsson, A. 2006. Identification of correct regions in protein models using structural, alignment, and consensus information. *Protein science,* 15**,** 900-913.

Wang, H. C., Li, K., Susko, E. & Roger, A. J. 2008. A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny. *Bmc evolutionary biology,* 8.

Wang, L. S., Leebens-Mack, J., Wall, P. K., Beckmann, K., Depamphilis, C. W. & Warnow, T. 2011. The impact of multiple protein sequence alignment on phylogenetic estimation. *Ieee/acm transactions on computational biology and bioinformatics,* 8**,** 1108-1119.

Waterhouse, A., Procter, J., Martin, D., Clamp, M. & Barton, G. 2009. Jalview version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics,* 25**,** 1189-1191.

Wiederstein, M. & Sippl, M. J. 2007. Prosa-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic acids research,* 35**,** w407-410.

Wlodawer, A. & Vondrasek, J. 1998. Inhibitors of hiv-1 protease: a major success of structure-assisted drug design.

Wonderlich, E. R., Leonard, J. A. & Collins, K. L. 2011. HIV immune evasion. Diruption of antigen presentation by the hiv nef protein.

Wu, T. D., Schiffer, C. A., Gonzales, M. J., Taylor, J., Kantor, R., Chou, S., Israelski, D., Zolopa, A. R., Fessel, W. J. & Shafer, R. W. 2003. Mutation patterns and structural correlates in human immunodeficiency virus type 1 protease following different protease inhibitor treatments. *Journal of virology,* 77**,** 4836-4847.

Xiong, J. 2006. *Essential Bioinformatics*, Cambridge University Press.

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**,** 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 of the epidemic. *Nature,* 391**,** 594-597.

Zvelebil, M. J. & Baum, J. O. 2008. *Understanding Bioinformatics,* New York, Garland Science.