MOLECULAR DYNAMICS OF SUBSTRATE RECOGNITION AND CO-EVOLUTION IN HIV-1 PROTEASE

by

Ayşegül Özen

B.S., Chemical Engineering, Boğaziçi University, 2004

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Chemical Engineering

Boğaziçi University

2008

ACKNOWLEDGEMENTS

This research was funded by EU-FP6-ACC-2004-SSA-2 contract No. 517991 and TUBİTAK-BİDEB SSA-2 Project Fellowship.

First and foremost, I would like to express my profound gratitude to Professor Türkan Haliloğlu, my thesis advisor, for her encouragement, careful and wise guidance of my project. I am most grateful as well to Professor Celia Schiffer, whose supervision and advice have contributed substantially to my study.

I would also like to thank Professor Viktorya Aviyente and Professor Pemra Doruker Turgut as members of the Examining Committee. Their thoughtful comments and suggestions were illuminating.

Thank you to my colleagues in PRC who have always supported me during the long process of completing this project.

A special thanks to Duygu Başaran for her help and kind friendliness.

And to my wonderful family, the best part of my life, especially my mother for always being there for me with boundless patience, love, and tenderness.

ABSTRACT

MOLECULAR DYNAMICS OF SUBSTRATE RECOGNITION AND CO-EVOLUTION IN HIV-1 PROTEASE

Human Immunodeficiency Virus Type 1 (HIV-1) protease recognizes at least ten cleavage sites as its natural substrates. There is little sequence homology between these substrates and they are asymmetric around the cleavage site in both charge and size distribution. Thus, understanding of the molecular determinants of substrate recognition is a challenging task as well as of great importance in design of effective drugs. The protease-substrate complex crystal structures indicate that the substrates occupy a remarkable uniform region within the binding site, which has been termed as the substrate envelope. Nevertheless, the activities of proteins are intimately related to the dynamics, from local to global motion of the structure. To this end, an elaborated analysis on both structural and dynamic features of seven HIV-1 protease-substrate complex structures are carried out by molecular dynamics (MD) simulations in the present thesis. The conformations of the complex structures in time have been analyzed with respect to the interaction of the substrate with the protease in terms of the substrate volume, the changes in the van der Waals (vdW) contacts between the two, and the dynamics of both substrate and the protease in general. On the other hand, the co-evolution of the substrate peptides with the drug-resistant protease variants is also analyzed. The MD simulations for the p1-p6 substrates (wild-type and LP1'F) in complex with the protease variants (D30N, N88D, and D30N/N88D) were run and similar analysis to those in wild-type complex structures were made. In this work, the substrate recognition has been observed to be an interdependent event and the recognition mechanism may not be the same for all natural substrates. Also, the dynamic substrate envelope has been found to be smaller than the crystal structures suggest. The analyses of the mutant structures have shown that the substrate recog-

nition is altered when there is drug resistance and this alteration is compensated by co-evolution. The results reveal that the conservation of the peptide conformational preferences and dynamic behavior of the complex structure appears to be important for protease substrate recognition.

ÖZET

HIV-1 PROTEAZDA SÜBSTRAT TANIMA VE EŞEVRİMİN MOLEKÜLER DİNAMİĞİ

İnsan Bağışıklık Yetmezliği Virüsü Tip 1 (HIV-1) proteaz Gag ve Gag-Pol poliproteininin en az on kesim bölgesini doğal sübstrat olarak tanır. Bu sübstratların arasında az miktarda dizi benzetisi olup, sübstratlar kesim bölgesi etrafında yük ve büyüklük bakımından asimetriktir. Bu yüzden, sübstrat tanımanın moleküler belirteçlerini anlamak zor olduğu kadar etkili ilaç tasarımında büyük önem taşımaktadır. Proteaz-sübstrat kristal bileşik yapıları sübstratların bağlanma bölgesinde benzer bir bölgeyi kapladıklarını göstermektedir. Bu bölge sübstrat zarfı olarak adlandırılmıştır. Ancak, proteinlerin aktiviteleri, hem yerel hem global olarak, yapının dinamik hareketleriyle çok yakın ilişkilidir. Bu tezde, yedi proteaz-sübstrat bileşik yapının yapısal ve dinamik özellikleri moleküler dinamik (MD) simülasyonları ile incelenmiştir. Kompleks yapıların zamana bağlı konformasyonları proteaz -sübstrat etkileşimi çerçevesinde analiz edilmiştir. Sübstrat hacmi, proteaz-sübstrat van der Waals (vdW) kontaktları ve genel olarak bileşik yapının dinamik hareketleri olarak özetlenebilir. Diğer yandan, sübstratların ilaca dirençli proteaz çeşitleriyle eşevrimi de çalışılmıştır. p1-p6 sübstratının (yaban tipi ve LP1'F) proteaz variantlarıyla (D30N, N88D, and D30N/N88D) bileşik yapılarının MD simülasyonları yapılmış ve benzer analizler yinelenmiştir. Doğal yapıların analizi sübstrat tanımanın birbirine bağımlı bir süreç olduğunu ve tanıma mekanizmasının tüm doğal sübstratlar için aynı olmayabileceğini; dinamik sübstrat zarfının kristal yapıların gösterdiğinin aksine daha küçük olduğunu, göstermektedir. Mutant yapıların analizi sonuçları, sübstrat tanımanın ilaç direnişi olduğunda değişime uğradığını ve bu değişimin eşevrimle telafi edildiğini desteklemektedir. Sonuçlar genelde peptit konformasyon tercihlerinin ve kompleks sistemin dinamik davranışının korunumunun proteaz-sübstrat tanıma süreci açısından önemli olduğunu ortaya koymaktadır.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS/ABBREVIATIONS

C_α	Alpha-carbon
CO_{ij}	Orientational cross-correlation of positions i and j
d	3-D grid cell edge length
F_i	Force acting on particle i
$g_{i,j,k}$	3-D grid cell with indices i, j, and k
l_i	Lengh of the virtual bond i
L	3-D grid length
m_i	Mass of particle i
N	Number of atoms
ns	nanosecond
N_1	Number of frames in an individual substrate density
N_2	Number of frames in the dynamic substrate envelope
R_i	Position of particle i
S_i	Side chain of the virtual bond i
V	Potential energy
V_{in}^{dyn}	Volume inside the dynamic substrate envelope
V_{in}^{stat}	Volume inside the static substrate envelope
V_{out}^{dyn}	Volume outside the dynamic substrate envelope
V_{out}^{stat}	Volume outside the static substrate envelope
$\mathring{A}$	Angstrom
ϵ	Well depth
σ	Collision diameter
τ	Time frame
ϕ	Pseudodihedral angle
A	Ala, Alanine
AIDS	Acquired Immunodeficiency Syndrome
APV	Amprenavir

ATV	Atazanavir
C	Cys, Cysteine
D	Asp, Aspartic acid
DRV	Darunavir
DNA	Deoxyribonucleic acid
E	Glu, Glutamic acid
F	Phe, Phenylalanine
FDA	Food and Drug Administration
G	Gly, Glycine
HAART	Highly Active Anti-retroviral Therapy
HIV	Human Immunodeficiency Virus
H	His, Histidine
I	Ile, Isoleucine
IDV	Indinavir
K	Lys, Lysine
L	Leu, Leucine
LPV	Lopinavir
M	Met, Methionine
MD	Molecular Dynamics
N	Asn, Asparagine
NFV	Nelfinavir
NMR	Nuclear Magnetic Resonance
MSF	Mean squared fluctuations
P	Pro, Proline
PDB	Protein Data Bank
PI	Protease Inhibitor
PME	Particle Mesh Ewald
PR	Protease
Q	Gln, Glutamine
R	Arg, Arginine
RMSD	Root mean squared deviation

RNA	Ribonucleic acid
RT	Reverse transcriptase
RTV	Ritonavir
S	Ser, Serine
SQV	Saquinavir
T	Thr, Threonine
TPV	Tipranavir
W	Trp, Tryptophan
WT	Wild-type
Y	Tyr, Tyrosine
V	Val, Valine
3-D	Three-dimensional

1. INTRODUCTION

Since the discovery of Acquired Immunodeficiency Syndrome (AIDS) in the early 1980s, the AIDS epidemic has resulted more than 25 million deaths, including women and children. According to the World Health Organization, 2.5 million (1.8 - 4.1 million) people were infected with the human immunodeficiency virus (HIV) in 2007 alone, and 33.2 million (30.6 - 36.1 million) people are currently living with HIV infection (UNAIDS and WHO, 2007). It is thus one of the major research areas in biological sciences where immediate contribution to the understanding of the function of this virus is of great urgency.

In the replication cycle of HIV-1, HIV-1 protease (PR) plays a very important role in the maturation of the virus by cleaving the viral polyproteins to yield infectious mature virions, thus it has been one of the major drug-targets in HIV infection treatment.

First, the function of this enzyme as well as its structure is explained in detail in the first part of this chapter. The biological background of why the long-term inhibition of HIV-1 PR is an extremely challenging task is also given along. Following, the computational design of the present thesis in elucidating the substrate recognition and co-evolution of substrate and protease by molecular dynamics is summarized as a plan of attack with the background literature.

1.1. Biological Background

1.1.1. Human Immunodeficiency Virus

HIV was identified as the causative agent for AIDS in 1983 (Barre-Sinoussi *et al.* 1983). HIV-1 is a member of the retrovirus family (see Figure 1.1). Retroviruses are small envelope viruses that contain a diploid, single-stranded RNA genome and therefore highly prone to mutation. The virus particle contains an inner core that

contains the viral nucleic acids (RNA), as well as enzymes required for early replication events (PR and RT). This inner core is surrounded by capsid proteins. The capsid itself is surrounded by a lipid membrane. A virus matrix protein is inserted into the inner surface of the membrane. An integral membrane protein, the envelope glycoprotein, protrudes through the membrane and forms the outer surface of the virus particle.

The HIV-1 genome has three reading frames: *gag*, *pol*, and *env*, which code several proteins that are essential for virus assembly and replication. Of these genes, *gag*: codes for proteins that make up the viral core, *pol*: encodes an enzyme called reverse transcriptase (RT), and *env*: encodes proteins that make up the viral envelope. RT is present within the virus particle and copies the retroviral RNA sequence into single-stranded DNA when infection of a host cell has occurred. A complementary strand of DNA copy of the retroviral genome integrates into the host cell's DNA. The integrated proviral genome can persist in this state for many years before AIDS symptoms finally develop due to destruction of helper T-cells. Expression of the retroviral genes by the host cell leads to the formation of new copies of the virus. The *gag*, *pol*, and *env* reading frames are expressed as polyproteins, which eventually have to be separated in order for the individual protein molecules to function. This cleavage is performed by the HIV PR, which belongs to the aspartyl protease family and is encoded by part of the *pol* gene.

1.1.2. Structure and Function of HIV-1 Protease

HIV-1 PR is a symmetric homodimeric aspartyl protease that cleaves a series of at least ten non-homologous and asymmetric substrate sites in the Gag and Gag-Pol polyproteins allowing the virus to mature and become infectious (see Figure 1.2). The proteolytic maturation of the virus particles continues after they have budded from the host cell. It has been observed that mutant viruses containing catalytically inactive protease fail to mature and are not infectious (Cooper, 2002). Thus, HIV-1 protease is the target of the most anti-viral drugs for the treatment of HIV-1 infection.

PR was the first HIV-1 protein to be structurally characterized (Miller *et al.*,

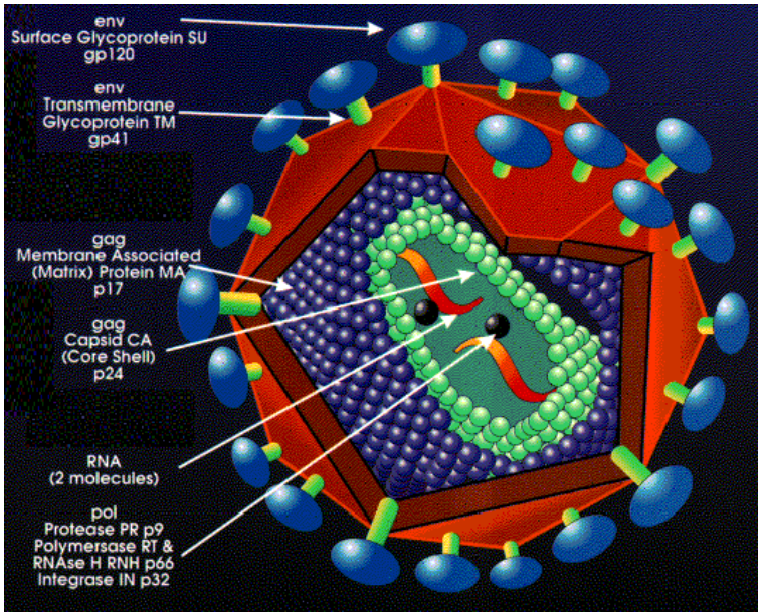


Figure 1.1. The Human Immunodeficiency Virus Type-1 (HIV-1). HIV-1 protease is enclosed in a capsid (green) along with viral RNA and the Reverse Transcriptase enzyme (RT).

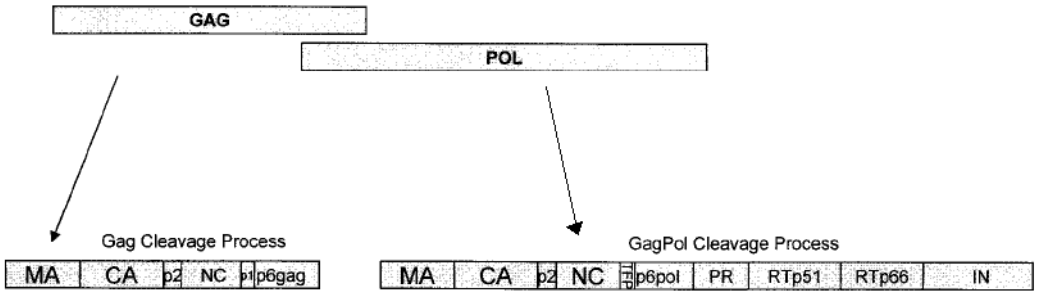


Figure 1.2. Cleavage sites on *gag* and *pol* polyproteins.

1989; Navia *et al.*, 1989; Wlodawer *et al.*, 1989). The enzyme is composed of two identical subunits of 99 amino acid residues each. The sequence of residues is commonly numbered from 1 (N-terminus) to 99 (C-terminus) in subunit 1 and from 1' to 99' in subunit 2. The PR structure is stabilized by a four-stranded antiparallel β-sheet formed by N- and C-terminal β-strands. The enzyme active site is formed at the interface of the two subunits and contains a catalytic triad (Asp25-Thr26-Gly27) responsible for the cleavage reactions. Each monomer contains a "flap" comprising two antiparallel β-strands connected by a loop (residues 49 to 52) and situated on top of the catalytic site (see Figure1.3).

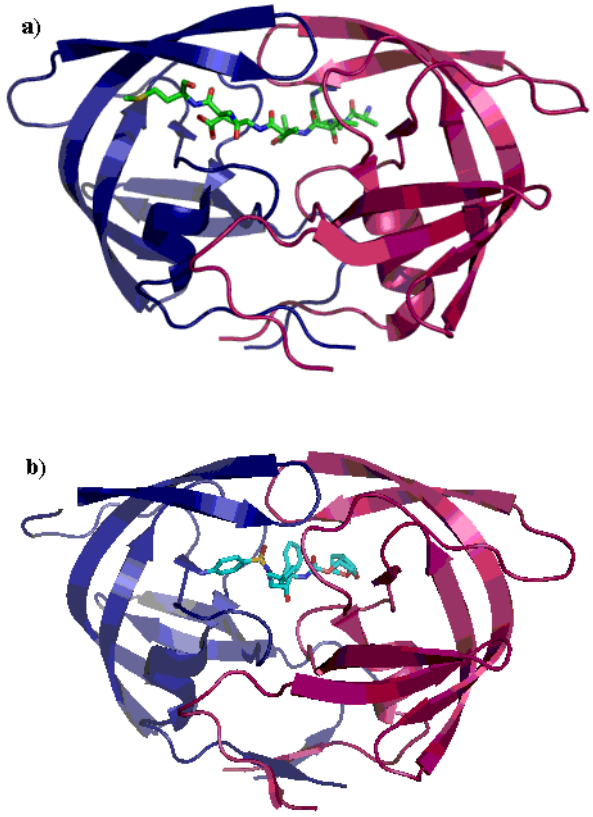


Figure 1.3. X-ray structures of HIV-1 PR in complex with a) cleavage site ca-p2 (PDB code: 1F7A) b) inhibitor Darunavir (PDB code: 1T3R).

1.1.3. Cleavage sites

Figure 1.4 shows the three-dimensional (3-D) crystal structures of seven substrate peptides that the HIV-1 PR recognizes as a cleavage site. Five of these peptides

correspond to substrate sequences within the Gag polyprotein (matrix-capsid [ma-ca], capsid-p2 [ca-p2], p2-nucleocapsid [p2-nc], nucleocapsid-p1 [nc-p1] and p1-p6) and two correspond to substrate sequences within the Pol polyprotein (reverse transcriptase-RNaseH [rt-rh] and RNaseH-integrase [rh-in]) (See also Figure 1.2).

Despite the symmetry conferred on the active site of the PR because of it is a homodimer, the enzyme recognizes asymmetric substrate sites within the *gag* and *pol* polyproteins. The position in the peptide sequence where the cleavage occurs is shown with an asterisk in Table 1.1. The peptides are cleaved around the positions P1 and P1'. The residues from the cleavage position to the N-terminal of the peptide are referred to the unprimed site, where the residues from this point towards the C-terminal are referred to the primed site. The amino acid sequences of these substrates are asymmetric around the cleavage sites in both size and charge distribution (see Figure 1.4). In addition, these share little sequence homology (see Table 1.1). There has been analysis of protease-substrate complex structures indicating that the protease recognizes an asymmetric shape adopted by the substrate peptides rather than a particular amino acid sequence (King *et al.* 2004).

Table 1.1. Amino acid sequences of the substrate cleavage sites of HIV-1 PR.

		P4	P3	P2	P1		P1'	P2'	P3'	P4'
matrix - capsid	ma-ca	S	Q	N	Y	*	P	I	V	Q
capsid - p2	ca-p2	A	R	V	L	*	A	E	A	M
p2 - nucleocapsid	p2-nc	A	T	I	M	*	M	Q	R	G
nucleocapsid - p1	nc-p1	R	Q	A	N	*	F	L	G	K
p1 - p6	p1-p6	P	G	N	F	*	L	Q	S	R
rev. trans. - RNaseH	rt-rh	A	E	T	F	*	Y	V	D	G
RNase-integrase	rh-in	R	K	I	L	*	F	L	D	G

1.1.4. Inhibiting HIV-1 PR in AIDS therapy

As the worldwide AIDS pandemic continues, a cure for HIV-1 still eludes the medical community. In the absence of a cure for HIV-1 pathogenesis, suppression of

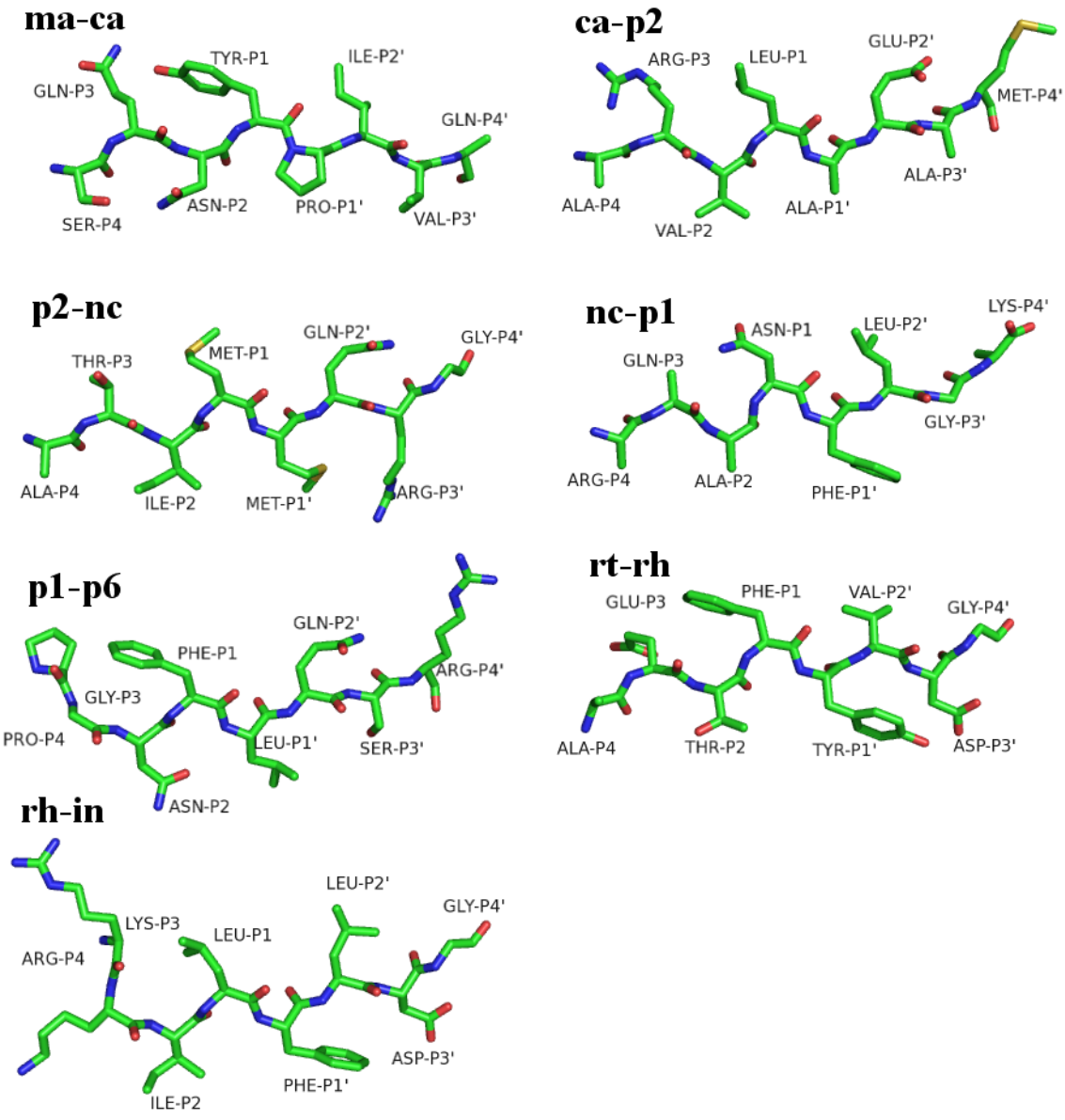


Figure 1.4. Conformation of six substrate peptides as observed in complexes with an inactive yet isosteric variant of HIV-1 protease, D25N. The peptides are colored by atom type.

viral replication and maintaining it at low to undetectable levels have become critical goals in the field of HIV-1 research. To this end, highly active antiretroviral therapy (HAART), which is a treatment approach where combinations of typically three or more antiretroviral drugs are prescribed, has become a successful strategy in providing long, quality life for infected individuals (Hoggs *et al.* 1998). However, viral resistance has been recognized as one of the most important factors involved in therapeutic failure.

HIV-1 protease inhibitors (PI) were the first success of the structure-based drug design (Wlodawer and Erickson 1993) and they are essential components of most HAART therapies (Flexner, 1998; McDonald and Kuritkzkes, 1997). PR is an ideal target for drug therapy since it processes the Gag and Gag-Pol polyproteins in ten or more unique non-homologous sites that are critical for virion maturation and thus the spread of the virus.

There are currently nine US Food and Drug Administration(FDA)-approved HIV-1 PI's; indinavir (IDV), nelfinavir (NFV), amprenavir (APV), saquinavir (SQV), ritonavir (RTV), lopinavir (LPV), atazanavir (ATV), tipranavir (TPV), and most recently darunavir (DRV). All of these are competitive, peptidomimetic inhibitors binding at the active site. All of them have large, generally hydrophobic, moieties that interact with the mainly hydrophobic P2-P2' pockets in the active site (Wlodawer and Erickson, 1993). Although chemically different the 3-D shape and electrostatic character of these drugs are fairly similar as seen in Figure 1.5, therefore a small set of mutations can result in a protease variant with multi-drug resistance. This evolution of drug resistance in HIV-1 PR presents a new challenge to future structure-based drug design efforts.

As HIV-1 PR binds inhibitors in the same active site, understanding how the enzyme recognizes substrates may be crucial for the next generation of inhibitor design (Prabu-Jeyabalan *et al.*, 2002).

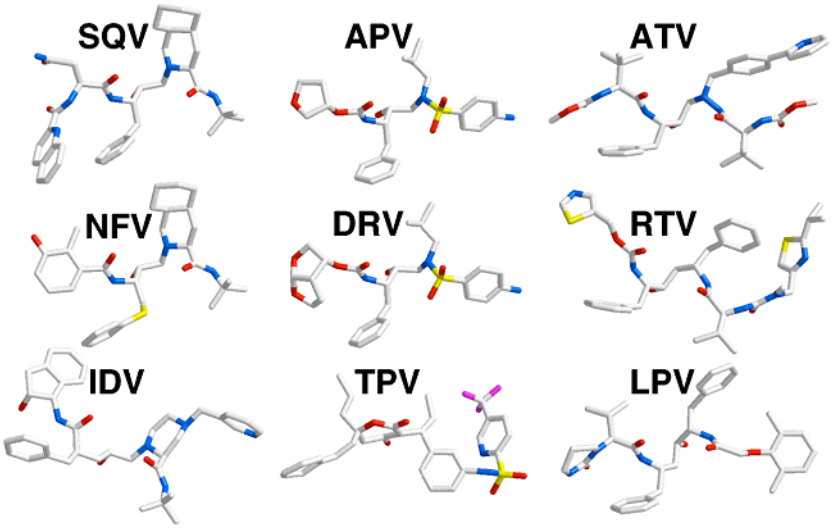


Figure 1.5. The nine FDA-approved HIV-1 PR inhibitors. The structures are colored by atom type.

1.1.5. A Major Problem: Drug Resistance

The introduction of PI's in 1990s revolutionized the treatment of HIV-1 infection. As the basis for combination regimens constituting HAART, PI's turned HIV infection from a death sentence to a manageable chronic disease, which under optimal circumstances, can be held in check indefinitely. Thousands of individuals suffering from the ravages of HIV infection were able to recover immune competence and return to normal lives.

Despite the above unequivocal success, a dilemma remains. Because HIV is a retrovirus with a high rate of replication (Coffin, 1995) , it exists as a quasispecies or swarm of viral variants in pseudoequilibrium, where potential drug resistant mutants are likely to preexist prior to therapy. Also a high intrinsic rate of mutation due to the infidelity of the HIV-1 RT adds to the rapid appearance of resistant viral mutants among treated HIV-1 patients (Ji and Loeb, 1992; Roberts *et al.*, 1988; Roberts *et al.*, 1989)

Due to drug therapy, in general, mutations occur in the target protein. Mutant proteases no longer bind to the inhibitors effectively but maintain their ability to rec-

ognize and cleave substrates. This phenomenon is termed as drug resistance. Since the introduction of PI's, drug-resistant mutations in the protease have become widespread. These mutations render the variant protease resistant to the inhibitor while allowing it to maintain its function in cleaving its natural substrates.

1.1.6. Co-evolution of cleavage sites with HIV-1 PR

HIV-1 PI's bind competitively to the active site of the enzyme. Mutations in the protease that alter inhibitor binding and cause drug resistance can also affect substrate recognition by changing the enzyme's substrate specificity. To compensate, the virus will be under selective pressure to co-evolve the substrate sequence, thereby allowing the protease to retain activity. Recently, it was observed that D30N/N88D mutations within the protease, which are a signature of NFV resistance, have a higher propensity to co-occur with mutations within the p1-p6 substrate. Co-evolution of the p1-p6 substrate with D30N of either L to F or S to N mutations were observed at the P1' and P3' positions, respectively. The previous modeling studies suggest compensating van der Waals contacts as a result of these mutations.

1.1.7. "Substrate Envelope" Hypothesis

Drugs that target key enzymes in the life cycle of HIV-1 have revolutionized the treatment of HIV infection in 1990's. Nevertheless, drug resistance remains a major problem. A novel structure-based strategy was proposed for combating drug resistance, using several HIV-1 PR-substrate and -inhibitor complex structures (Prabu-Jeyabalan *et al.*, 2002).

As described earlier, HIV-1 PR cleaves the HIV-1 *gag* and *pol* precursor proteins in at least ten different locations to allow viral maturation and the amino acid sequences of these substrates share little sequence homology. It was originally speculated (Kirkpatrick, 2004) that this multiplicity for enzyme substrate would make the development of resistance to PI's unlikely. Because the PR would be unable to accommodate mutations necessary to decrease the affinity of drug binding in its active

site without seriously compromising its ability to bind at least one of its substrates. However, this has not turned out to be the case.

Therefore, in order to understand how drug resistance can emerge while the PR retains its ability to recognize its substrates, the crystal structures of an inactive variant of HIV-1 PR in complex with six of its known substrates are solved. Using computational analysis to assess the overlap between the volumes in the active site occupied by these substrates, they identified an "envelope" within the active site that apparently must be made available for substrate binding (Prabu-Jeyabalan *et al.*, 2002).

An analogous analysis (King *et al.*, 2004) with HIV-1 PR in complex with eight different inhibitors also revealed an envelope within the active site occupied by the inhibitors. However, overlaying the "inhibitor envelope" and the "substrate envelope" revealed several regions where the inhibitor envelope protrudes beyond the substrate envelope. And crucially, many of the residues that have been reported to mutate in drug-resistant HIV-1 strains contact these regions. In Figure 2.7, Prabu-Jeyabalan *et al.* (2002) shows the substrate and inhibitor envelopes and how inhibitor envelope fits within the substrate envelope. In Panel A, the substrate envelope is visualized by overlapping van der Waals volume of four or more substrate peptides. The colors of the substrate peptides are red, ma-ca; green, ca-p2; blue, p2-nc; cyan, p1-p6; magenta, rt-rh; and yellow, rh-in. The substrate envelope as it fits within the active site of HIV-1 PR is illustrated in Panel B. The α-carbon trace is of the ca-p2 substrate peptide complex. The inhibitor envelope is given in Panel C from overlapping van der Waals volume of five or more of eight inhibitor complexes. The colors of the inhibitors are yellow, NFV; gray, SQV,; cyan, IDV; light blue, RTV; green, APV; magenta, LPV; blue, ATV; and red, DRV. In Panel D, the inhibitor envelope as it fits within the active site of HIV-1 protease is pictured. Finally in Panel E, superimposition of the substrate envelope (blue) with the inhibitor envelope (red) are shown. Residues that contact the inhibitors where the inhibitors protrude beyond the substrate envelope and confer drug resistance when they mutate are labeled.

It therefore seems that if future drug design focuses on developing inhibitors that

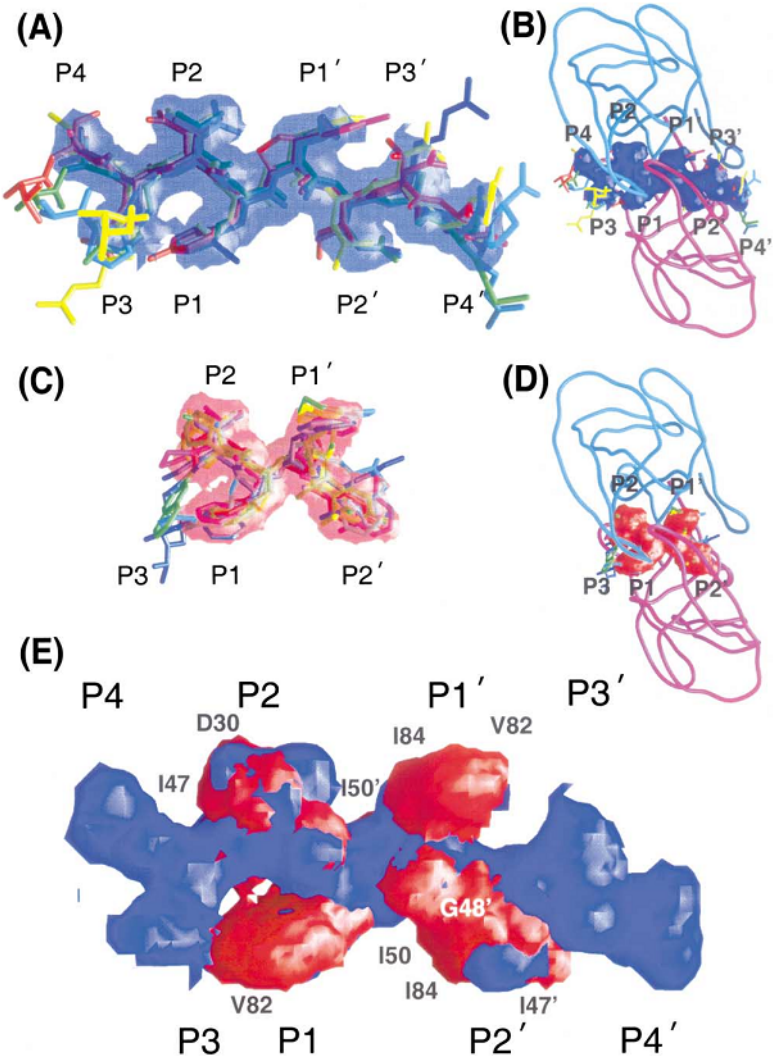


Figure 1.6. A) The substrate envelope. B) The substrate envelope as it fits within the active site of HIV-1 PR. C) The inhibitor envelope. D) The inhibitor envelope as it fits within the active site of HIV-1 protease. E) Superimposition of the substrate envelope (blue) with the inhibitor envelope (red)

interact strongly only with the residues within the substrate envelope of HIV-1 PR, the chances of the drug resistance developing would be reduced. Indeed, recent inhibitors indicate that this might be possible (King *et al.*, 2004) Furthermore; this strategy could also be extended to other enzymes important in viral pathogenesis, such as the serine protease of the hepatitis C virus (Yao *et al.*, 1999).

1.2. Plan of Attack

1.2.1. Protease-Substrate Complex Structures

There are currently seven crystal structures of the substrates which are available in the Protein Data Bank (PDB); ca-p2 (Prabu-Jeyabalan *et al.*, 2000), ma-ca, p2-nc, p1-p6, rt-rh, and rh-in (Prabu-Jeyabalan *et al.*, 2002), and nc-p1 (Prabu-Jeyabalan *et al.*, 2004) in complex with the wild-type protease. Here, these structures are used to study the structural and dynamic basis of the substrate recognition and co-evolution of protease and its substrate.

Nevertheless, crystal structures of D30N and N88D protease variants in complex with WT p1-p6 and D30N/N88D protease variant in complex with either WT or LP1'F substrate variants were not available at the time this study was initiated. Therefore, *in silico* mutants are created by the mutagenesis wizard of the molecular visualization software, PyMol (DeLano, 2002). The WT protease in complex with WT p1-p6 substrate complex is used as template for all mutants.

1.2.2. Substrate Recognition

In the first part of this work, dynamic and structural similarities and/or differences of the wild-type protease-substrate complexes are analyzed. The analyses have three main focuses: (1) constructing the dynamic substrate envelope, comparing that with the static one, and assessing different substrate peptides in terms of their fitness into these envelopes, (2) assessing the van der Waals contacts between the protease and the substrate as an other measure of fitness of substrate into the enzyme ac-

tive site, and (3) investigating the dynamics of the complex system in terms of the fluctuations of atomic positions and pseudodihedral angles in order to elucidate the similarities and/or differences in the flexible/mobile regions of different complexes and the cross-correlations of these fluctuations of different residues.

1.2.2.1. Constructing the Dynamic Substrate Envelope. The substrate envelope was first defined by overlapping van der Waals volume of four or more substrate peptides out of six (Prabu-Jeyabalan *et al.*, 2002). Then the substrate density in the binding site was modeled with a 3-D grid using these six crystal structures (Chellappan *et al.*, 2007). However, proteins are not rigid structures and they fluctuate and undergo conformational changes around their native states taking different conformations. Molecular dynamics (MD) simulations are used in this study to take the full atomistic protein dynamics into account in calculating the shape adopted by the substrates in the active site of the protease. In stead of the crystal structures alone, the coordinate sets of conformations in time, which are the outputs of the MD simulations are used to define the consensus shape adopted by the substrates. Multiple complex conformations are superimposed (see Section 3.1.2) and the peptide structures from all trajectories are used as the input to the original grid model by Chellappan *et al.* (2007). The major difference of the approach presented here from the original model is that the twofold rotational chemical symmetry of the protease is not operated in order to be able to consider the asymmetry of the peptide conformation preference.

1.2.2.2. Evaluating of the Individual Substrates with respect to the Envelope. The protrusion of each substrate density beyond the substrate envelope is calculated using the same grid model (see Equation 2.10). These results are compared with those obtained from the crystal structures. The occurrence frequency of a particular quantity of substrate protrusion beyond the envelope is analyzed in order to understand how frequent a particular substrate lies outside of the envelope or not. The statistical significance of the differences between the volumes of the substrates outside the substrate envelope is assessed by computing the standard deviations.

1.2.2.3. Assessing the van der Waals contacts. Another measure that reflects how well a substrate fits into the active site of the protease is its van der Waals contacts with the protease residues. For the crystal structures, van der Waals potentials of protease and substrate residues were calculated by a simplified Lenard-Jones potential, which will be described in Chapter 2. How these contacts are dynamically conserved is also an important question to be addressed. The calculations are repeated for the conformations obtained from MD frames and the results are compared with the crystal structures. The contacts between the protease and the substrate are focused to study the effect of van der Waals contacts on the substrate recognition, therefore the total van der Waals contact energy of the solvated complex systems reported by AMBER 8 is not used in this study. Protease residues are compared in terms of the contact potentials with the substrate. This analyses was performed for the substrate residues as well. The assessment of van der Waals contacts gives information about the protease residues that plays an important role in substrate recognition and/or binding. Contact potentials of the substrates, on the other hand, may be informative about the crucial residues for binding to the enzyme.

1.2.2.4. Fluctuation dynamics. The atomic positional fluctuations and pseudodihedral angle fluctuations are calculated to see the highly flexible/mobile regions in different protease-substrate complexes. In order to assess how the protease dynamics is coupled with the substrate dynamics, the cross-correlations of these fluctuations are also computed. Time-evolution of these cross-correlations are investigated to see if there is any difference between the dynamic memory of different protease-substrate complex systems.

1.2.3. Co-evolution of p1-p6 substrate with D30N/N88D protease variant

Here, LP1'F mutation co-occuring with the D30N/N88D mutations were studied by the molecular dynamics simulations based on a well-developed forcefield. Previous studies (Kolli *et al.*, 2006) dealt with the structural basis of these co-occuring mutations by means of molecular modeling.

After an energy minimization, MD simulations are set up with the same simulation parameters as the wild-type complexes. The trajectories are analyzed for the substrate volumes protruding beyond the dynamic substrate envelope, the loss or gain in the van der Waals contacts and how frequent a particular value of a parameter being analyzed is seen in the trajectories. The atomic positional fluctuations and pseudodihedral angles are also calculated for the protease-p1-p6 complex variants in order to study the structural and dynamic basis of the co-evolution phenomenon. These analyses would likely elucidate the conservation of structural and dynamic behavior in a protease-substrate complex via co-occuring mutations in protease and substrate.

2. MATERIALS and METHODS

The basic concepts of the methods used in this work, which are molecular dynamics simulations, modeling of the substrate envelope, assessing the van der Waals contacts between the protease and substrate, and cross-correlations of the atomic positional and pseudodihedral angle fluctuations are introduced here.

2.1. Wild-type and Mutant Protease-Substrate Complex Systems studied

Seven wild-type (WT) substrate peptide structures in complex with the inactive variant (D25N) of HIV-1 PR are analyzed in this study. Five of these peptides correspond to substrate sequences within the *gag* polyprotein (matrix-capsid [ma-ca], capsid-p2 [ca-p2], p2-nucleocapsid [p2-nc], nucleocapsid-p1 [nc-p1] and p1-p6) and two correspond to substrate sequences within the *pol* polyprotein (reverse transcriptase-RNaseH [rt-rh] and RNaseH-integrase [rh-in]). The PDB codes for the crystal structure of each substrate-protease complex are listed Table 3.1.

In silico mutations are created on the WT protease-p1-p6 complex structure (1KJF) to get four different mutants (D30N, N88D, D30N/N88D, D30N/N88D/LP1'F) using PyMOL's mutagenesis wizard. The rotamers in PyMOL's library are ordered according to their frequencies of occurrence in proteins and the highest probable side-chain orientation was selected for each mutant residue from the rotamer library.

2.2. Molecular Dynamics Simulations

2.2.1. Theoretical Background

In order to predict the time-dependent events occurring in a molecular system on the atomistic scale, molecular dynamics simulations are widely used. In MD simulations, atoms are allowed to interact with each other using empirical potential energy functions or forcefields, from which the forces on atoms are extracted for a given con-

figuration. Successive configuration of the system is obtained by the integration of Newton's equation of motion, which is:

$$\frac{d^2\mathrm{R_i}}{dt^2} = \frac{\mathrm{F_i}}{m_i} \tag{2.1}$$

In equation 2.1, the motion of a particle with a mass of m_i along the direction of R_i under the force of F_i in that direction is described.

Forcefields describe the potential energy of a system as a function of the atomic positions/coordinates. MD simulations are based on an empirical model of interactions within a system involving stretching and rotation of bonds, as well as non-bonded interactions within a system.

$$\begin{aligned}(\mathrm{R}_1, ..., \mathrm{R}_N) = &\sum_{bonds} \frac{k_i}{2}(l_i - l_{i,0})^2 + \\ &\sum_{angles} \frac{k_i}{2}(\theta_i - \theta_{i,0})^2 + \\ &\sum_{torsions} \frac{V_n}{2}(1 + cos(nw - \gamma)) + \\ &\sum_{i=1}^{N}\sum_{j=i+1}^{N}\left(4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}\right)\end{aligned} \tag{2.2}$$

In equation 2.2, denotes the potential energy, which is a function of the positions (R_i) of N atoms or particles. The first term in the equation describes the interaction of pairs of bonded atoms, where l_i is the bond length. The second term is similarly the summation over all the angles in the molecule modeled using a harmonic potential, where θ_i is the angle between the three successive atoms. Torsional potential describes the change in energy when a bond rotates, and is depicted with the third term in the

equation. The fourth contribution in the equation is for the non-bonded atoms, which are separated by at least three atoms. The non-bonded interactions are defined by two different potentials. The former one is the Lennard-Jones 12-6 potential function that accounts for van der Waals interactions, whereas the latter one is the Coulomb potential for electrostatic interactions. There may be terms that are more complicated in the force fields other than these basic four components (Leach, 2001).

Potential energy of a macromolecular system is a multi-dimensional function of the atomic coordinates; hence, protein fluctuates in a multi-dimensional energy surface. To predict the geometries of the system at the minimum points, minimization algorithms are used. Energy minimization prior to the MD simulation provides a better starting conformation, removes the steric overlaps in the structure and relaxes the bond lengths and angles. The task is to minimize the energy of the system according to 3N atomistic coordinates, which is not a trivial task. Steepest descent and conjugate gradient are widely used minimization algorithms to solve this nonlinear optimization problem.

Steepest descent method is performed prior to the conjugate gradient method due to its quick convergence ability in finding the minima. To determine the exact location of the minimum point, the minimization is then switched to conjugate gradient method.

In classical MD simulations, the initial configuration of the system should be introduced by specifying 3N atomistic coordinates (R_i) of the structure. This structure is generally obtained from the experimental data, such as X-ray or NMR structure of a protein. It is meaningful to select a starting conformation that is close to the desired state of the protein, generally minimum energy/native state. Furthermore, any high-energy interactions in the system may cause instabilities during the simulation; therefore, an energy minimization is performed prior to the simulation.

In order to emphasize boundary effects in the simulation, periodic boundary conditions are used. By the utilization of periodic boundary conditions it is possible to include the solvent effect with a relatively small number of particles. In periodic

boundary conditions, particles are enclosed in a solvent box; this box is replicated to infinity by rigid translation in all the three cartesian directions, completely filling the space. The shape of the solvent box may be a truncated octahedron, a cube or a hexagonal prism depending on the shape of the initial structure.

When the initial configuration of the system is minimized in a solvent box, it is required to assign initial velocities at $t = 0$. The initial velocities of the system are assigned according to the Maxwell-Boltzmann distribution at the initial temperature (Leach, 2001).

After the system is initialized, i.e. put in a solvent box and assigned initial velocities, the potential energy of the system can be calculated, and hence the force on each atom from the derivative of potential energy is determined by,

$$\mathrm{F}_i = - \bigtriangledown V_i(\mathrm{R}_1, ..., \mathrm{R}_N) = -\frac{\partial V(\mathrm{R}_1, ..., \mathrm{R}_N)}{\partial \mathrm{R}_i} \tag{2.3}$$

Once the forces on each atom at the current time t are calculated, the next step is to produce the new conformation at time t+Δt by integrating Newton's equation of motion (Equation 2.1). Various algorithms have been designed to do this. Verlet algorithm is one of those integration schemes. The Verlet algorithm is fast, although this may not be the highest concern in many cases. Second, it is not particularly accurate for long time steps. Third, it requires about as little memory as possible. The leapfrog algorithm, which is equivalent to the Verlet algorithm, is used in this study. This integrator evaluates the velocities at half-integer time steps and uses the velocities to compute new positions. To derive the leapfrog algorithm from the Verlet scheme, the velocities at half-integer time steps are defined as:

$$v\left(t-\Delta t/2\ \right)\equiv\frac{r\left(t\right)-r\left(t-\Delta t\right)}{\Delta t} \tag{2.4}$$

and

$$v\left(t+\Delta t/2\ \right)\equiv\frac{r\left(t+\Delta t\ \right)-r\left(t\right)}{\Delta t}. \tag{2.5}$$

From the latter equation an expression is obtained for the new positions, based on the old positions and velocities:

$$r\left(t+\Delta t\ \right)=r\left(t\right)+\Delta t v\left(t+\Delta t/2\right) \tag{2.6}$$

From the Verlet algorithm, the following expression is obtained for the update of the velocities:

$$v\left(t+\Delta t/2\ \right)=v\left(t-\Delta t/2\right)+\Delta t\frac{F_i\left(t\right)}{m_i} \tag{2.7}$$

As the leapfrog algorithm is derived from the Verlet algorithm, it gives rise to identical trajectories. The trajectories are not defined at the same time as the positions. As a consequence, kinetic and potential energy are also not defined at the same time, hence the total energy can not be computed directly in leapfrog integrator (Frenkel and Smit, 1996).

2.2.2. Simulation Parameters

The AMBER (Case *et al.*, 2004; Case *et al.*, 2005) simulation package with the ff03 forcefield was used in all simulations. The protein was solvated explicitly in a truncated octahedron box using the TIP3P water model (Jorgensen *et al.*, 1983). Bonds involving hydrogens are constrained by the use of the SHAKE algorithm (Ryckaert *et al.*, 1977) with a relative geometrical tolerance of 10E-5. Initial atom velocities corresponding to a temperature of 10 K were generated from a Maxwellian distribution and the temperature was gradually raised to 300 K. The temperature was maintained at 300 K, and the pressure was maintained at 1 bar by the Berendsen weak-coupling approach (Berendsen *et al.*, 1984). Constant pressure periodic boundary conditions are used with isotropic position scaling. The Particle Mesh Ewald (PME) method (Essman *et al.*, 1995) was used to calculate the full electrostatic energy of a periodic box, bypassing pairlist creation and nonbonded force and energy evaluation by calling special PME functions to calculate the Lennard-Jones and electrostatic interactions with a cutoff distance of 9 Å time step of 2 fs was employed in the Leapfrog integrator. Coordinates and energies are written every 0.4 ps.

2.3. Modelling of the Substrate Envelope

2.3.1. Static Substrate Envelope

The static substrate envelope is computed based on the crystal structures of seven substrates in complex with HIV-1 PR. The static substrate envelope and the fit of each substrate within this envelope were evaluated by a 3-D grid-based method based on the model by (Chellappan *et al.*, 2007). Crystal structures are overlaid. The amino acid sequences of the substrates are asymmetric around the cleavage sites in both size and charge distribution. In order to have a more realistic model, the chemical C2 rotational symmetry of the receptor structure was not used during superimposition in this study in contrast to the original model by (Chellappan *et al.*, 2007).

A cubic 3-D grid with side length, L, and grid spacing, *d*, was centered on the

active site. An initial value of 0 was assigned to each grid point. Then a value of g_{ijk} was incremented by 1 for every substrate structure that contains the grid point (i,j,k), where a grid point was considered to be contained by a substrate if it lay within the AMBER van der Waals radius of any nonhydrogen atom of the substrate. The resulting grid values vary between 0 (outside all substrates) and N (inside all substrates), which is the total of structures used in the calculation.

The effective volume of a single substrate structure outside the static substrate envelope defined by the seven crystal structures, V_{out}, is computed by summing the values of the grid points g_{ijk} that lie within the van der Waals volume of the substrate, normalizing the sum by N and converting to a volume by multiplying by the volume of a single grid cell (d^3):

$$V_{out}^{stat} = \frac{d^3}{N} \sum_{i,j,k}^{inside} (N - g_{ijk}) \tag{2.8}$$

As a control, the effective volume of a substrate that lies within the static substrate envelope is computed as:

$$V_{in}^{stat} = \frac{d^3}{N} \sum_{i,j,k}^{inside} (g_{ijk}) \tag{2.9}$$

The total volume of a single substrate structure is computed by adding these two quantities.

2.3.2. Dynamic Substrate Envelope

Instead of crystal structures, several snapshots from trajectories of MD simulations with crystal structures as initial structures are used in the computation of the dynamic substrate envelope. The dynamic substrate envelope and the fit of each substrate density within this envelope were evaluated by the same 3-D grid-based method. All frames from different trajectories are superimposed. The cubic 3-D grid has the same side length, L, and grid spacing, d, and it is located at the same position as in the static envelope. An initial value of 0 was assigned to each grid point. Then a value of $g_{ijk,2}$ was incremented by 1 for every snapshot that contains the grid point (i,j,k), where a grid point was considered to be contained by a substrate frame if it lay within the AMBER van der Waals radius of any nonhydrogen atom of the substrate. The resulting grid values vary between 0 (outside all substrates) and N_2 (inside all substrates), which is the total of frames from multiple MD simulations.

The density of an individual substrate is computed from N_1 frames from MD simulation trajectory for that substrate. An initial value of 0 was again assigned to each grid point. Then a value of $g_{ijk,1}$ was incremented by 1 for every snapshot that contains the grid point (i,j,k), where a grid point was considered to be contained by a substrate frame if it lay within the AMBER van der Waals radius of any nonhydrogen atom of the substrate. The resulting grid values vary between 0 (outside all substrates) and N_1 (inside all substrates), which is the total of frames from a single MD simulation.

The effective volume of an individual substrate density outside the dynamic substrate envelope defined by the MD frames is computed by summing the values of the grid points $g_{ijk,1}$ multiplied by the value (N_2 - $g_{ijk,2}$), which shows how dense the dynamic substrate envelope is at that grid point, and normalizing the sum by N_1N_2. The result is obtained after converting to a volume by multiplying by the volume of a single grid cell (d^3):

$$V_{out}^{dyn} = d^3 \sum_{i,j,k}^{allgridpoints} \frac{N_2 - g_{ijk,2}}{N_2} \times \frac{g_{ijk,1}}{N_1} \tag{2.10}$$

As a control, the effective volume of an individual substrate density that lies within the dynamic substrate envelope is computed as:

$$V_{in}^{dyn} = d^3 \sum_{i,j,k}^{allgridpoints} \frac{g_{ijk,2}}{N_2} \times \frac{g_{ijk,1}}{N_1} \tag{2.11}$$

The total volume of an individual substrate density is computed by adding these two quantities.

2.4. Assessment of van der Waals contacts

HIV-1 PR is a relatively "plastic" enzyme that adjusts in conformation depending on what mutations are present and what ligand is bound in the active site. These adjustments take the form of changes in both the back-bone and side-chain conformations. Cross-correlations of atomic fluctuations give a pair-wise representation of how the positional fluctuation of a single back-bone or side-chain atom correlates with another in a particular structure. This representation for one protein-substrate complex structure can be compared to another as a general assessment of what is different in two complex structures. How a particular ligand interacts with a particular protease variant can be assessed in terms of detailed molecular interactions i.e., hydrogen bonds, local water structure and van der Waals contacts. While hydrogen bonds and interactions with water molecules can be easily counted based on distance and geometry assessing van der Waals contacts accurately is a more subtle analysis; often a simple distance criterion of 4.2 - 5.0 Å does not accurately distribute the energetic contributions. The

van der Waals interaction energy can be calculated as a "6-12" or "Lennard-Jones" potential, with a long range shallow attractive interaction and a short range repulsive one (see Figure 2.1).

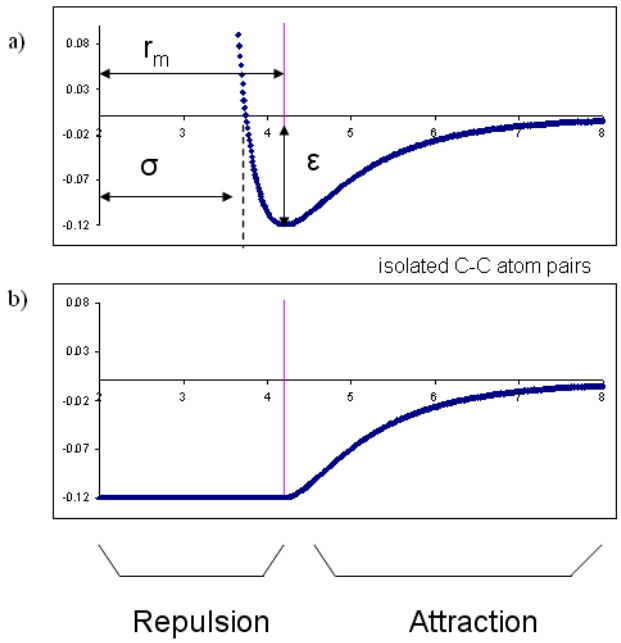


Figure 2.1. Plot of the a) the van der Waals equation and b) simplified function.

This approach allows the van der Waals potentials to be calculated and compared for each substrate in complex with the wild-type and mutant protease. Also the distribution of the van der Waals contacts is shown on a residue-by-residue basis. Thus the contacts can be quantitatively compared for a series of structures, such as a particular substrate with a series of protease variants or series of different substrates with the same protease. In addition, the role of a particular amino acid residue can be compared for a series of substrates. During our MD simulations the movements of the amino acid residues are handled by taking also the repulsive term into account, it is not likely to have unfavorable van der Waals interactions in MD frames. Therefore, we apply this technique to the analysis of both the crystal structures and MD simulation frame snapshots to be consistent and to have comparable results.

2.5. Atomic Positional Fluctuations

The normalized time-delayed orientational cross-correlations between residue pairs are defined as

$$CO_{i,j}(\tau) = \frac{\langle \Delta R_i(t) \Delta R_j(t+\tau) \rangle}{\langle \Delta R_i(t)^2 \rangle^{1/2} \langle \Delta R_j(t+\tau)^2 \rangle^{1/2}} \tag{2.12}$$

where $\Delta R_i(t)$ is the fluctuation in the position vector R_i of site i at time $t$and $\Delta R_j(t+\tau)$ is the fluctuation in the position vector R_j of site j at time $t+\tau$. The brackets represent time averages over recorded snapshots. The cross-correlations vary in the range [-1, 1] with the lower and upper limit indicating fully anti-correlated and correlated atomic fluctuations, respectively. $\tau = 0$ gives the equal-time cross-correlations of atomic fluctuations.

2.6. Pseudodihedral Angle Fluctuations

Schematic presentation of the virtual bond model (Flory, 1969) for a protein segment between backbone units: $\alpha - C_{i-2}$ and $\alpha - C_{i+1}$ is given in Figure 2.2. Side-chain attached to the j^{th} α-carbon is represented by S_i. l_i is the i^{th} virtual bond extending from $\alpha - C_{j-1}$ to $\alpha - C_j$. ϕ_i is the rotational angle of the i^{th} virtual bond, defined by the respective locations of the four backbone units $\alpha - C_{i-2}$,$\alpha - C_{i-1}$, $\alpha - C_i$, $\alpha - C_{i+1}$. θ_i is the bond angle between l_i and l_{i+1}. θ_i^s is the bond angle between l_j and l_i^s, where l_i^s is the side-chain virtual bond connecting $\alpha - C_i$ to S_i.

The average of the rotational angle change of the virtual bond i occurring within a time period τ is calculated as $\langle \Delta\phi_i(\tau) \rangle = \langle \phi_i(t+\tau) - \phi_i(t) \rangle$.

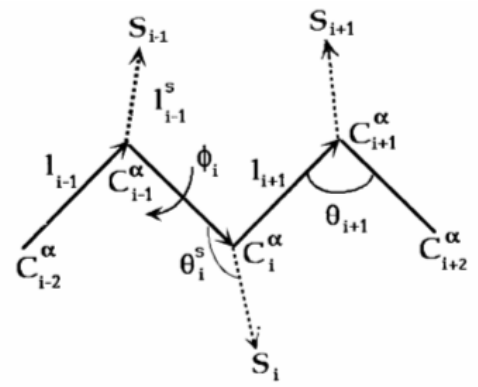


Figure 2.2. Virtual bond model.

3. RESULTS AND DISCUSSION

This chapter starts with a general assessment on how multiple conformations of protease-substrate complexes by MD simulations should be overlaid to define a realistic model of the substrate envelope. After this general assessment, the fit of each substrate into the substrate envelope is given together with a comparison of static and dynamic envelopes. Following, van der Waals contacts between the substrate and the protease are elaborated for an additional measure to reflect how well a particular substrate fits into the active site of the enzyme and identify the conserved and crucial contacts for the binding of substrates. Along, the dynamics of the complex structures are analyzed in terms of fluctuations of the residues and the pseudihedral angles. Following, the structural and dynamic basis of the co-evolution phenomenon is elucidated by the level of protrusion beyond the substrate envelope, the assessment of van der Waals contacts, and the fluctuation dynamics.

3.1. Conformational Data from MD Simulation Trajectories

3.1.1. Sampling

Table 4.1 lists 11 WT and mutant structures for which MD simulations are set up as described in Chapter 2. The crystal structures of the substrates; ca-p2 (Prabu-Jeyabalan *et al.*, 2000), ma-ca, p2-nc, p1-p6, rt-rh, and rh-in (Prabu-Jeyabalan *et al.*, 2002), and nc-p1 (Prabu-Jeyabalan *et al.*, 2004) in complex with the wild-type protease are used in this work. PDB codes for each structure and simulation length is also given in this table.

The coordinates were recorded at every 0.4 ps, resulting 27,500 frames for a 11 ns simulation. For the dynamic substrate envelope calculations this number increases to 192,500 (7 x 27,500), which makes the calculations time-costly. It is desired to include the information in the full trajectory as much as possible in the dynamic substrate envelope calculations with minimum computation time. Therefore, data is extracted

Table 3.1. Complexes studied and the simulation lengths.

	Substrate	PDB code	Simulation length (ns)
WT structures			
1	ma-ca	1KJ4	11
2	ca-p2	1F7A	11
3	p2-nc	1KJ7	11
4	nc-p1	1TSU	11
5	p1-p6	1KJF	11
6	rt-rh	1KJG	11
7	rh-in	1KJH	11
p1-p6 mutant structures			
8	D30N	1KJF	15.5
9	N88D	1KJF	15.5
10	D30N/N88D	1KJF	15.5
11	D30N/N88D/LP1F	1KJF	40.5

from the full trajectory with equal intervals at every 20 ps instead. This results a sample coordinate set of 550 frames. Potential energy values, the root-mean-square deviation (RMSD) values, and mean-square fluctuations (MSF) are investigated for both the full trajectory and the sample set of 550 frames in order to make sure that information lost is at minimum. Figure 3.1 shows the potential energy variation throughout the simulation for the cleavage site, maca, for a) a sample set of 550 frames with 20 ps intervals and b) the full trajectory with 0.4 ps intervals. In Figure 4.2, RMSD with respect to the initial structure. In both figures, it is visually inspected that the reduced sample set represents the full trajectory. The potential energy and RMSD plots for the other substrates are given in Appendix A.

The MSF for $\alpha-$carbon of each residue was calculated for the data recorded at every 0.4 ps and 20 ps for the whole PR-substrate complex structure and the substrate region only. Correlation coefficients of fluctuations from two sets are calculated. The results, which are given in Table 3.2 , show that the atomic positional fluctuations in the full trajectory are represented very well in the reduced set of 550 frames.

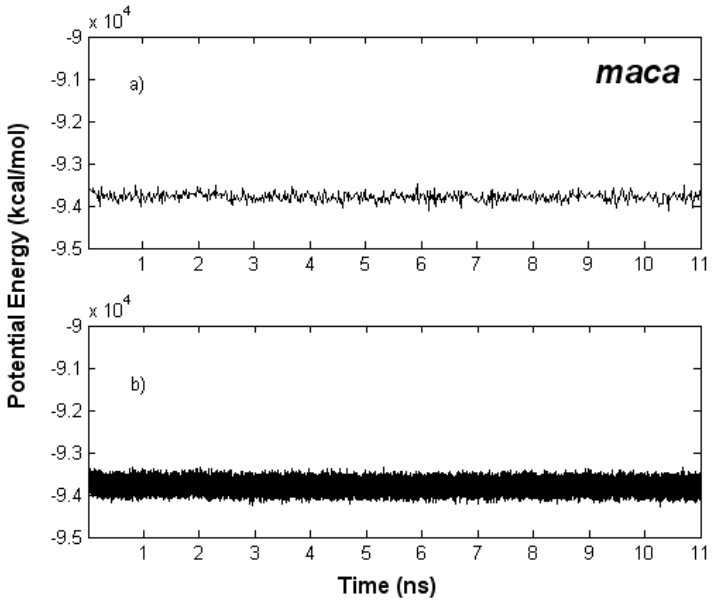


Figure 3.1. Potential energy of maca. a) 550 sample frames taken with equal intervals. b) Samples taken at each 5 time steps.

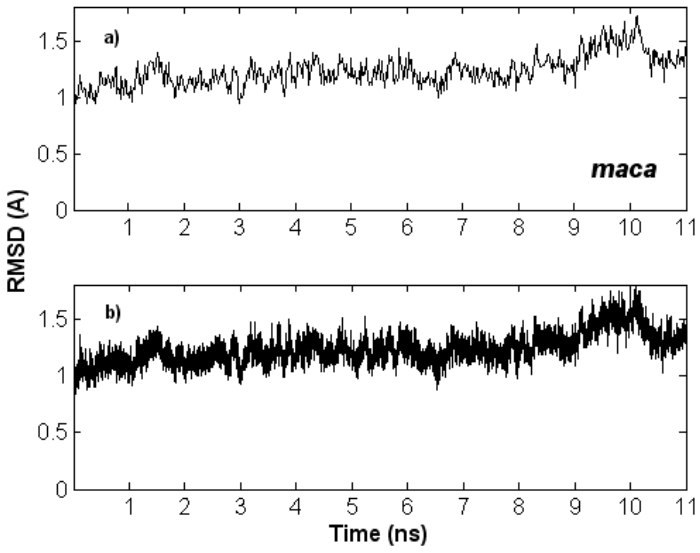


Figure 3.2. Root-mean-square distance of protease-ma-ca substrate structure using backbone carbons for a) 550 sample frames taken with equal intervals and b) Samples taken at each 5 time steps.

Table 3.2. Correlation between the mean square fluctuations of full trajectory and a sample set of 550 frames.

	ma-ca	ca-p2	p2-nc	nc-p1	p1-p6	rt-rh	rh-in
Whole complex	0.9994	0.9991	0.9993	0.9992	0.9994	0.9992	0.9996
Substrate only	0.9998	0.9995	0.9998	0.9999	0.9992	1.0000	0.9991

Although the peptides used for crystallization span a slightly different region of the cleavage sites (see Table 3.3), these differences were previously reported to be likely to have minimal or no effect on the bound conformation of the substrate. The presence or absence of positions P5 and P5', which are outside of P4-P4', are unlikely to impact how the substrate is packed in the active site (Prabu-Jeyabalan *et al.*, 2004). Therefore, these positions, if they exist in the crystal structure, were kept in the simulations, however they were not included in the substrate envelope calculations.

Table 3.3. Cleavage sites spanned in the crystal structures.

Substrate	Cleavage site
ma-ca	P5-P4'
ca-p2	P5-P4'
p2-nc	P4-P4'
nc-p1	P4-P4'
p1-p6	P5-P5'
rt-rh	P4-P5'
rh-in	P5-P5'

3.1.2. Superimposition

Substrate envelope hypothesis claims that the natural substrates of HIV-1 PR adopt a specific consensus shape in 3-D space (King *et al.*, 2004). How the structures are superimposed is an important factor that determines the substrates indeed takes the same shape or not. The best way to see the conformational differences in a series of structures is to overlay those structures using the most rigid residues; otherwise the motion could be propagated to actually immobile regions. For that purpose, the mean-

square fluctuations of each PR α-carbon are plotted for each trajectory (See Figure 3.3). The most immobile residues are observed to be 24-26, 85-90, 24'-26', and 85'-90'. The trajectories are overlaid again using only these very immobile regions in order to see only the movement of the more mobile parts. Both the crystal structures and MD simulation frames are superimposed on the WT PR-cap2 complex crystal structure by the RMSD trajectory tool of the molecular visualization software, VMD (Humphrey *et al.*, 1996).

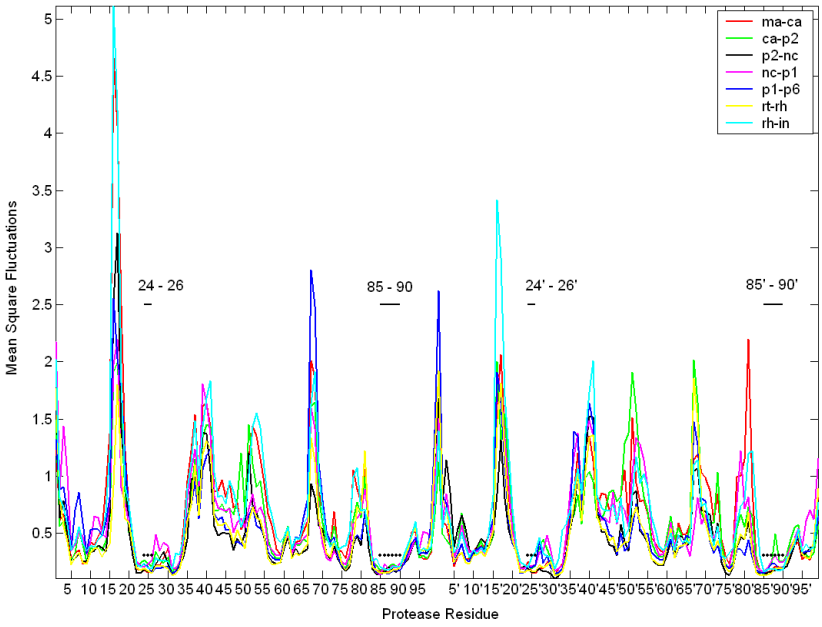


Figure 3.3. Mean square fluctuations of protease residues in different substrate complex simulations

3.2. Wild-type Substrate Complexes

3.2.1. Static and Dynamic Envelopes

The static and dynamic substrates envelopes and the fit of each substrate within these envelopes were evaluated by the 3-D grid-based method that we developed (see Section 2.3). The side-length and grid spacing of the cubic grid, which is centered on the active site, are 54 Å and 0.2 Å, respectively.

Seven crystal structures are used in the calculation of the static substrate envelope and 3850 conformations of the complex structure (550 frames for each of total seven

trajectories) for the dynamic substrate envelope.

The substrate volume lying within the dynamic substrate envelope, V_{in}^{dyn}, and the static substrate envelope, V_{in}^{stat}, and the substrate volume lying outside of the dynamic substrate envelope, V_{out}^{dyn}, and the static substrate envelope, V_{out}^{stat}, are given in Figure 3.4, together with the total volumes, V_{tot}^{dyn} and V_{tot}^{stat}. It is observed in Figure 3.4 that the dynamic substrate envelope is smaller than the static one. The increase in the number of conformations in dynamic case causes the dynamic substrate envelope smoother than the static one. The substrates adopt various conformations slightly different than the crystal conformations, thus the volume that is in consensus becomes smaller. V_{in} and V_{out} values get effected from this difference in the two envelopes. Since the total substrate envelope gets smaller dynamically, the volume that lies outside of this consensus volume gets larger. The total volume of the static substrate envelope is 1552.3 $\pm$ 99.6 $\mathring{A}^3$ and that of the dynamic one is 1412.5 $\pm$ 66.4 $\mathring{A}^3$. The standard deviation is lower in the dynamic case, indicating that the error when higher number of conformations is taken into account is less than the case where the protein is considered to be static. In the static case, the substrate that protrudes the most beyond the substrate envelope is ma-ca whereas nc-p1 is the substrate sticking out of the dynamic substrate envelope the most. This indicates that the potential protrusion of either drugs or substrates beyond the envelope differs when the protein dynamics is taken into account and the consensus shape that the substrates adopt is also dynamic.

The V_{in} and V_{out} values plotted in Figure 3.4 are tabulated in Table 3.4 so that the values and their standard deviations can be investigated relatively quantitatively. Ideally, V_{tot} should have a standard deviation of zero for each substrate since the number of atoms in each frame is the same. However, the use of a discrete 3-D grid causes errors in calculating the total substrate volume due to the possibility that in one particular frame any two substrate atom fill the same grid cell while in another frame they may not. This error never exceeds 1% for either of the substrates; thus, it is neglected. The largest substrate is rh-in, yet it is not the most protruding substrate. Nc-p1, which is known to co-evolve with the protease, is the most protruding substrate beyond the envelope. Its larger volume outside the envelope will likely make the nc-p1 substrate

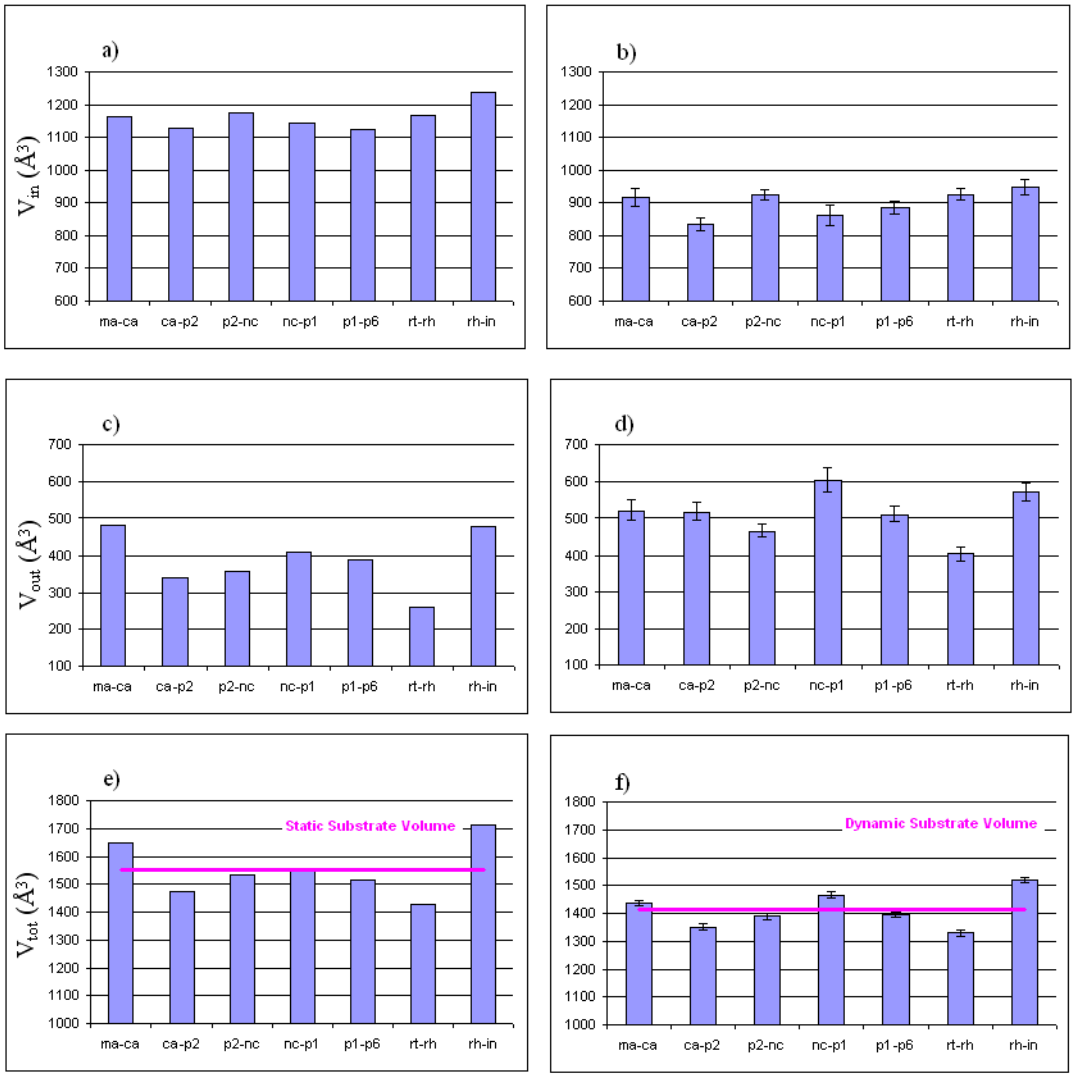


Figure 3.4. For WT substrates a)V_{in}^{stat} b) V_{in}^{dyn} c) V_{out}^{stat} d) V_{out}^{dyn} e) V_{tot}^{stat} f) V_{tot}^{dyn}

vulnerable to structural changes in the protease. On the other hand, p1-p6, which is another co-evolving substrate, does not significantly protrude more than the other substrates. This may be likely due to the possibility that a substrate is susceptible to co-evolve with the protease when it already protrudes beyond the substrate envelope in wild-type complex or it may move outside of the envelope as a result of the structural changes in the protease mutant with respect to wild-type and a need for co-evolution may arise. Table 3.4 also includes the number of nonhydrogen atoms as a control measure of molecular size. V_{out} shows a larger variation than the other molecular size measure.

Table 3.4. Numerical values of V_{in}^{stat}, V_{in}^{dyn}, V_{out}^{stat}, V_{out}^{dyn}, V_{tot}^{stat}, and V_{tot}^{dyn} .

	ma-ca	ca-p2	p2-nc	nc-p1	p1-p6	rt-rh	rh-in
NA	67	59	61	66	64	63	67
static	crystal structures						
V_{in}	1165	1129	1175	1142	1125	1166	1236
V_{out}	485	342	358	412	390	262	479
V_{tot}	1650	1471	1533	1554	1515	1428	1715
dynamic	MD frames excluding crystal structures						
V_{in}	916±26	834±20	925±16	862±31	886±19	926±17	949±23
V_{out}	522±27	517±25	464±18	603±33	510±21	404±19	571±25
V_{tot}	1438±10	1351±11	1389±10	1465±12	1396±11	1330±12	1520±11

Figure 3.5 shows the distributions of the V_{in}^{dyn}, V_{out}^{dyn}, and V_{tot}^{dyn} values throughout the trajectory for each substrate. The abscissas and ordinates show the volume in $\mathring{A}^3$ and number of occurrence among the 550 frames, respectively. Although rh-in is the largest substrate, its protrusion beyond the dynamic substrate envelope is not significantly higher than the rest of the substrates. V_{out} of nc-p1 is higher than the other substrates most of the time throughout the trajectory, being exposed to the protease the most. This may explain the high mutation rate of nc-p1 in correlation with the protease drug-induced mutations in terms of the substrate envelope hypothesis. On the other hand, rh-in and p2-nc lie within the dynamic substrate envelope in a more robust manner than the rest of the substrates. Based on the substrate envelope hypothesis,

these two substrates are less likely to co-evolve with the protease unless a mutation in the protease causes a dramatic structural/conformational change in the protease-substrate complex and results higher V_{out}^{dyn} values for these two substrates as it is the case for p1-p6 substrate. The wild-type p1-p6 substrate seems not to protrude beyond the envelope to a high extent in a complex with a wild-type protease. Still we know that it co-evolves with D30N/N88D variant of protease. The structural basis for that co-evolution will be elucidated in Section 3.3.

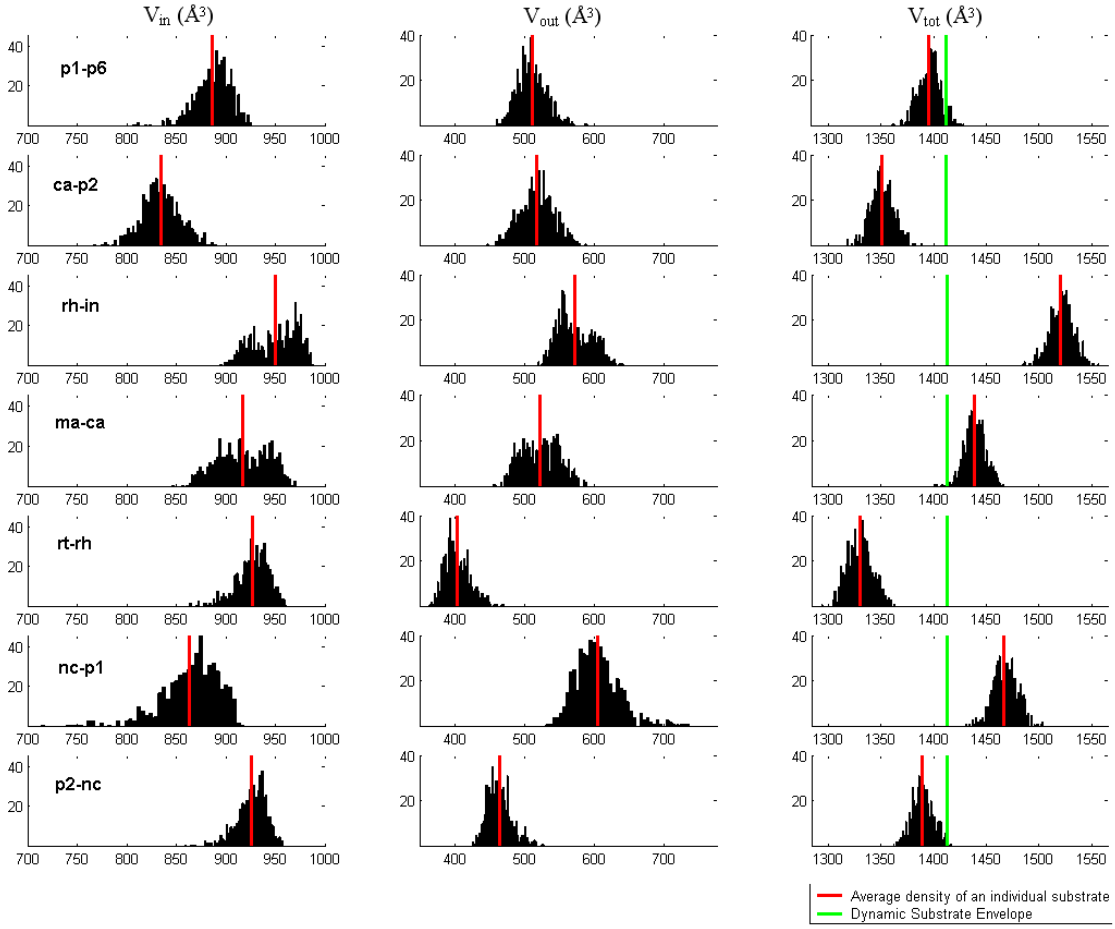


Figure 3.5. Volume distribution of the substrates throughout the trajectories.

The protrusion beyond the substrate envelope may be related to translational dynamics of the substrate lying at the active site of the protease as well as its conformational mobility. In order to see which one has a higher contribution to the protrusion, the correlation of substrate center of mass fluctuations with V_{out}^{stat} and V_{out}^{dyn} are calculated for the WT substrates in complex with the protease. Figure 3.6 shows the correlation coefficients of substrate peptide center of mass mean square fluctuations

and their V_{out} values for static and dynamic substrate envelopes. It is observed that the substrate volume outside of the dynamic substrate envelope is correlated up to a coefficient of 0.67 with the substrate center of mass mean square fluctuations, whereas the volume outside of the static substrate envelope is not (with a correlation coefficient of 0.22).

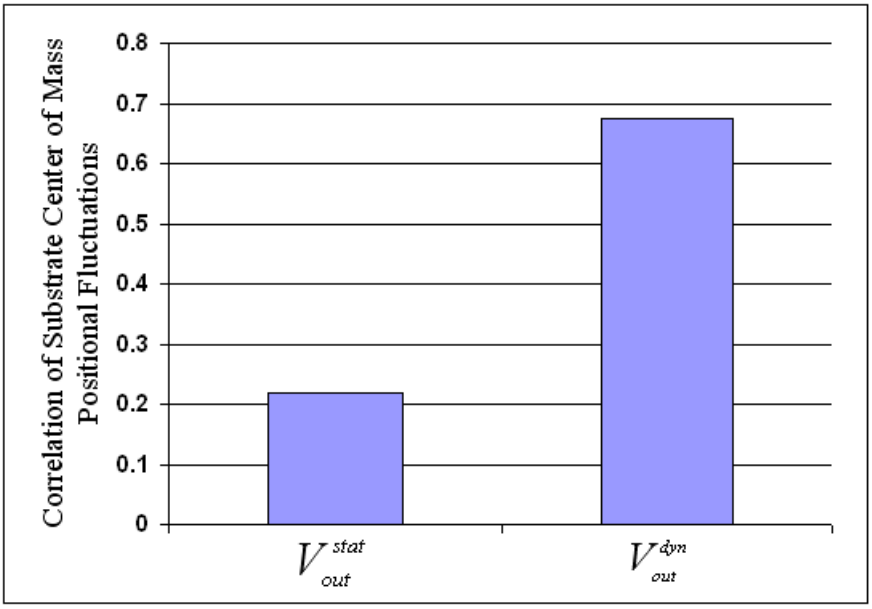


Figure 3.6. The correlation coefficients of substrate center of mass MSF and the V_{out} values for static and dynamic substrate envelopes.

For the contribution of the conformational dynamics of the substrate into the protrusion, the fluctuations of pseudodihedral angles are calculated. Figure 3.7 shows the correlation coefficients of substrate virtual bond (pseudodihedral) angle fluctuations and the substrate volume that lie outside of both the static and dynamic substrate envelopes. The angles around the cleavage site are most positively correlated with V_{out} and the extent of their correlation is almost the same as that of substrate center of mass fluctuation (with a correlation coefficient of 0.62). This may suggest that the transitional and rotational fluctuations may be dependent and a flip in a pseudodihedral angle may cause a change in the substrate center of mass coordinates as well. Therefore, it may be concluded that protrusion beyond the substrate envelope is driven both by transitional and conformational fluctuations.

The correlation of the substrate volume outside of the dynamic substrate envelope

and the substrate pseudodihedral angle fluctuations is far higher than that of the static substrate envelope. This may suggest that the crystal structures may not be enough to define the substrate envelope given that the proteins are not rigid structures and they fluctuate around their native states.

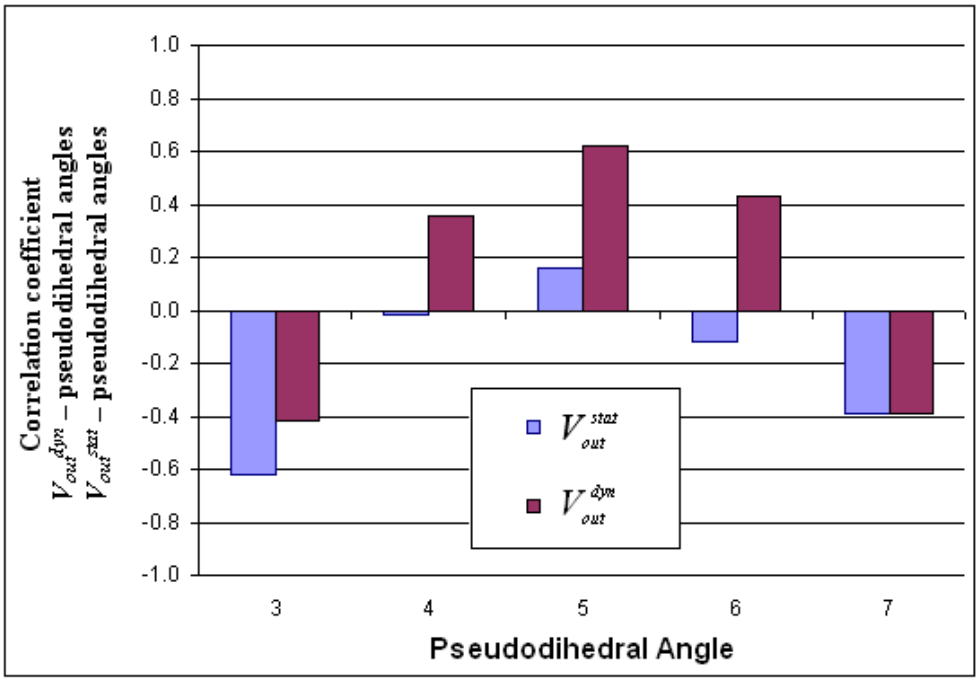


Figure 3.7. The correlation coefficients of substrate pseudodihedral angle fluctuations and V_{out} values for static and dynamic substrate envelopes.

3.2.2. van der Waals Contact Analysis

The P2-P2' pockets in the active site of HIV-1 PR are mainly hydrophobic. Therefore, van der Waals contacts of substrate with the protease play a very important role in substrate binding. Substrate-protease van der Waals contacts are analyzed in the crystal structures for seven cleavage sites. The frames from MD simulation trajectories are also assessed in terms of the substrate-protease van der Waals contacts in order to see if the residue-based contacts which are observed in the crystal structures are preserved throughout the simulations.

Figure 3.8 shows the protease residues having van der Waals contacts with the substrate. The plot gives the residue-by-residue average contacts calculated from a) seven crystal structures and b) 550 frames from simulations of those seven initial struc-

tures. The standard deviation for each protease residue having a contact with the substrate is also given in this figure. Investigation of panel a and b of Figure 3.8 enables assessment of how the van der Waals contacts differ from static to dynamic picture.

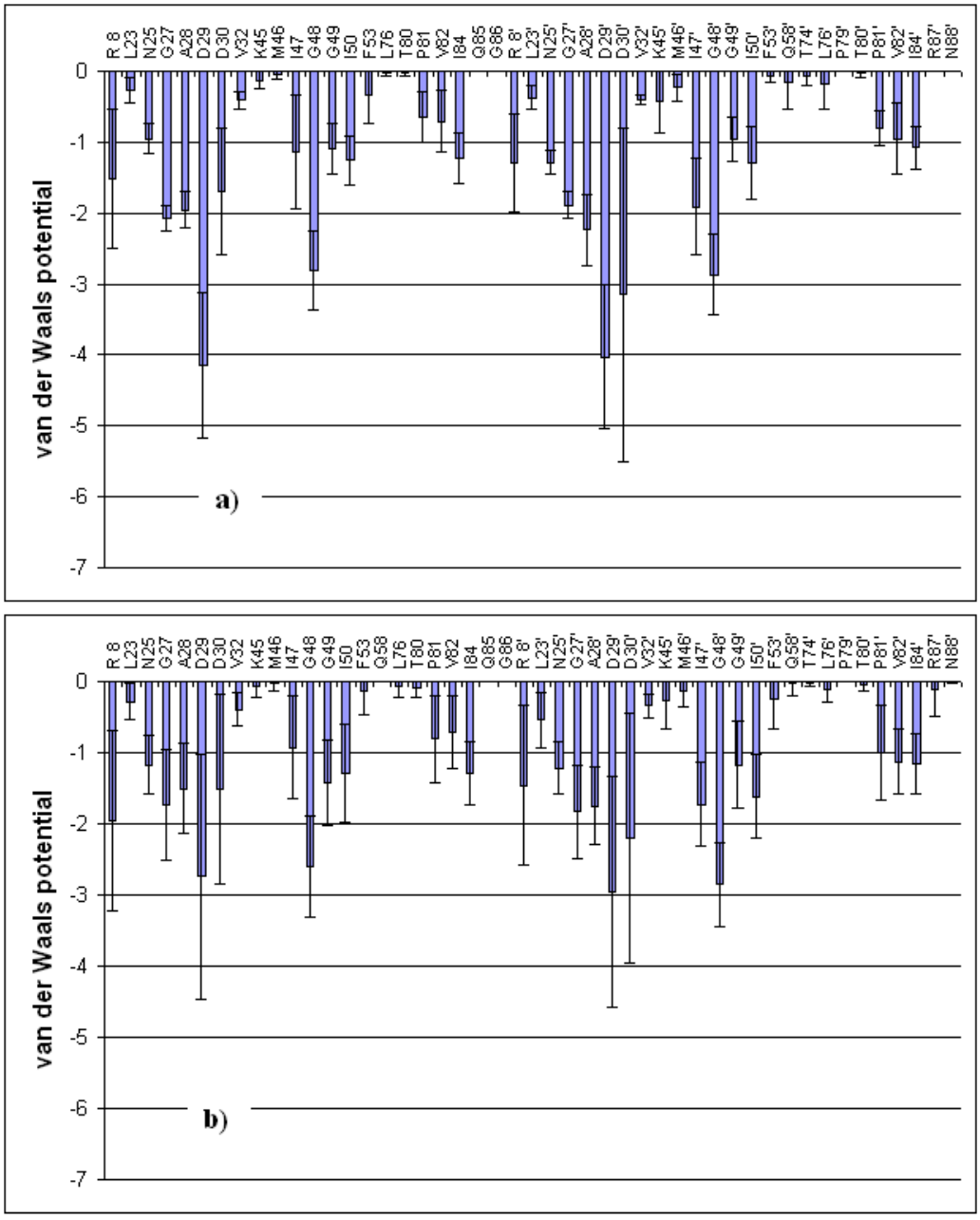


Figure 3.8. Residue-by-residue average van der Waals contacts for a) seven crystal structures and b) MD frames for those seven initial structures.

The residue contacts that show the highest variation among the crystal protease-substrate structures are R8, D29, and D30 in both monomers. The van der Waals contacts for each substrate are also investigated individually. As an example, Figure

3.10 shows the crystal and dynamic van der Waals contacts of p1-p6 substrate with the wild-type protease. The plots of static and dynamic van der Waals contacts of the other substrates are given in Appendix B.

In ma-ca and ca-p2 crystal structures R8 does not seem to have favorable contacts with the substrate but there seems to be an increased and consistent (low standard deviation with respect to mean value) contact in MD frames. In p1-p6 substrate, R8 contacts are almost the same in crystal structure and MD frames whereas R8' contacts increase from crystal structure to MD frames. In p2-nc, R8 has a large number of contacts both in crystal structure and MD frames. The van der Waals contact potentials of R8 and R8' are very low in rt-rh and very high in nc-p1 when compared to the rest of the substrates in both static and dynamic cases. R8-R8' residues in rh-in substrate have moderate contacts in both static and dynamic cases.

In general, D29' and D30' are the protease residues that have the lowest van der Waals potentials with p1-p6 and ca-p2 substrates in the crystal structures with the simplified Lennard-Jones potential function. However, these contacts are lost to a certain extent in ca-p2 complex structure. D30' contacts are still maintained in p1-p6 dynamically whereas D29' contacts have a large standard deviation (the least conserved contact). I84 seems to have the lowest variation in p1p6 frames. D29 is the best contact in both ma-ca and nc-p1 whereas in both the variation of D29' is significantly higher than that of D29. D29 is a peak in crystal p2-nc whereas it is almost lost (having a standard deviation larger than the mean) dynamically. D29' and D30' have favorable contacts. The dynamic and static van der Waals contacts of rt-rh and rh-in are very similar. The van der Waals contacts of the substrate sites are also important to be able to see the residues, which are most likely to be effected by a structural change in the protease. This will also help us to study the co-evolution phenomenon. The conservation of D29-D29' and D30-D30' contacts with the substrate from the static picture to the dynamic one are very important in this regard.

Figure 3.9 shows the van der Waals contacts of the substrate residues individually in the crystal structures and in the MD frames. The absolute values of the potentials

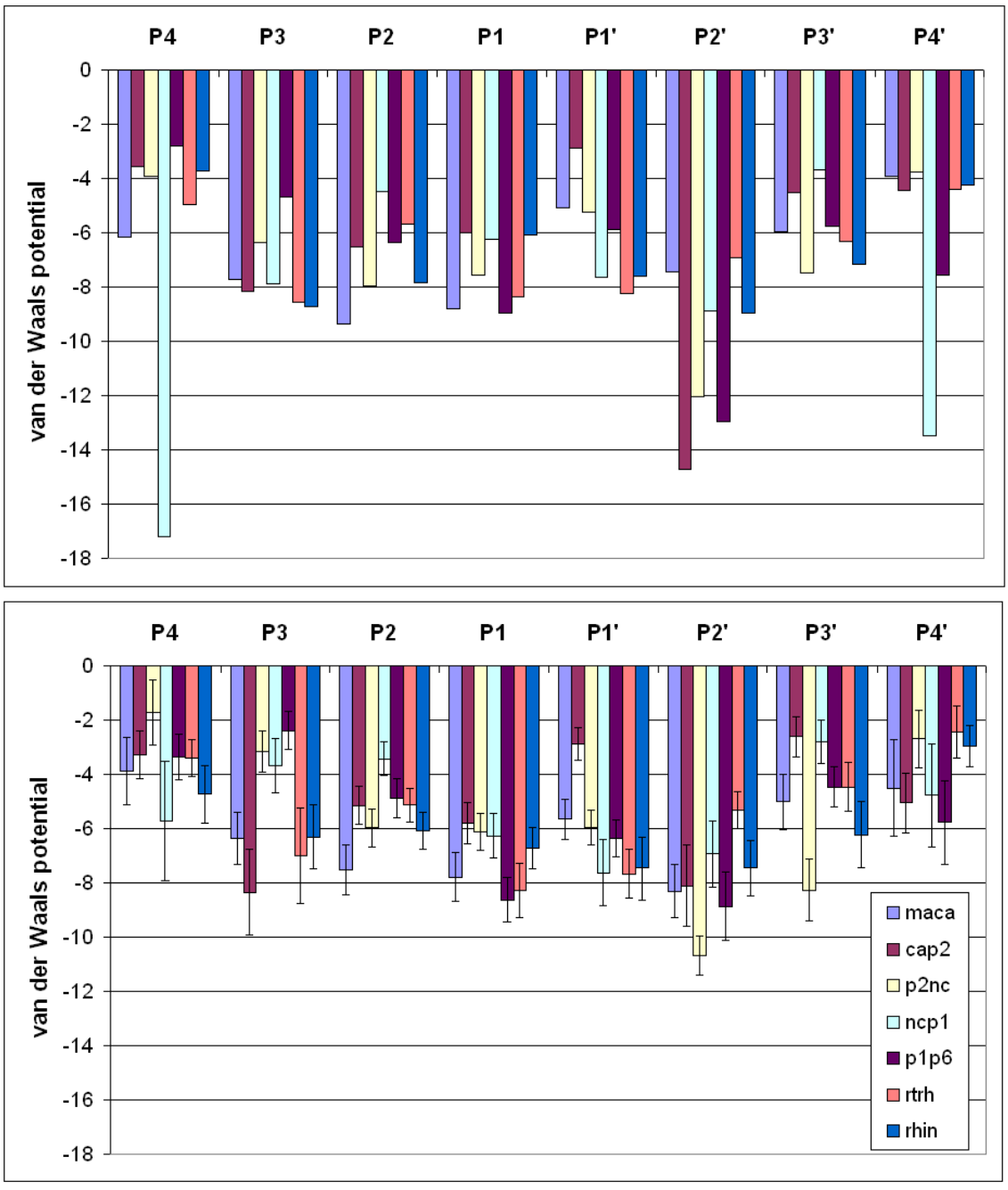


Figure 3.9. The van der Waals contact potentials of substrate sites with the protease in a) the crystal structures and b) MD frames.

are much lower in the MD frames suggesting that the bad contacts, which are not penalized by the simplified Lennard-Jones function, are removed during MD simulations. Especially P4 contacts of nc-p1 is an example for that.

On the other hand, the most conserved contacts among the substrate complexes are mainly around the active site (N25, G27, A28, N25', G27', A28', and V32'), flap region (G48, G49, I47', G48', and I50'), and the substrate cleft (I84, V82', and I84'). This shows that these residues play a role in substrate recognition or binding due their consistent appearance in different substrate complexes. Some of these residues which are located at the active site and the flaps are highly evolutionary conserved but the ones on the substrate cleft are not. Therefore, highly conserved van der Waals contacts of these residues with the substrate does not necessarily require a sequence conservation as well.

The general trend of consensus static van der Waals contacts is very similar to that of consensus dynamic van der Waals contacts. the contacts of the flaps (G48-I50 and G48'-I50') and the substrate cleft (R8, V82-I84 and R82, V82'-I84') with the substrate seem to be conserved from crystal structures to MD frames. On the other hand, the active site (N25-D30 and N25'-D30') contacts (having lower mean and larger variation) tend to decrease dynamically in both monomers. MD simulations after a minimization process may have helped the structures to readjust themselves so that the bad crystal contacts are removed.

3.2.3. Atomic Positional Fluctuations

The atomic positional (mean-square) fluctuations were previously given in Figure 3.3. In this section, the cross-correlations of nonhydrogen atomic fluctuations between all the protease-substrate complex system are reported. The correlations of monomers of the protease, the monomers and the substrate, and within the substrate peptide are analyzed separately.

Figure 3.11 shows the equal-time cross-correlations of protease and substrate

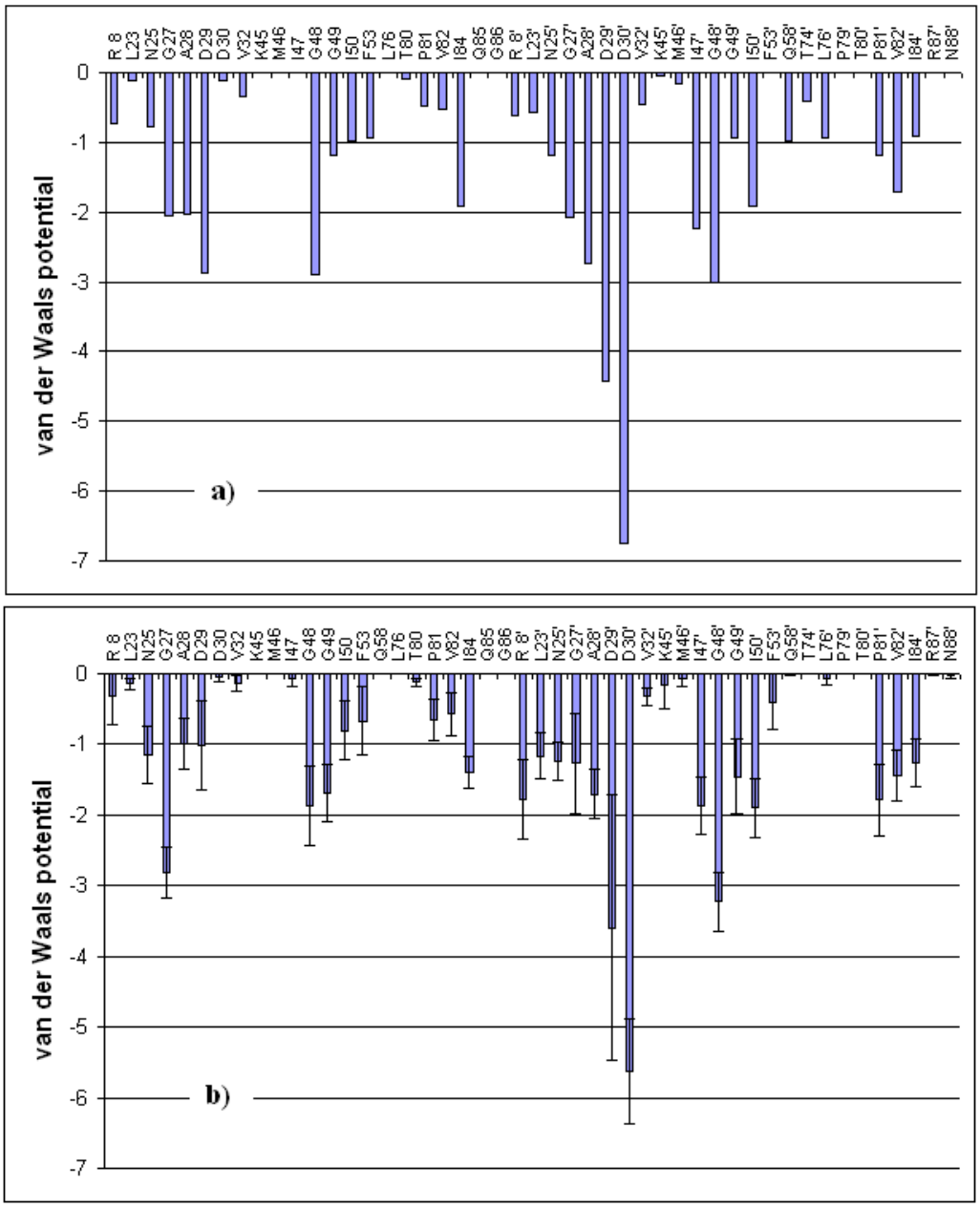


Figure 3.10. a) Crystal and b) Dynamic van der Waals contacts of p1p6 substrate with the wild-type protease.

atomic fluctuations for all nonhydrogen atoms. The protease residues that are highly correlated with any substrate residues and the extent of their correlation are very similar in all seven trajectories. These residues are mainly D30-D30', I50-I50', and V82-V82'. The time-delayed calculations also support this result from the equal-time cross-correlations. The high correlation of atomic fluctuations of D30-D30', I50-I50', and V82-V82' with the substrate residues survive even after 1.1 ns. However, the substrate positions that fluctuate in correlation with these protease positions in particular are different. This may suggest that substrate recognition is an interdependent event and the recognition phenomenon may be different for all natural substrates.

The cross-correlations of substrate nonhydrogen atomic fluctuations with each other also support that (see Figure 3.12-Figure 3.18). The intrinsic dynamics of each substrate may be different from another. The time evolution of with-in-the-substrate cross-correlations are given in Appendix C. The correlations that survive in the course of the simulations are, in general, on the sites P4 and P4' except for rh-in substrate. For this substrate, the auto-correlation of P1' site seems to survive longer times than P4 and P4' sites. In cap2 substrate, P4-P2' and P3'-P4' regions are anti-correlated; they fluctuate almost like two different domains. p2nc is observed to have a shorter memory than the rest of the substrates. This difference may be an influence of the interdependent substrate recognition and the substrates may have different intrinsic dynamics to adopt the required conformations in order not to protrude beyond the envelope.

Figure 3.19 shows the equal-time cross-correlations of atomic fluctuations of the monomers of substrate-bound and unbounded protease structures. The residues that fluctuate with the highest correlation with the other monomer are again D30-D30', I50-I50', and V82-V82'. This suggests that the residues that play an important role in binding are also effective in dimerization.

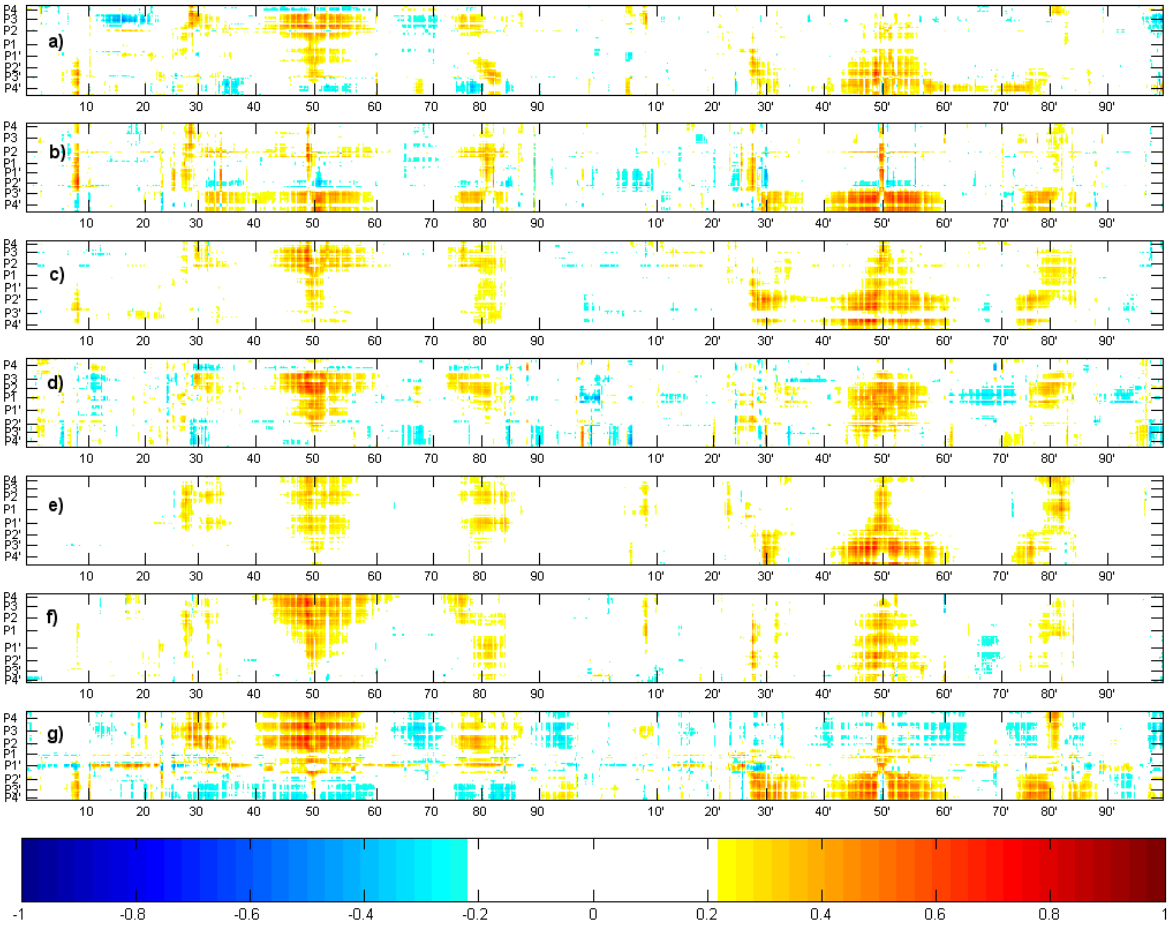


Figure 3.11. Equal-time correlations of protease atomic fluctuations with substrate atomic fluctuations for the substrates a) ma-ca, b) ca-p2, c) p2-nc, d) nc-p1, e) p1-p6, f) rt-rh, and g) rh-in.

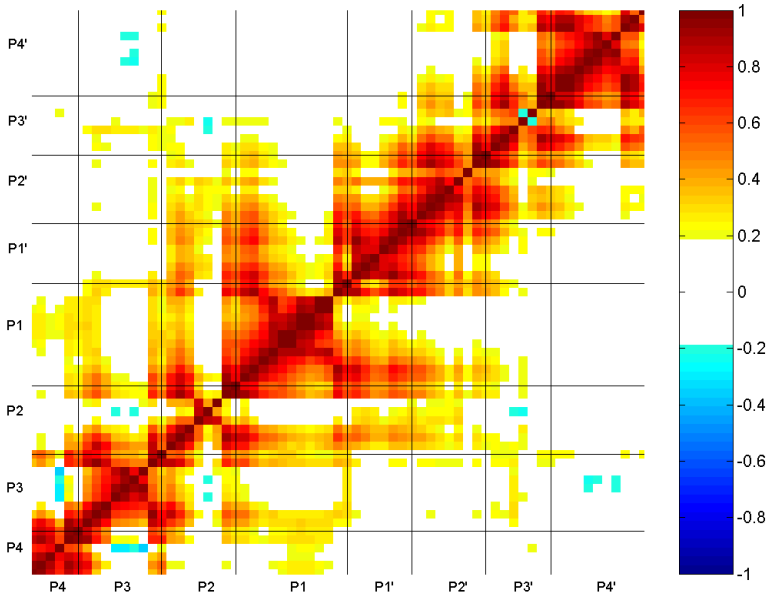


Figure 3.12. Equal-time cross-correlations of ma-ca substrate.

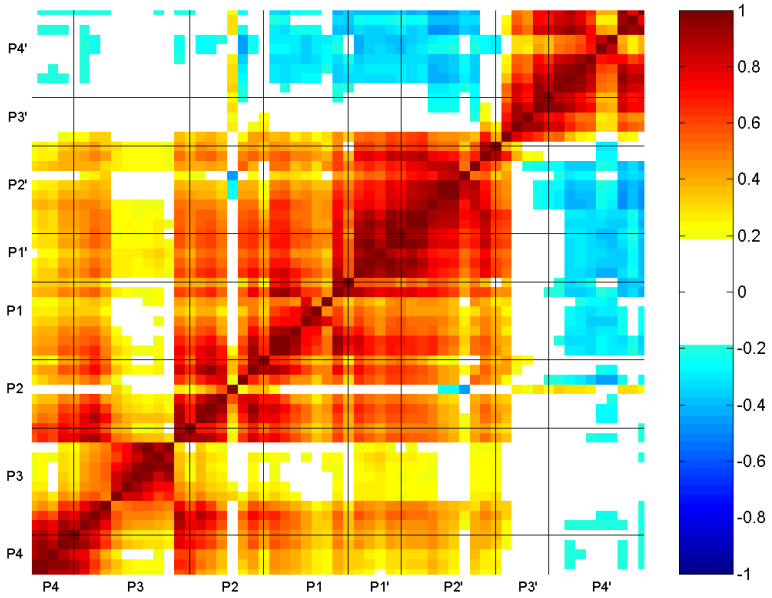


Figure 3.13. Equal-time cross-correlations of ca-p2 substrate.

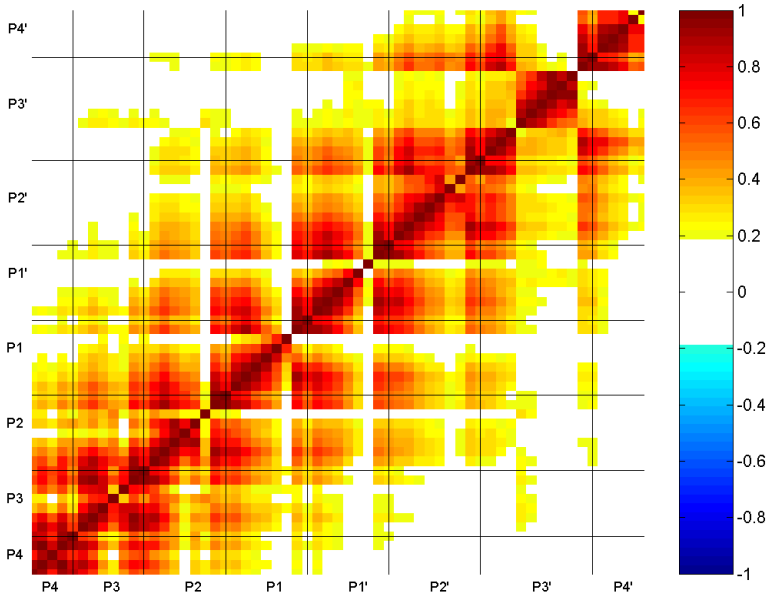


Figure 3.14. Equal-time cross-correlations of p2-nc substrate.

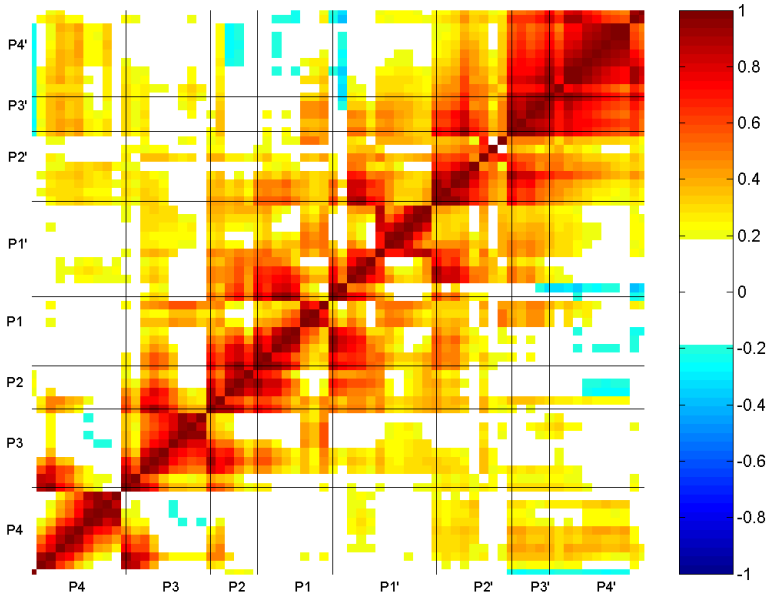


Figure 3.15. Equal-time cross-correlations of nc-p1 substrate.

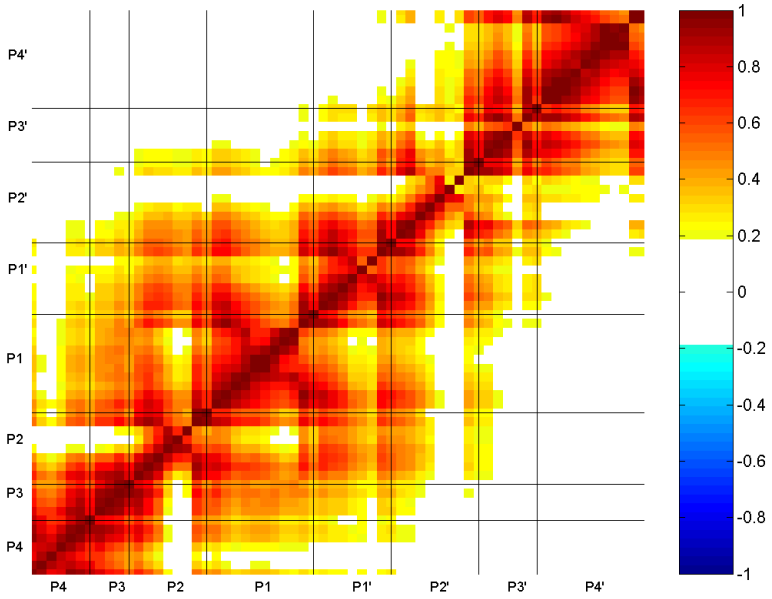


Figure 3.16. Equal-time cross-correlations of p1-p6 substrate.

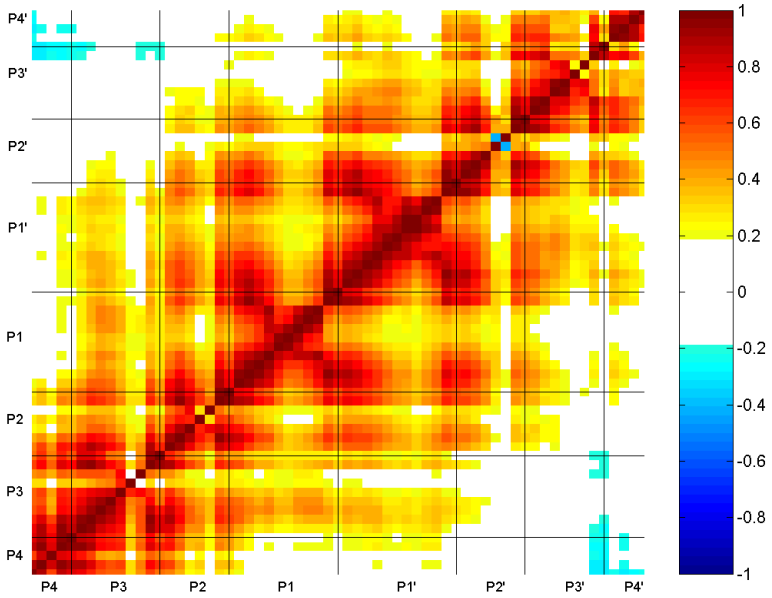


Figure 3.17. Equal-time cross-correlations of rt-rh substrate.

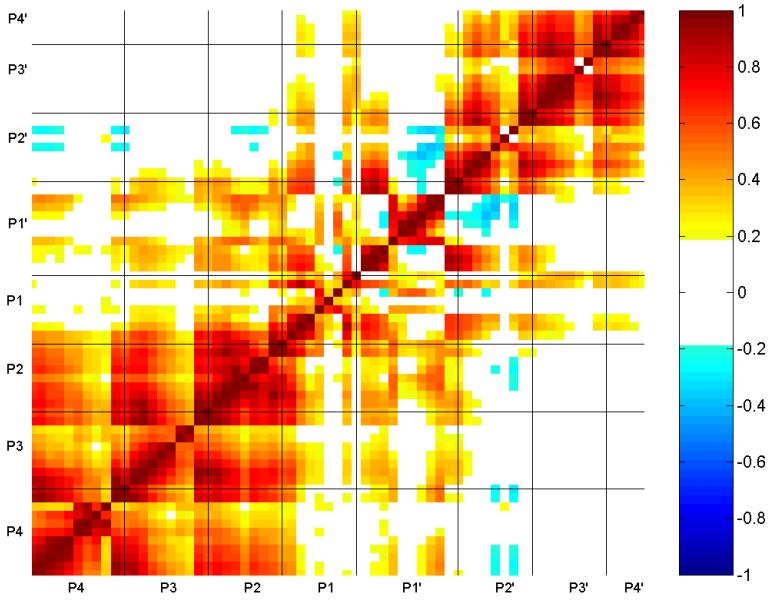


Figure 3.18. Equal-time cross-correlations of rh-in substrate.

3.2.4. Pseudodihedral Angle Fluctuation

Figure 3.20 shows the average $\Delta\phi_i$ values for the protease residues in complex with ma-ca substrate. Very similar behavior is observed in different complex systems. The plots for the other substrates are given in Appendix D. The figures show the fluctuations of the pseudodihedral angles of the protease for a given time of interval $\tau = 20$ ps. The fluctuations are also calculated for longer time scales, yet the general trend of the curve for each complex remains very similar. Therefore, only the average of the difference between the current and the previous angle are plotted together with it standard deviation.

The protease pseudodihedral angle fluctuations are almost the same for the wild-type protease-substrate complexes. The correlation coefficient matrix (see Table 3.5) shows that the least alike pseudodihedral angle behaviors are seen in proteases in complex with ma-ca and ca-p2 with a correlation coefficient of 0.88. This suggests that the the changes in the pseudodihedral angles of the protease residues are not dramatically effected by different substrates in the active site.

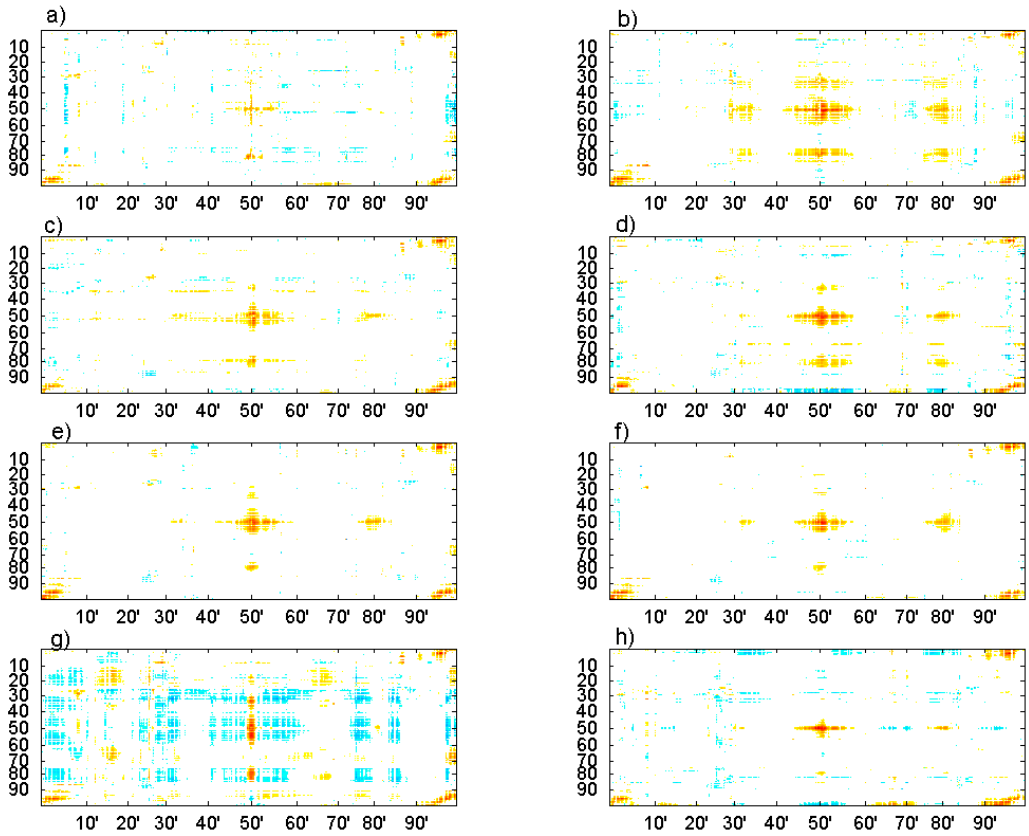


Figure 3.19. Equal-time correlations of atomic fluctuations of monomers of protease in complex with the substrates a) ma-ca, b) ca-p2, c) p2-nc, d) nc-p1, e) p1-p6, f) rt-rh, g) rh-in, h) unbounded protease structure.

Table 3.5. Correlation coefficients of the pseudihedral angle fluctuations of the protease residues in different substrate complexes.

	ca-p2	p2-nc	nc-p1	p1-p6	rt-rh	rh-in
ma-ca	0.88	0.94	0.91	0.94	0.93	0.95
ca-p2		0.93	0.94	0.92	0.92	0.93
p2-nc			0.94	0.95	0.90	0.95
nc-p1				0.96	0.96	0.94
p1-p6					0.96	0.95
rt-rh						0.93

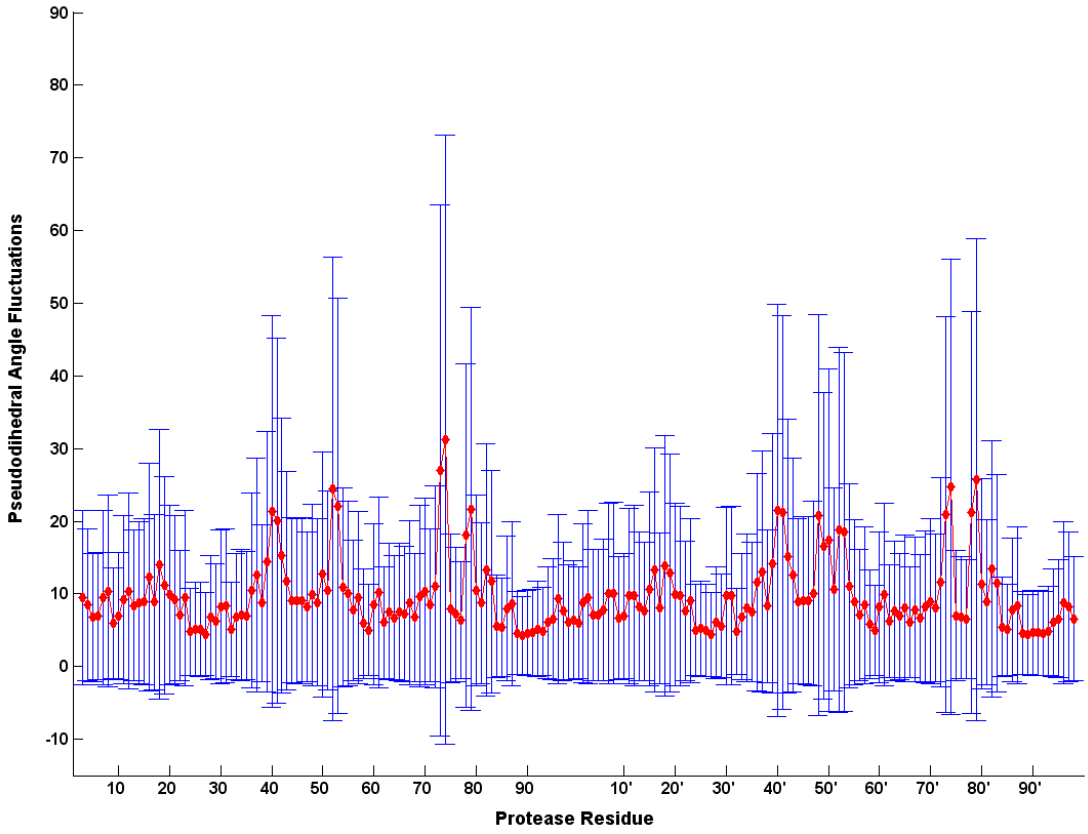


Figure 3.20. $\langle \Delta\phi_i\,(\tau = 20 \text{ ps }) \rangle$ of the protease residues in complex with ma-ca.

On the other hand, the pseudodihedral angle fluctuations in the substrate residues are very different from each other. The correlation coefficient of the P4-P4' sites are given in Table 3.6. The fact that the positions of different substrates have different rotational patterns supports that different substrates may have different intrinsic dynamics. In most of the substrates (ma-ca, ca-p2, nc-p1, rt-rh, and rh-in), it is observed that the most stationary rotation angles are around the virtual bond P2-P1, while in the remaining substrates it is the P1-P1' that is the least flexible. This shows that the substrate tends to stay immobile at the positions it is cleaved. Also the fluctuations of the atomic positions of the substrate residues supports this finding.

3.3. Co-evolution of p1-p6 substrate with D30N/N88D protease variant

Earlier studies (Prabu-Jeyabalan *et al.*, 2002) have shown that substrate specificity of the protease is based on the shape adopted by the substrate sequences, de-

Table 3.6. Correlation coefficients of the pseudihedral angle fluctuations of the P4-P4' residues in different substrates.

	ca-p2	p2-nc	nc-p1	p1-p6	rt-rh	rh-in
ma-ca	-0.35	0.86	0.25	0.89	0.95	-0.06
ca-p2		-0.28	0.75	-0.61	-0.21	0.79
p2-nc			0.41	0.92	0.96	0.11
nc-p1				0.04	0.43	0.90
p1-p6					0.90	-0.22
rt-rh						0.09

fined as the substrate envelope. Most primary active-site mutations occur outside the substrate envelope and thereby preferentially impact inhibitor binding over substrate recognition. Therefore, most of the substrates do not co-evolve with the protease. However, some substrates protrude beyond the envelope, and it was reported that they are the ones which co-evolve with the protease.

D30N and N88D mutations, which is a signature of NFV resistance and the dynamic and structural basis for LP1'F mutation on the p1-p6 substrate to co-occur with those mutations are focused in this part.

Figure 3.21 shows the RMS deviations of protease structures in complex with p1-p6 substrate with respect to the corresponding initial conformation. Each curve has a smooth trend without major fluctuations, suggesting that the protease does not undergo major conformational changes in the simulations. The fact that D30N/N88D protease variant has a shifted RMSD curve above the others and it has still a smooth trend may suggest that this structure is the most likely to adopt a different conformation from the initial structure which was created from WT protease-substrate complex by *in silico* mutagenesis.

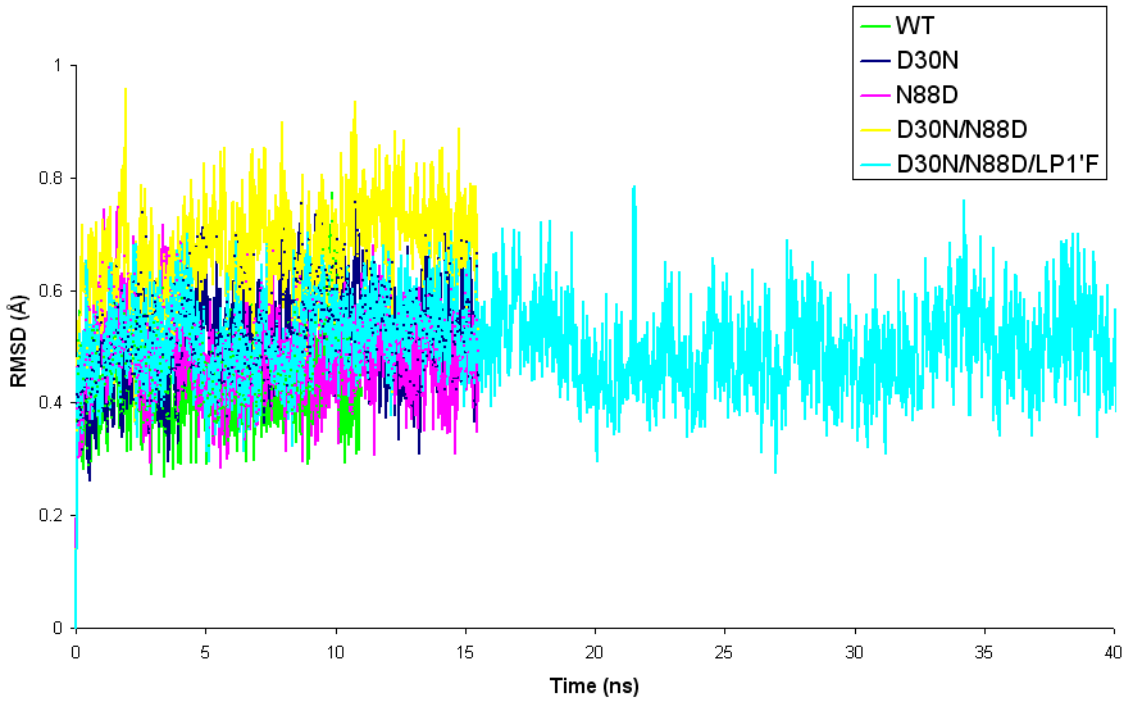


Figure 3.21. RMSD of variants of protease-p1-p6 substrate complex structures throughout the simulations.

3.3.1. Overall Fit within the Dynamic Substrate Envelope

The protease variants D30N, N88D, and D30N/N88D in complex with wild-type p1-p6 substrate and D30N/N88D variant in complex with LP1'F p1-p6 substrate variant are evaluated based on the dynamic substrate envelope hypothesis. Figure 3.22 shows the substrate volumes that a) stay within the dynamic substrate envelope, V_{in}^{dyn}, b) lies outside of the envelope, V_{out}^{dyn}, and c) the total substrate volumes, V_{tot}^{dyn}. From Panel c it is observed that the LP1'F mutation apparently increases the total substrate volume by substituting Leu to bulkier Phe. However, the p1-p6 substrate volume that lies outside of the dynamic substrate envelope, V_{out}^{dyn}, shows that the substitution of Leu to Phe helps the substrate to adopt a shape protruding beyond the envelope to a level which is very similar to that of wild-type substrate in complex with wild-type protease.

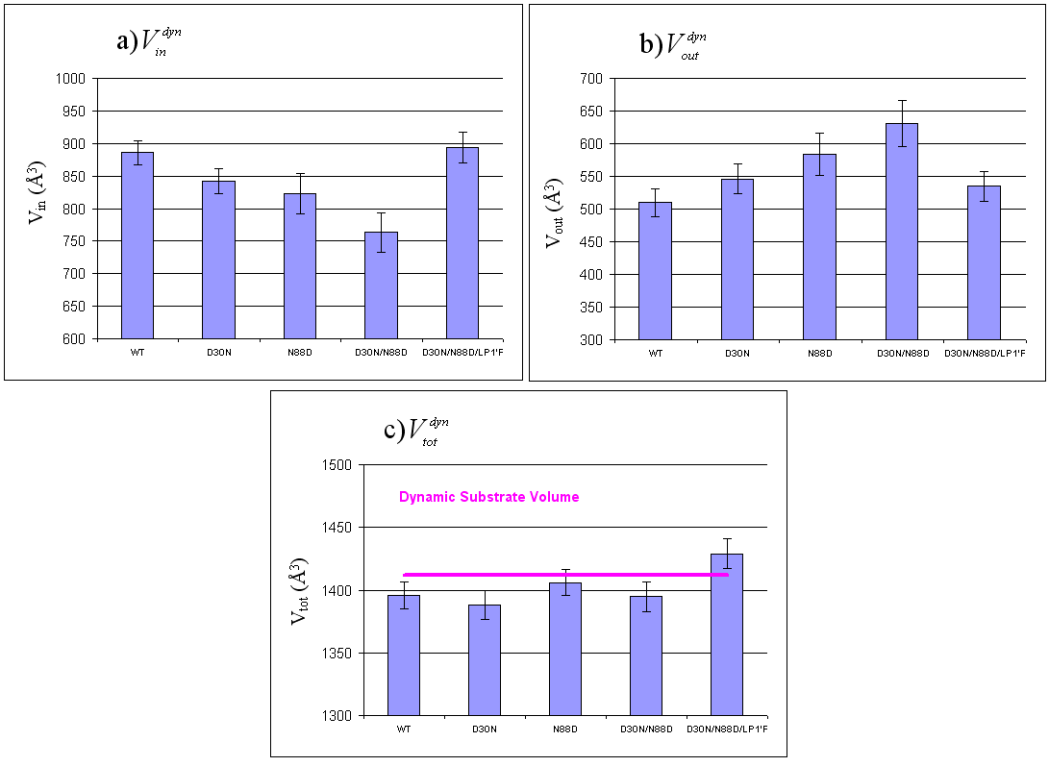


Figure 3.22. For variants of protease-p1-p6 complexes a)V_{in}^{dyn} b) V_{out}^{dyn} c)V_{tot}^{dyn}.

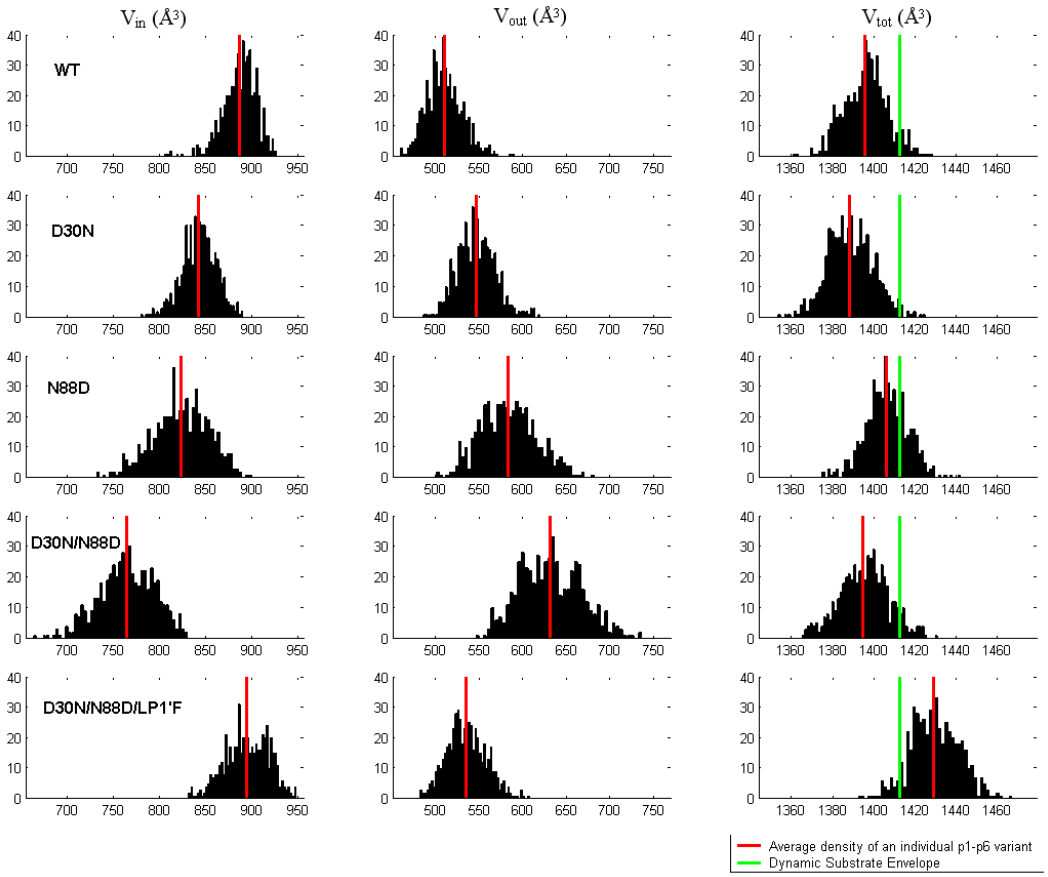
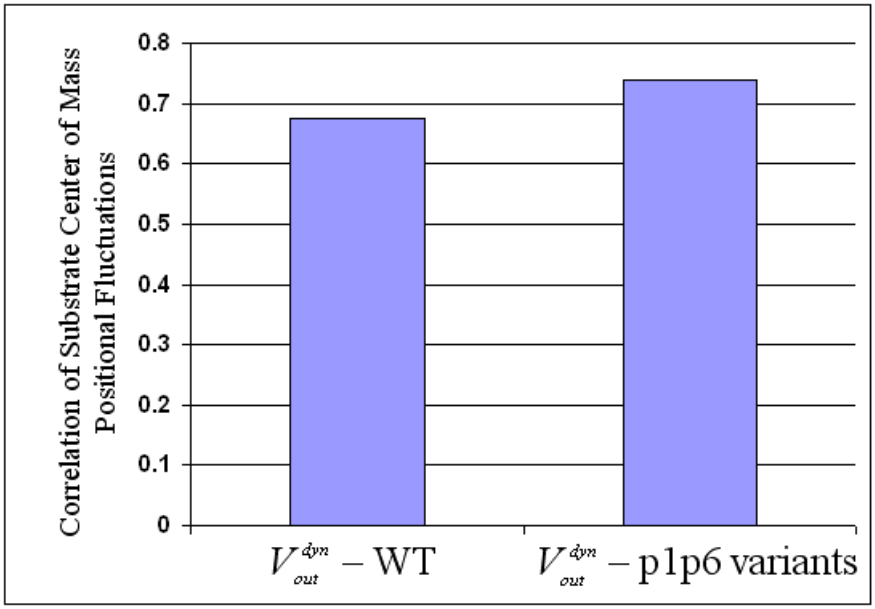


Figure 3.23. Volume distribution of p1-p6 substrates throughout trajectories.



3.3.2. Fit within the Dynamic Envelope by Residue

Figure 3.25 shows the volume of each p1-p6 substrate position which lies outside of the substrate envelope individually. The positions P4 and P4' are the most protruding residues as they are the farthest residues from the active site. They also have less van der Waals contacts than the other residues in the substrates (see Figure 3.27). Therefore, these two residues are less restricted than the rest of the peptide and likely to protrude beyond the envelope.

The P1 position is the most protruding residue among the remaining ones except P4 and P4'. P2' has extensive van der Waals contact potential (Figure 3.27), yet it unexpectedly stays within the substrate envelope in all variants. D30N/N88D mutations apparently changes the scaffold of the protease structure so that P2 position tends to have a high potential to protrude beyond the envelope. It also has extensive van der Waals contacts in the D30N/N88D variant. Also the overall protrusion of the WT substrate in this variant is highest among the others. It is observed in Figure 3.25 that the replacement of Leu at P1' to a bulkier Phe does not effect the extent of the protrusion of this position. It even decreases the total substrate volume which sticks out of the envelope (see Figure 3.22).

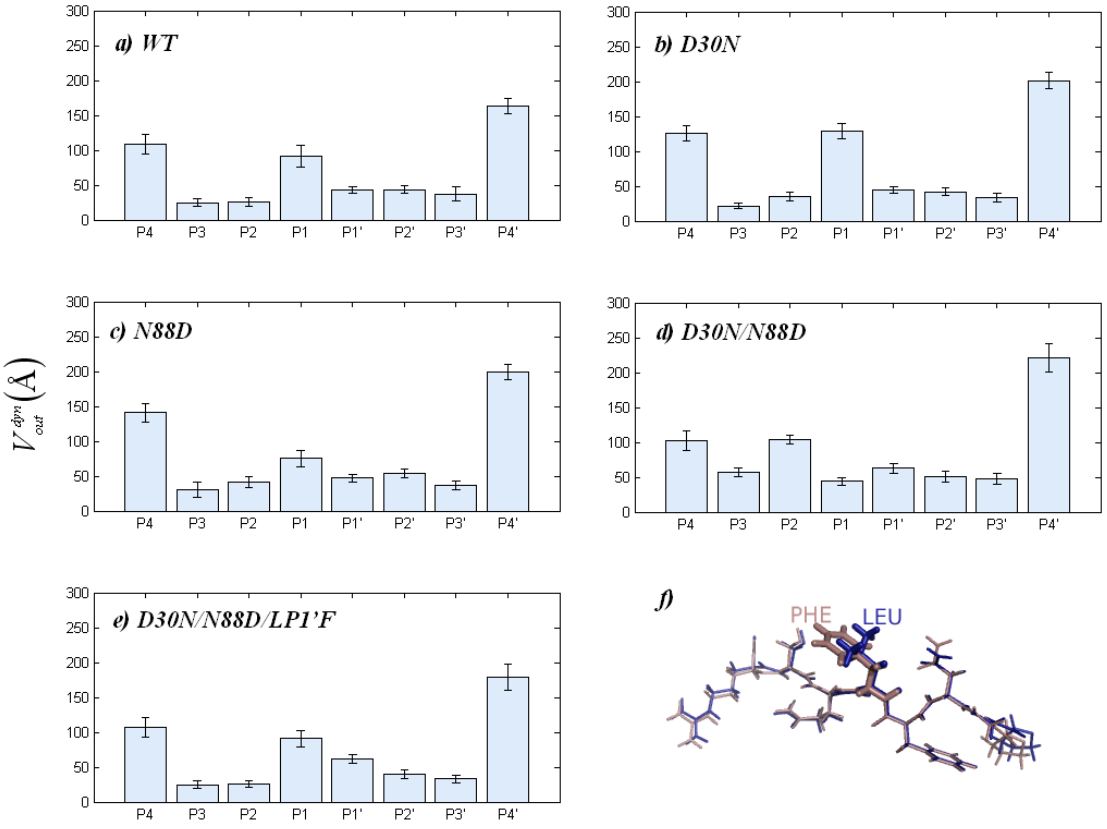


Figure 3.25. V_{out}^{dyn} values for individual substrate positions in protease-p1-p6 complex variants; a) WT PR - WT p1-p6 b) D30N PR - WT p1-p6 c) N88D PR - WT p1-p6 d) D30N/N88D PR - WT p1-p6 e) D30N/N88D PR - LP1'F p1-p6.

3.3.3. van der Waals Contact Analysis

It was previously reported (Kolli *et al.*, 2006) that as a result of modeling studies from the crystal structure of the WT protease bound to the p1-p6 substrate, Phe at P1' fits the active site much better than Leu at P1' (WT) due to more extensive van der Waals contacts. It was also reported that the van der Waals contacts of N30' that are lost in D30N/N88D variant are regained with a substitution of Leu with Phe.

After the wt protease-substrate complex, the most favorable contacts in the active site occur in N88D variant among the protease variants. This may suggest that N88D is as efficient as the wt protease in recognizing and cleaving p1-p6 wt substrate. Our van der Waals contacts analysis, showing the wt and N88D proteases have the same contact pattern with the wt substrate, is in agreement with the experimental catalytic efficiency findings (Kozisek *et al.*, 2007).

The inhibition constants reported by Kozisek *et al.* (2007) supports our finding that among the protease variants harboring only one of the single mutations D30N, N88D, L90M, and A71V, the most significant effect on binding of NFV was observed for the D30N mutant and the mutation N88D alone produced no important change in K_i for any of the inhibitors tested in that study by Kozisek *et al.* (2007).

The difference between the V_{out}^{dyn} values of WT and D30N proteases in complex with WT substrate may not be considered as significant. This may bring the question of the sufficiency of D30N single mutation, but D30N mutation decreases the van der Waals contacts of wt substrate with the protease suggesting a less tight fit into the active site. Recently Kozisek *et al.* (2007) reported that the D30N mutant protease showed a two-fold decrease in catalytic efficiency (k_{cat}/K_m) compared to wt protease using a chromogenic peptide substrate derived from the ca-p2 cleavage site.

D30N/N88D mutations increasing the substrate volume outside of the envelope contradictorily decreases the van der Waals contacts of the wt substrate with the active site (see Figure 3.26). The outcome of this lost in overall van der Waals contacts

and protrusion beyond the substrate envelope to a high extent results an unfavorable complex structure, which likely leads to a mutation co-occuring in the substrate.

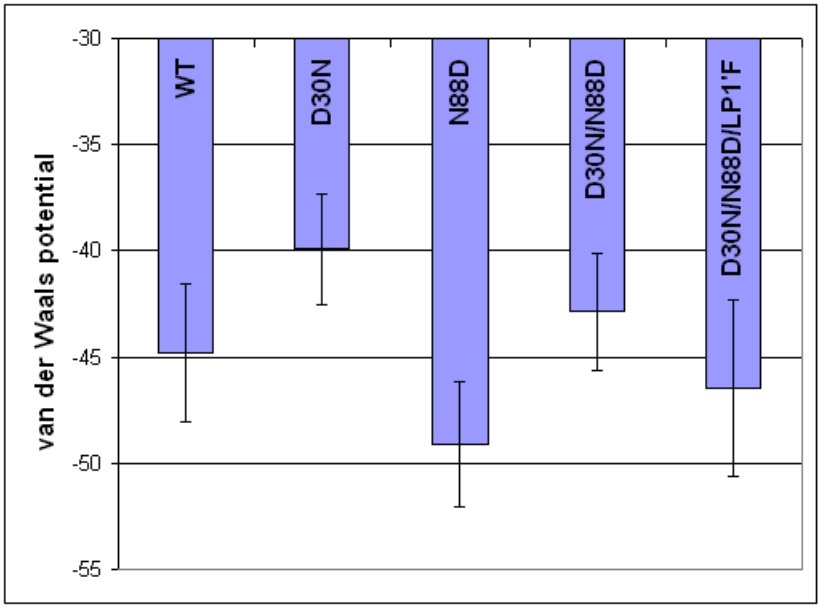


Figure 3.26. Total van der Waals contact potential of the substrate in complex with protease variants.

3.3.4. Atomic Positional Fluctuations

The cross-correlations of the atomic positional fluctuations of substrate and protease residues are given in Figure 3.28. In the protease variant D30N/N88D in complex with the LP1'F variant of the substrate sites, the correlations are far more obvious than the other variant complex structures. This is mainly due to the bulkier Phe in the P1' position and increased number of contacts in this position. The general correlation trend of the variant complex structures are similar to those of the WT complexes. D30N protease variant in complex with WT p1-p6 substrate has high negative correlations between the unprimed site and substrate cleft residues around L10. Similar negative correlations are also observed in between the primed site of the substrate and R8 of the protease. The co-occuring of LP1'F mutation with the double mutation D30N/N88D in the protease removes this anti-correlation, resulting a contour plot similar to that of the WT protease-substrate complex.

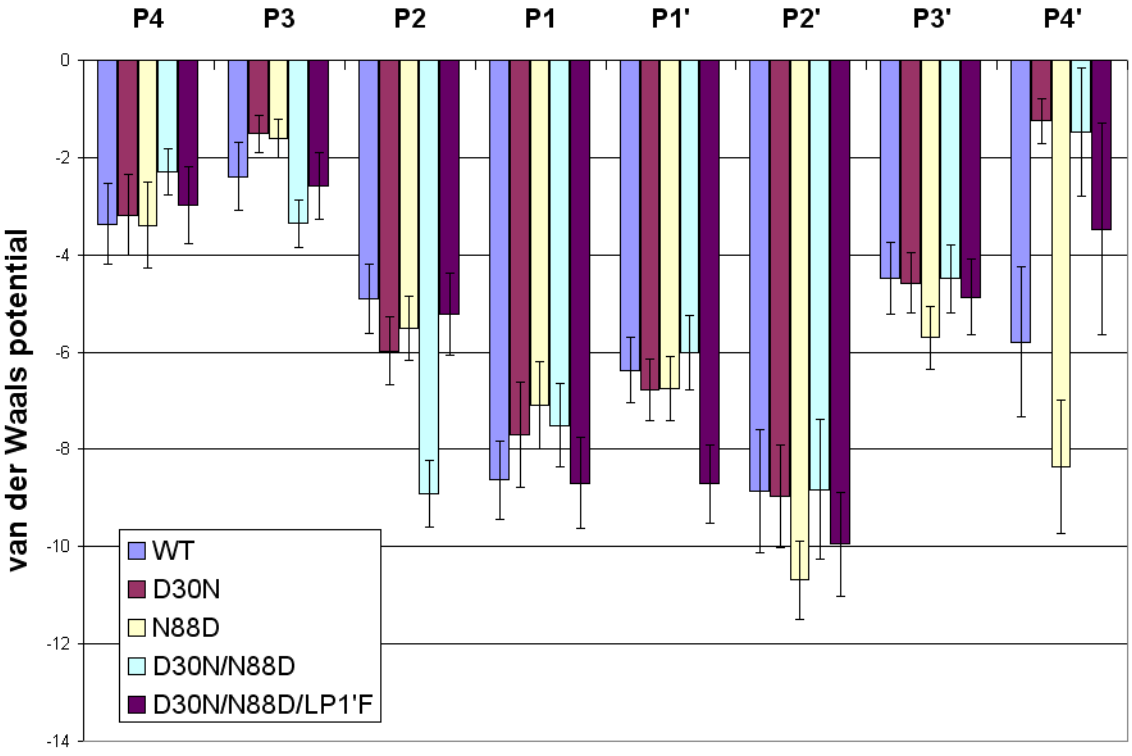


Figure 3.27. Residue-by-residue van der Waals contact potential of the substrate in complex with protease variants.

Figures 3.29 and 3.30 shows the atomic positional fluctuations of the protease residues in complex with p1-p6 substrate with the coordinate sets recorded at every 0.4 ps and 20 ps, respectively. As seen in these figures using 550 frames is representative. The general trends of the curves for different protease-substrate complex variants in Figure 3.29 are similar. The residues of the second monomer in the active site region (K20'-D30') are restrained most likely due to high packing in the active site. However, the corresponding residues in the second monomer (K20-D30) are far more mobile in D30N/N88D variant. The active site may be considered as the most important region of the substrate because the cleavage process takes place in this part. Therefore, deviation from the WT complex dynamic behavior in such an important region is likely one of the reasons of the need for a third compensating mutation in the complex, which is, in this case, LP1'F. In the co-evolved variant, the mobility in this region is suppressed by the substitution of Leu by Phe retaining the complex to preserve its dynamic behavior.

3.3.5. Pseudodihedral Angles

The pseudodihedral angles fluctuations of the protease residues were averaged over the trajectories and plotted with the corresponding standard deviation values (see Figures D.1-D.6). The general trends of these plots are similar. In order to understand, which complexes are more alike in terms of the pseudodihedral angle fluctuations, the correlation coefficients of the whole set of angles averaged over the trajectory were calculated for five different protease-p1-p6 complexes. This calculation was performed for the protease angles and substrate angles separately. Table 3.7 shows the correlation of protease pseudodihedral angle fluctuations of the five complexes (WT, D30N, N88D, D30N/N88D, D30N/N88D/LP1'F). D30N and D30N/N88D/LP1'F variants have the highest correlation with the WT protease-substrate complex in protease pseudodihedral angle fluctuations. The correlation coefficients of the substrate angles in the variant structures also show that the pseudodihedrals of WT and the D30N/N88D/LP1'F variant have the most alike general trend among the other complexes. Similarity between the WT and the D30N/N88D/LP1'F mutant is observed at all angles but since the angles around P2-P1 and P1-P1' are very restricted, the contribution of the angles of

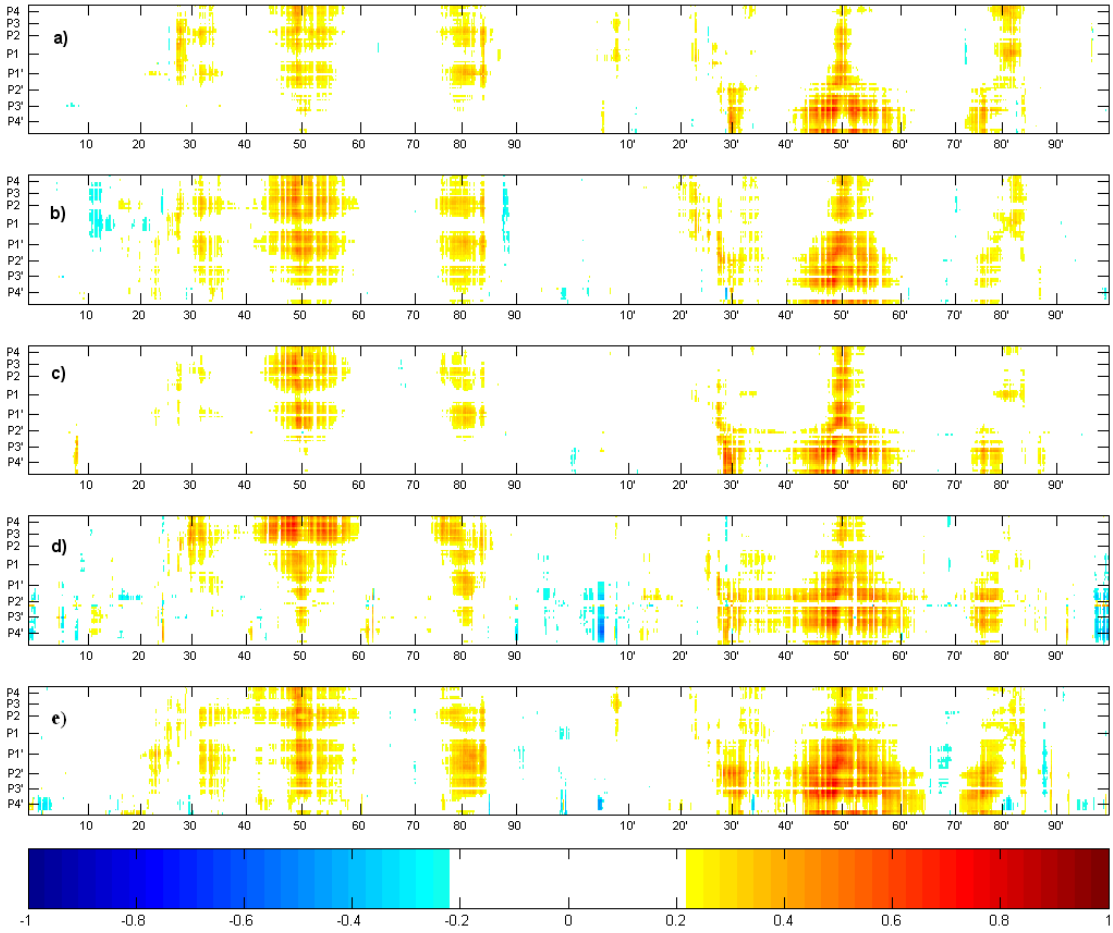


Figure 3.28. Equal-time correlations of protease atomic fluctuations with substrate atomic fluctuations for the substrates a) WT, b) D30N, c) N88D, d) D30N/N88D, e) D30N/N88D/LP1'F.

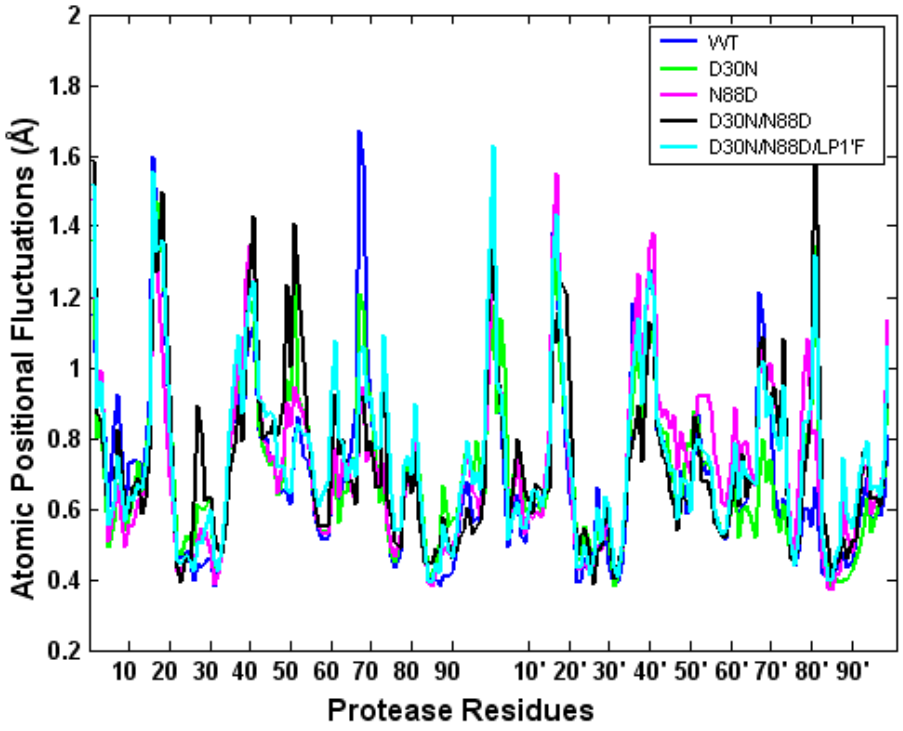


Figure 3.29. Atomic positional fluctuations of the protease residues in protease-p1-p6 complex variants (Data recorded at every 0.4 ps.).

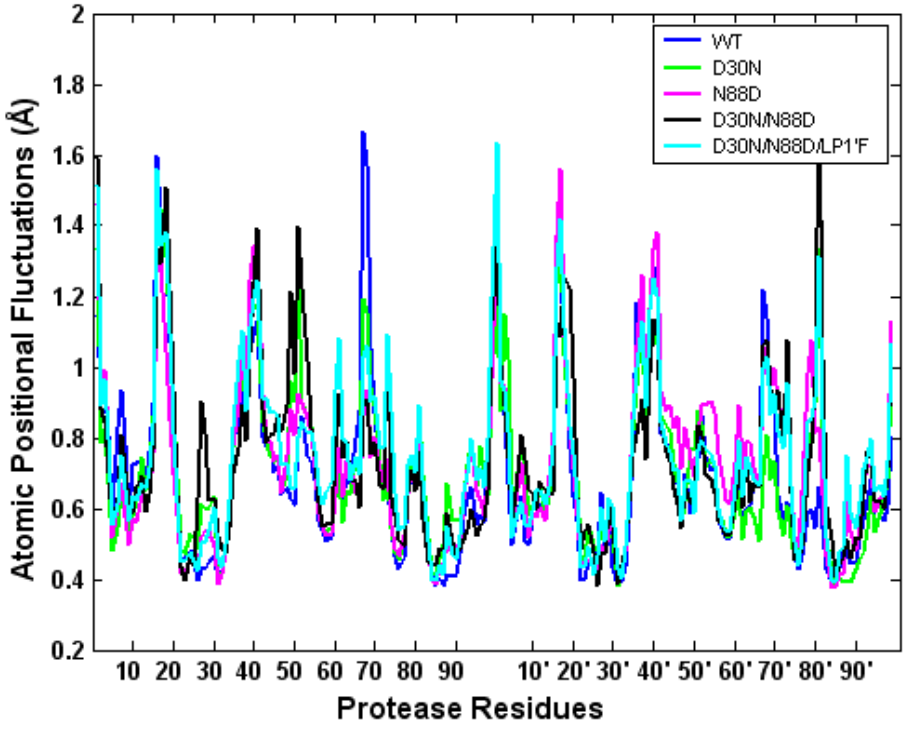


Figure 3.30. Atomic positional fluctuations of the protease residues in protease-p1-p6 complex variants (Data recorded at every 20 ps.).

the other residues, namely, P4-P3-P2-P1 and P1'-P2'-P3'-P4' are higher.

The correlation coefficients of the pseudodihedral angles of the substrate residues in various p1p6 complexes are given in Table 3.8. The highes correlation is seen between WT substrate bound to the WT protease and LP1'F p1-p6 substrate bound to D30N/N88D protease variant. This elucidates the necessity for LP1'F for the D30N/N88D double mutant. In this table, it can also be said that N88D compensates the alteration in the conformational dynamics of the substrate in D30N mutant. However, the double mutant does not resemble the WT complex as the D30N/N88D/LP1'F mutant does.

Table 3.7. Correlation coefficients of the pseudodihedral angles of the protease residues in protease-p1-p6 variant structures.

	D30N	N88D	D30N/N88D	D30N/N88D/LP1'F
WT	0.97	0.94	0.92	0.97
D30N		0.95	0.91	0.96
N88D			0.92	0.97
D30N/N88D				0.93

Table 3.8. Correlation coefficients of the pseudodihedral angles of the substrate residues in protease-p1-p6 variant structures.

	D30N	N88D	D30N/N88D	D30N/N88D/LP1'F
WT	0.83	0.71	0.96	1
D30N		0.94	0.92	0.79
N88D			0.82	0.68
D30N/N88D				0.94

4. CONCLUSIONS AND FUTURE STUDIES

4.1. Conclusions

In this thesis, the substrate recognition of HIV-1 protease and the co-evolution of the p1-p6 substrate with the enzyme have been studied by molecular dynamics simulations. Three main analyses have been conducted; (1) the fit of the individual peptide densities into the remarkably uniform region that the substrates occupy in the binding site, which is termed as the substrate envelope, (2) the conservation of the van der Waals contacts of the substrates with the protease residues dynamically, and (3) the fluctuation dynamics of the atomic positions and the pseudodihedral angles.

The re-modeling of the substrate envelope, which was previously defined based on the crystal structures, by taking the protein dynamics into consideration revealed the potential variance of the consensus substrate conformational preference. This is likely an effect of that the crystal structure conformations are not sufficient to span the possible conformations that the substrates can adopt. The importance of this finding arises from the fact that the dynamic substrate envelope occupies a smaller region than the static one. This is a significant point that should be addressed especially while designing novel drugs aimed to stay within the substrate envelope and not to induce mutations in the protease.

The calculation of dynamic substrate envelope, defined by multiple conformations, enables assessment of the significance of the potential difference between the substrates. The results show that protrusions of some substrates (nc-p1 and rt-rh) beyond the envelope are significantly different. However, this finding, on its own, is not enough to conclude about the potential co-evolution of a cleavage site with the protease. Because, although nc-p1 is a co-evolving and the most protruding site beyond the envelope, p1-p6 is another co-evolving site and it does not protrude significantly more than the other sites. The investigations of the p1-p6 variant structures elucidate this contradiction. In this study, it has been observed that the mutations in the protease

cause the p1-p6 peptide to protrude more than the wild-type peptide. That is likely the driving force for a compensating third mutation in the peptide, which has been found to help the protease-substrate complex to conserve the interaction preferences of the substrate.

Assessment of the van der Waals contacts between the WT protease and the WT substrates shows that there may be differences in the potential contacts of the substrates with the protease although, in general, they prefer to have contacts with mainly the active site, flap region, and the substrate cleft in common. This may imply that substrate recognition and binding may not be the same for all the cleavage sites. It has been also observed in the protease-p1-p6 variant structures that the van der Waals contacts that are lost by D30N/N88D double mutant are restored by the LP1'F substitution in the substrate. Being able to see the effects of the co-evolution in the protease-substrate van der Waals contacts highlights the importance of these contacts in substrate fit into the active site.

Analysis of the equal-time and time-delayed correlations of the atomic positional fluctuations shows that the protease residues that are highly correlated with any sb-starte residues (D30, I50, and V82) and the extent of their correlation are very similar in all WT complex trajectories. However, the substrate residues that they are correlated with seem to be exhibiting variety, suggesting that the substrate recognition is an interdependent event. The recognition mechanism may not be the same for all natural substrates. The fluctuation dynamics of the protease-p1-p6 complex variants has also been observed to be restored by the LP1'F mutation in the cleavage site after D30N/N88D double mutation in the protease.

In summary of this thesis, the substrate recognition may be an interdependent event and the recognition mechanism may not be the same for all natural substrates. The substrate envelope may actually be smaller than the crystal structures impose and this is crucial for drug design efforts with the substrate envelope hypothesis. Finally, the differences in the protrusions beyond the substrate envelope in the mutant and WT structures indicate that the substrate recognition is altered when there is drug

resistance. The better fit of LP1'F p1-p6 into the envelope may suggest that this alteration is compensated by co-evolution. The structural and dynamic conservation appears to be important (if not essential) for protease-substrate recognition.

4.2. Future Studies

D30N was previously reported to be an important residue both to the binding of NFV and also likely to the recognition of the p1-p6 cleavage site in a study by Kolli *et al.* (2006). In this study, analysis of the crystal structures have shown that both NFV and p1-p6 substrate have atoms that protrude beyond the substrate envelope and contacts D30, thus both the inhibitor and the p1-p6 substrate are likely to be affected by D30N mutation. This was reported to likely explain the particular co-evolution of the p1-p6 site with the D30N-resistant mutation and also why no other co-evolution with any of the other substrates occurs. In this study, the structural dynamic basis of the co-evolution of P1' position of the p1-p6 substrate with the D30N/N88D protease variant was explored. However, the structural dynamic basis of the D30N mutation to occur in the first place still needs to be elucidated. For that purpose, MD simulations of NFV in complex with protease variants D30N, N88D, and D30N/N88D may be set up. Analysis of these trajectories based on the substrate envelope hypothesis may help to elucidate the structural and dynamic driving forces for the D30N mutation to be induced by NFV binding.

The first drug-resistant mutation that often occurs in patients, V82A, is frequently associated with IDV or RTV therapy. In several studies of drug-experienced patient viral sequences (Prabu-Jeyabalan *et al.*, 2004) nc-p1 has been observed to mutate. The most frequently observed change occurs at P2, where the ALA mutates to a VAL in viral sequences that also contain the V82A drug-resistant protease mutation. Strikingly, it was reported that the WT HIV-1 PR cleaves AP2V in the nc-p1 cleavage site more efficiently than it cleaves the WT nc-p1 sequence (Feher *et al.*, 2002). However, despite the enzymatic advantage, the AP2V mutation is not observed in sequences of the WT virus, implying that a more efficient and presumably premature cleavage of this site may be detrimental to viral maturation. The crystal structure of

the V82A/AP2V protease-nc-p1 variant has been solved at a resolution of 2.0 Å by Prabu-Jeyabalan *et al.* (2004) and deposited to the Protein Data Bank with the accession code of 1TSQ. MD simulations of this structure together with *in silico* created V82A protease mutant in complex with WT nc-p1 substrate and AP2V nc-p1 mutant in complex with WT protease would be very helpful to further study the co-evolution phenomenon taking the dynamics of the protein structure into consideration.

The crystal structures of six substrate peptides in complex with the WT protease were solved by Prabu-Jeyabalan *et al.* (2002). The structural analysis of these complexes revealed that five waters are completely conserved between the five of the structures that were determined to 2.0 Å. These conserved waters include three that directly contact the peptides and two that stabilize them. Overall, these five waters stabilize the extended conformation of the peptides by hydrogen bonding to the carbonyl oxygens of the peptide to which the protease in the absence of water could not directly hydrogen bond. Several additional waters surround each of the peptides but do not make direct hydrogen bonds. The major role of these waters were said to be to occupy space. This cluster of water molecules might aid in the recognition of the peptide by the protease or be crucial for product release once the substrate is cleaved. These waters might as well be structurally a part of the substrate envelope. The analysis of these water molecules in the already simulated structures might help elucidate the structural basis of their conservation through different substrate-protease complexes.

In order to make sure that the LP1'F mutation at the p1-p6 substrate is structurally related to the D30N/N88D double mutations, simulations of wild-type protease in complex with a LP1'F p1p6 substrate variant should be studied. This will show the necessity of LP1'F mutation to compensate the affinity loss to the substrate caused by D30N/N88D mutations in the protease and to show that LP1'F is less likely to be seen in a WT protease.

The RMSD plot of ma-ca (Figure 3.2) has a peak at around 9^{th} ns. This brings up the possibility that other peaks may be observed in any of the simulations in this study. Therefore, the structures should be simulated for longer times in order to get

more reliable results.

The MD simulations of the drug-bound WT protease structures enable to define the dynamic inhibitor envelope. Since it has been observed that the static and dynamic substrate envelopes are different from each other in terms of volume size, it may not be meaningful to compute the fit of crystal structures of the inhibitors into the dynamic substrate envelope, because the inhibitor envelope is likely to change dynamically. Once the dynamic inhibitor envelope is computed, the fit of the inhibitor densities into the dynamic substrate envelope may reveal the potential protease positions where the drug-resistant mutation may occur.

APPENDIX A: SAMPLING

Two different coordinate sets of the protease-substrate complex systems are extracted from each trajectory. In the first one, data is recorded at every 20 ps where in the second one at every 0.4 ps the coordinates are saved. This chapter includes RMSD and potential energy plots for the two coordinates sets in order to show that the first set represents the second one for the substrate envelope modelling analysis.

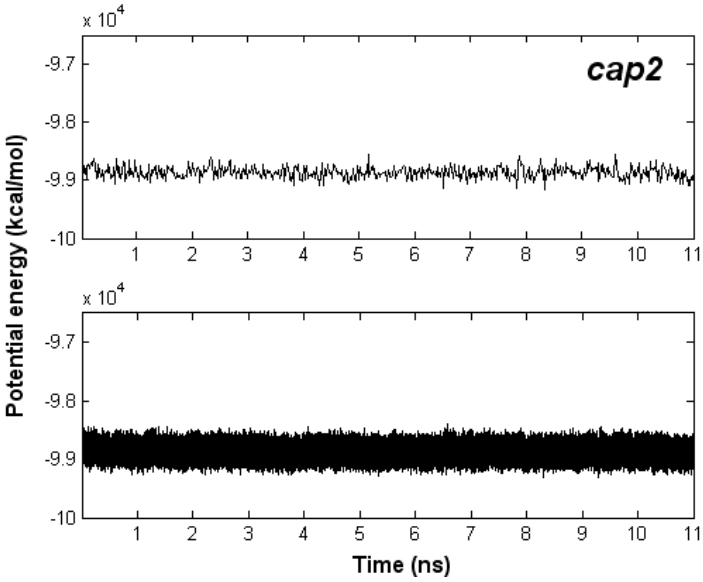


Figure A.1. Potential energy curve for WT protease-ca-p2 complex simulation

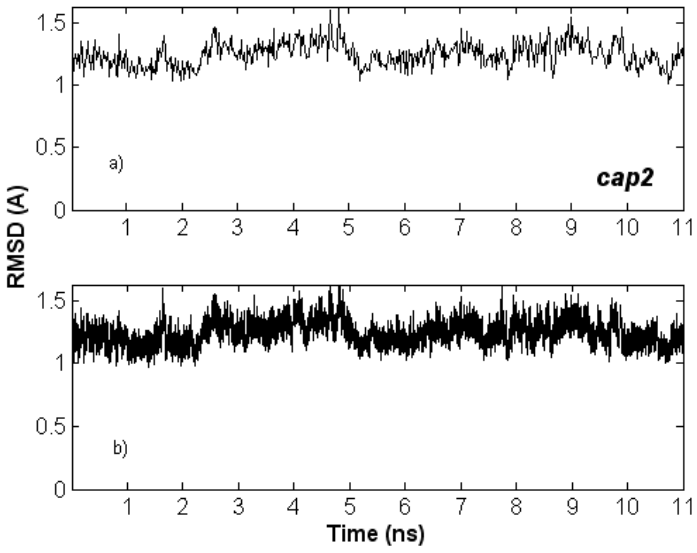


Figure A.2. Potential energy curve for WT protease-ca-p2 complex simulation

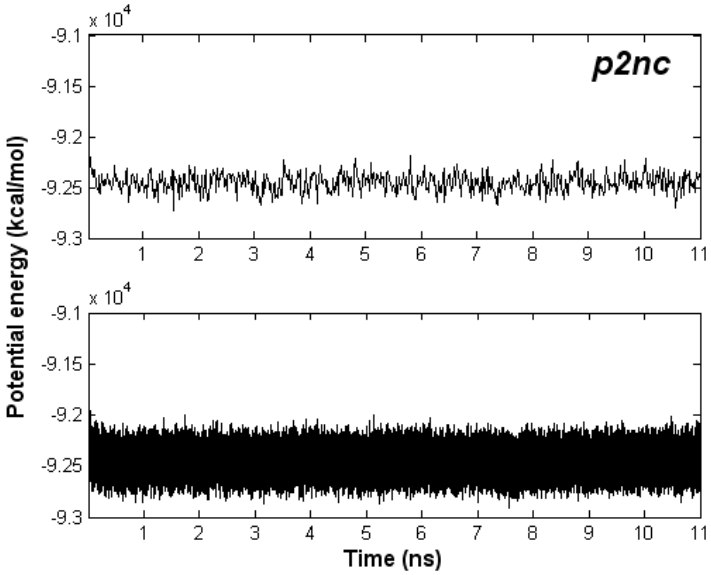


Figure A.3. Potential energy curve for WT protease-p2-nc complex simulation

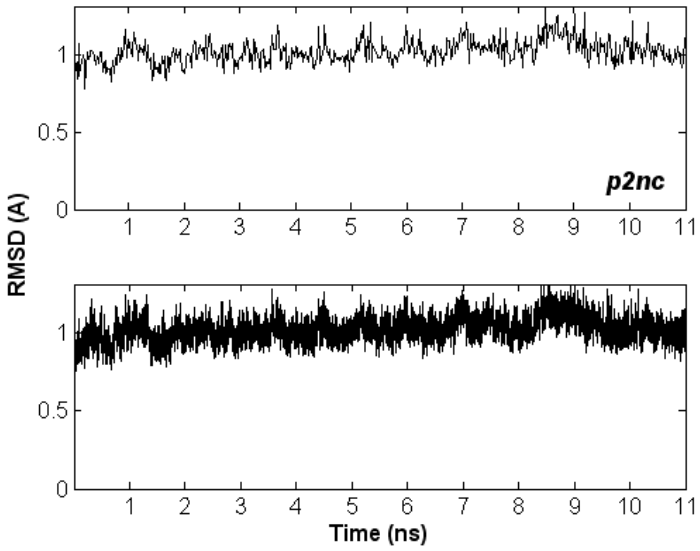


Figure A.4. Potential energy curve for WT protease-p2-nc complex simulation

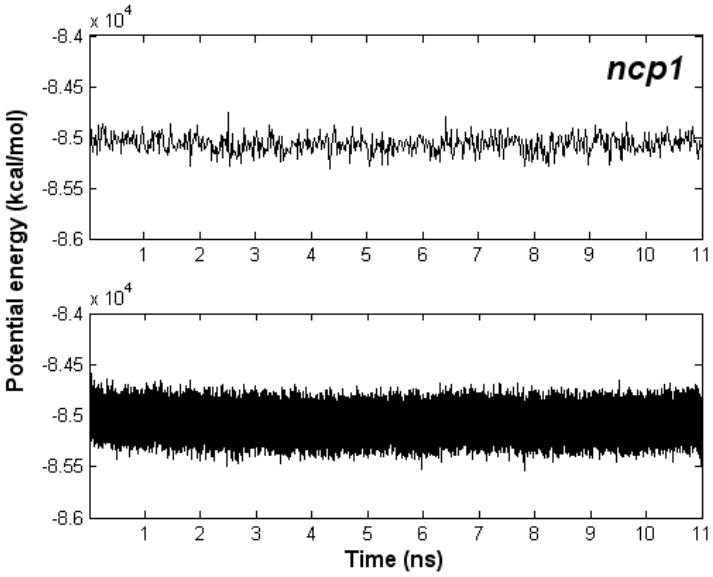


Figure A.5. Potential energy curve for WT protease-nc-p1 complex simulation

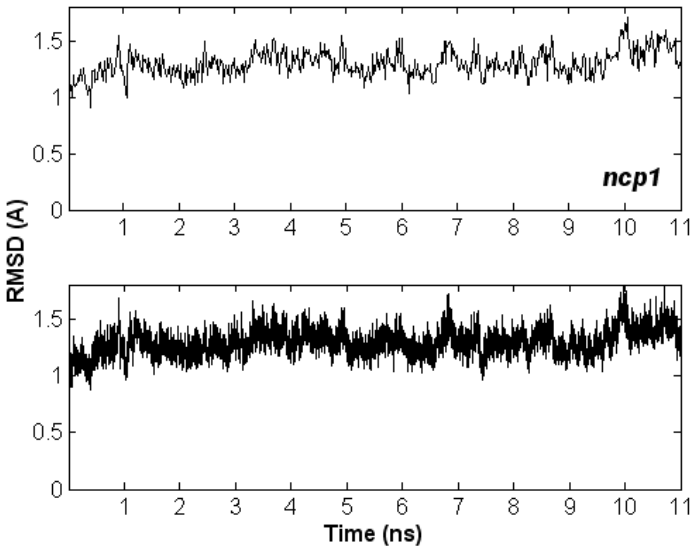


Figure A.6. Potential energy curve for WT protease-nc-p1 complex simulation

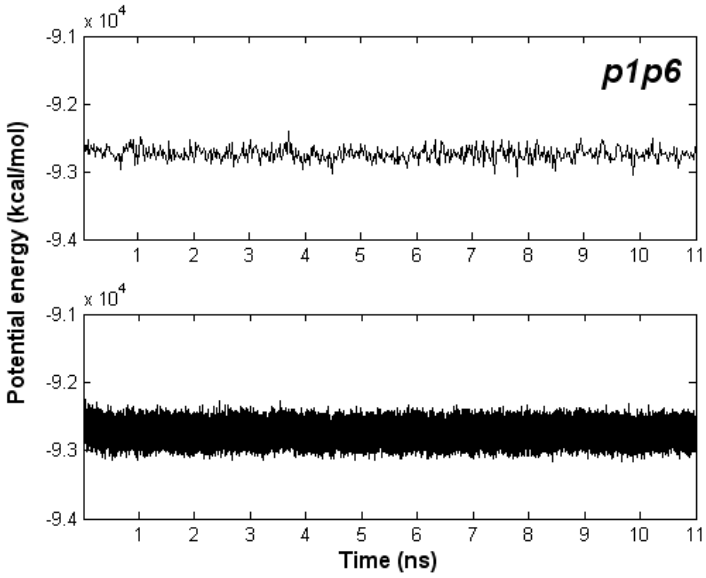


Figure A.7. Potential energy curve for WT protease-p1-p6 complex simulation

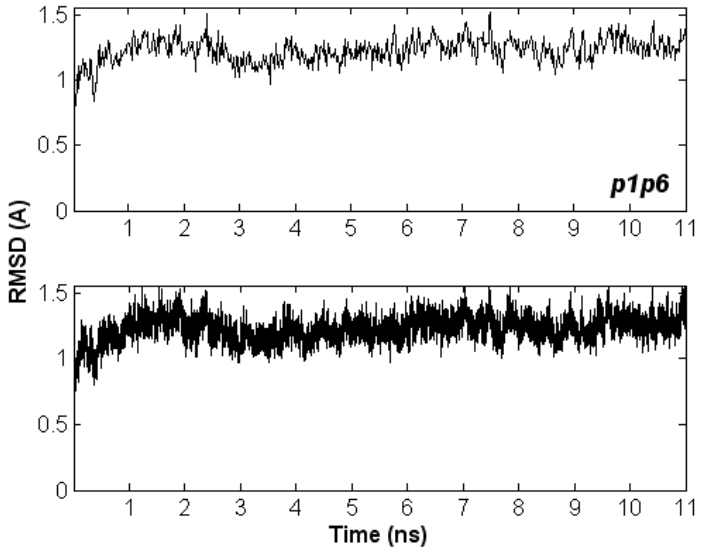


Figure A.8. Potential energy curve for WT protease-p1-p6 complex simulation

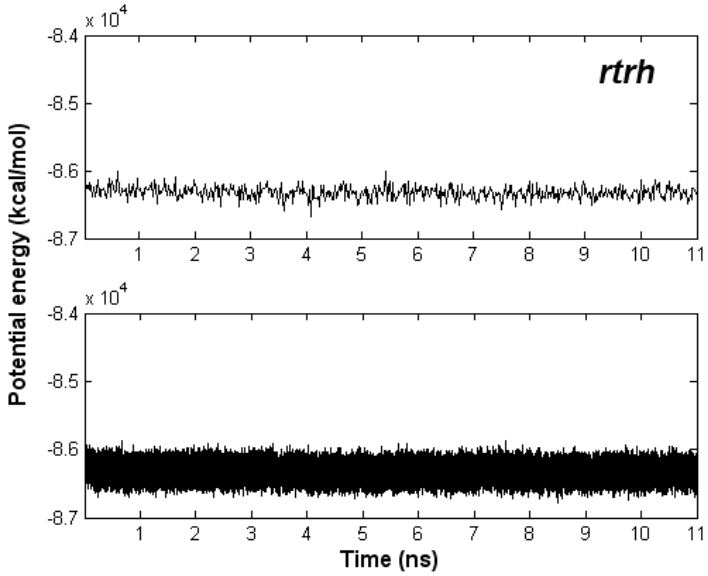


Figure A.9. Potential energy curve for WT protease-rt-rh complex simulation

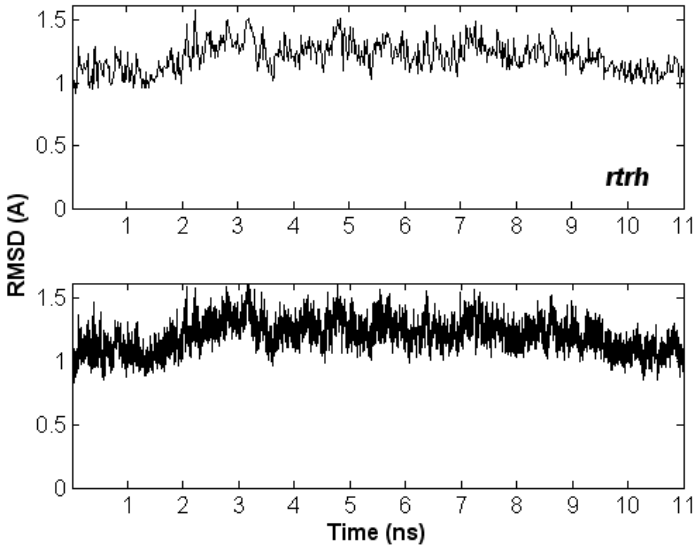


Figure A.10. Potential energy curve for WT protease-rt-rh complex simulation

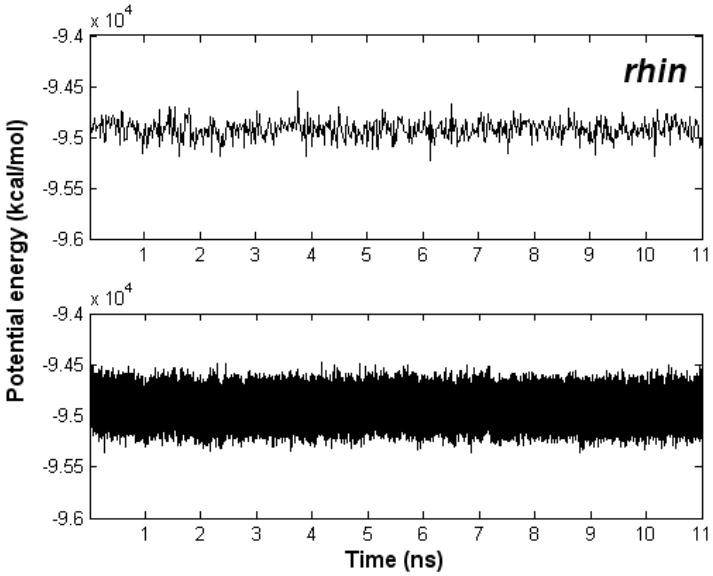


Figure A.11. Potential energy curve for WT protease-rh-in complex simulation

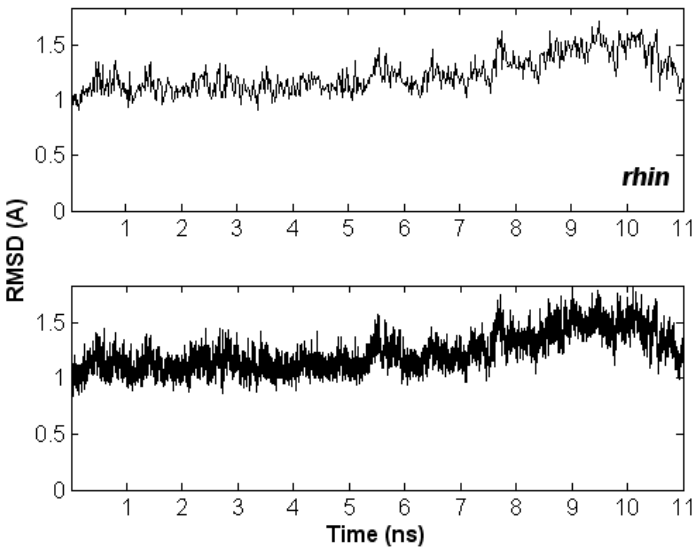


Figure A.12. Potential energy curve for WT protease-rh-in complex simulation

APPENDIX B: VAN DER WAALS CONTACTS

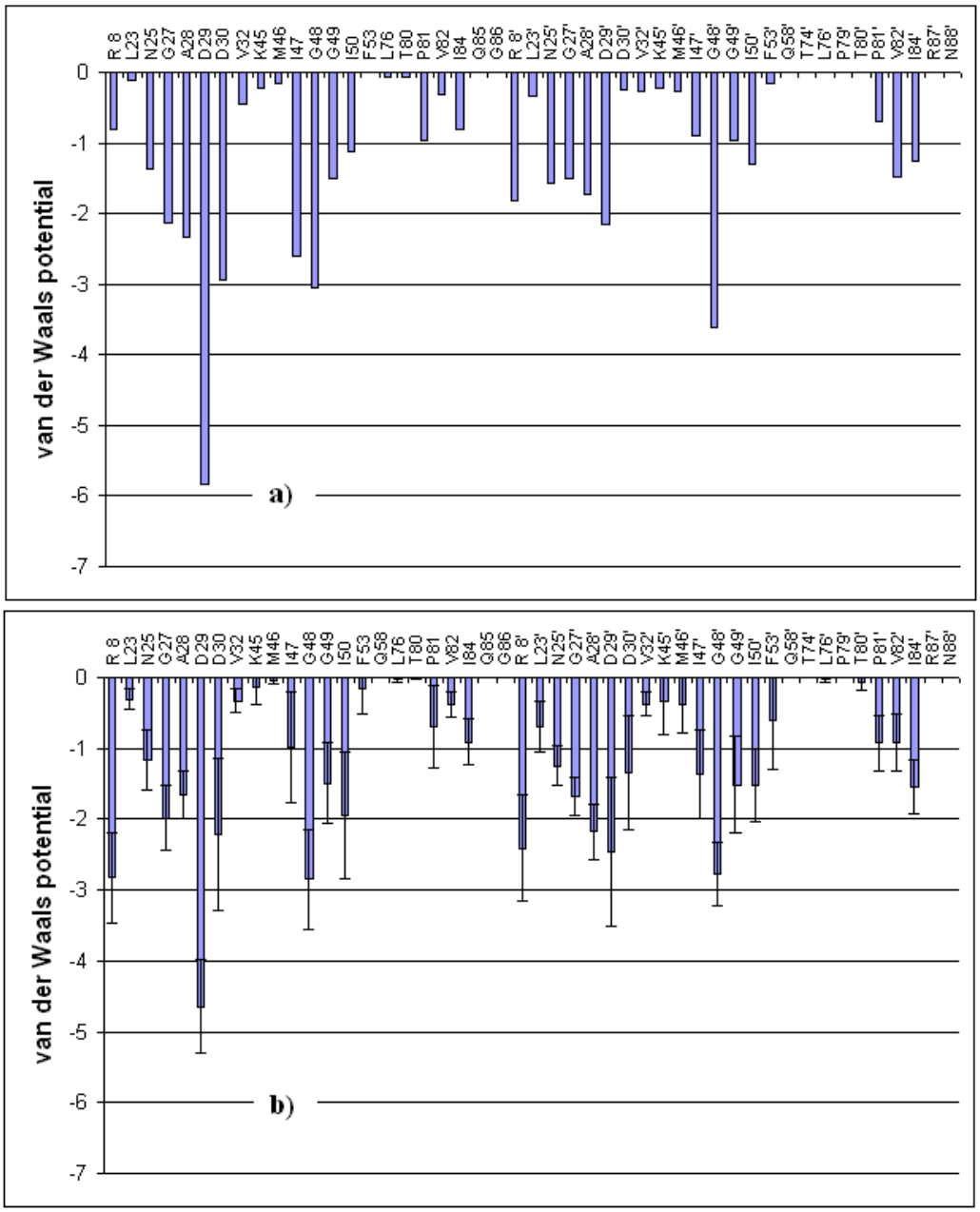


Figure B.1. a) Crystal and b) Dynamic van der Waals contacts of ma-ca substrate with the wild-type protease.

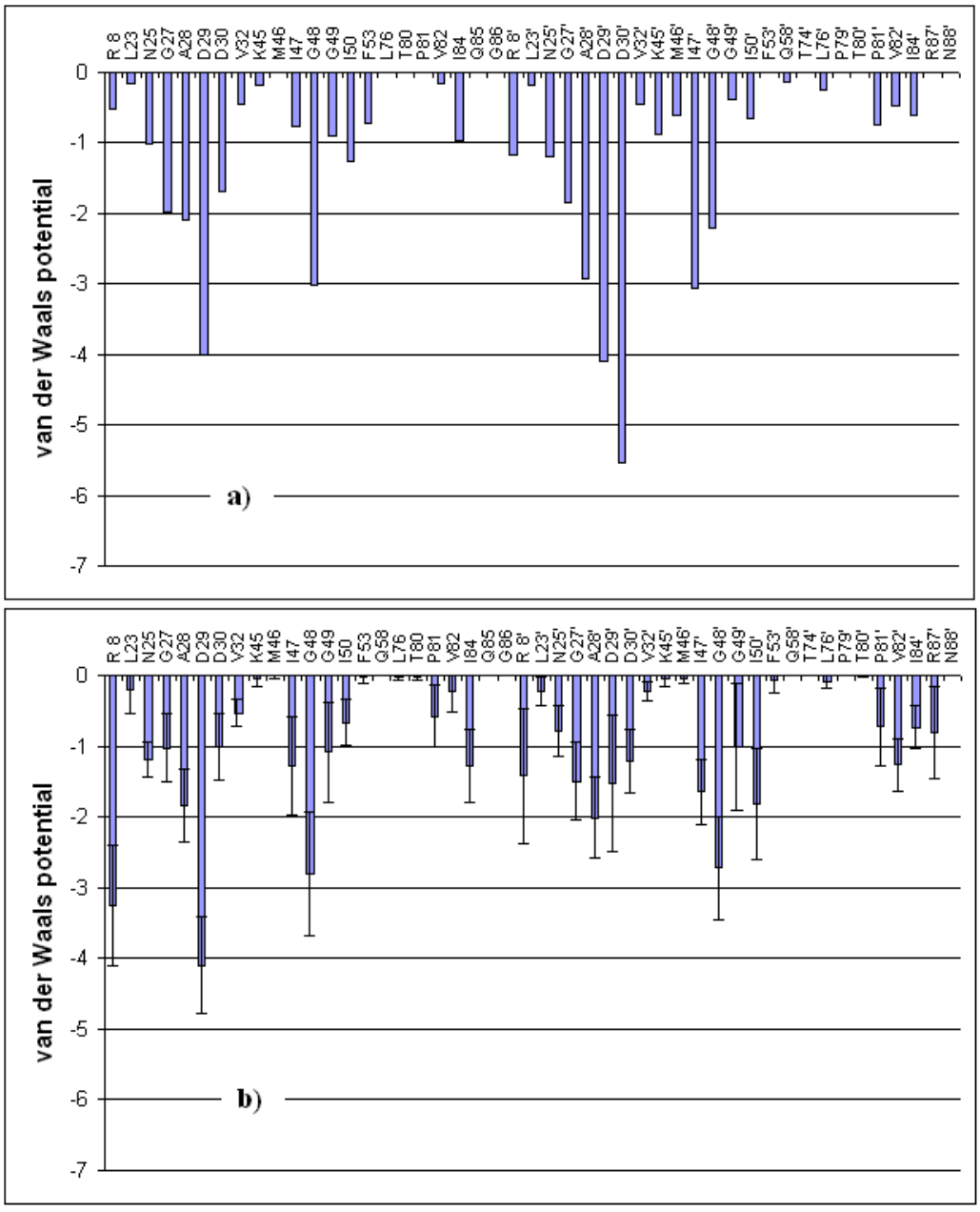


Figure B.2. a) Crystal and b) Dynamic van der Waals contacts of ca-p2 substrate with the wild-type protease.

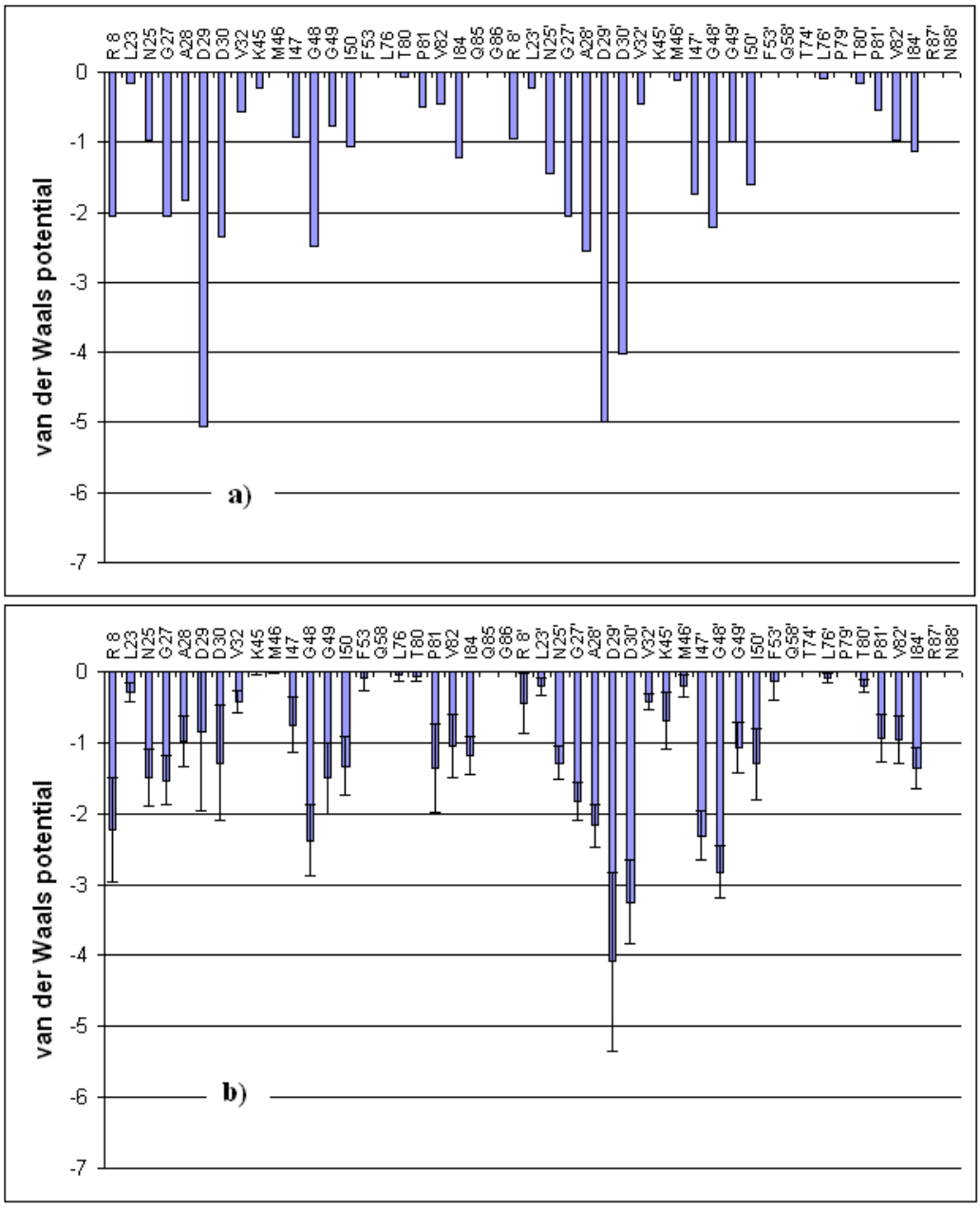


Figure B.3. a) Crystal and b) Dynamic van der Waals contacts of p2-nc substrate with the wild-type protease.

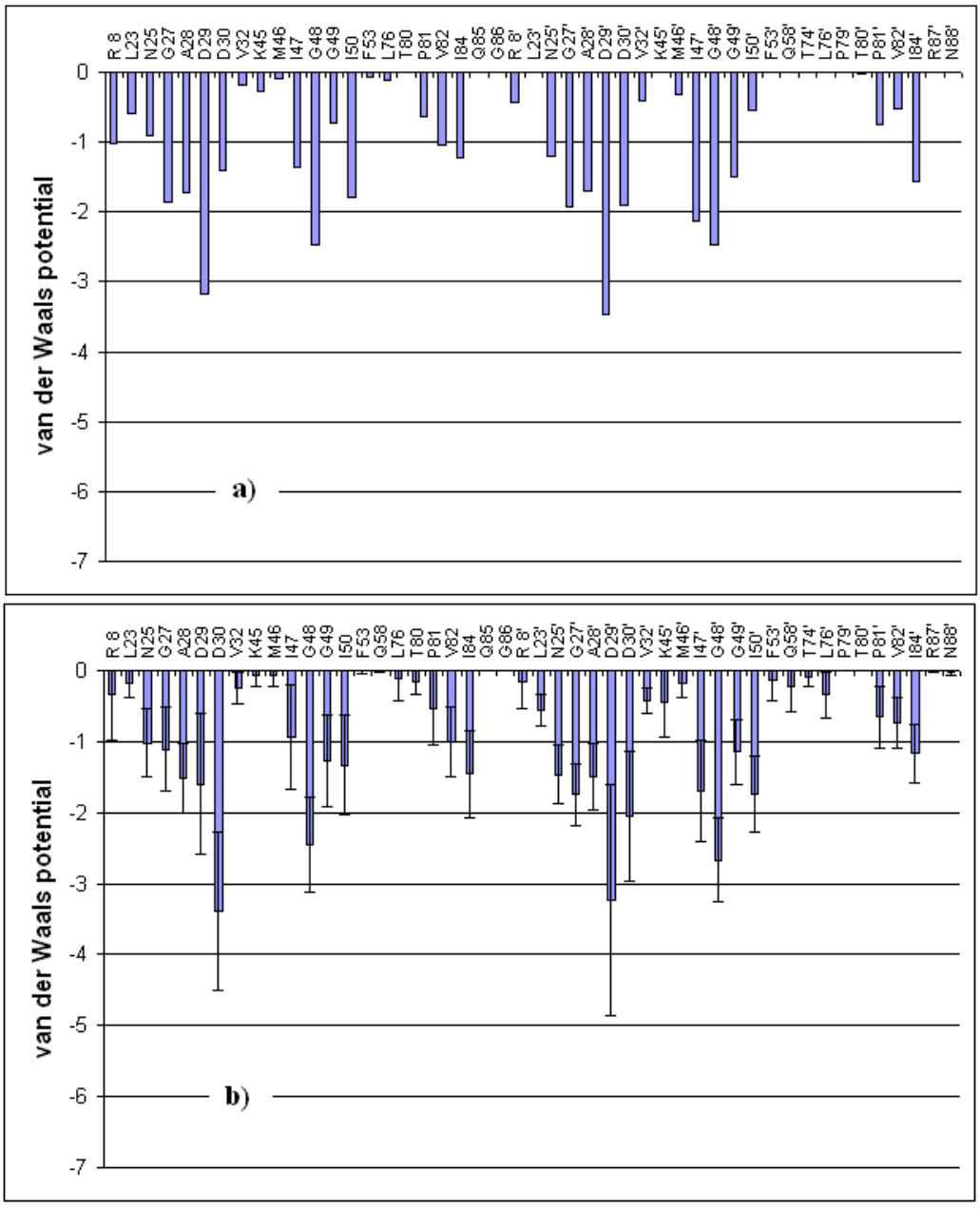


Figure B.4. a) Crystal and b) Dynamic van der Waals contacts of nc-p1 substrate with the wild-type protease.

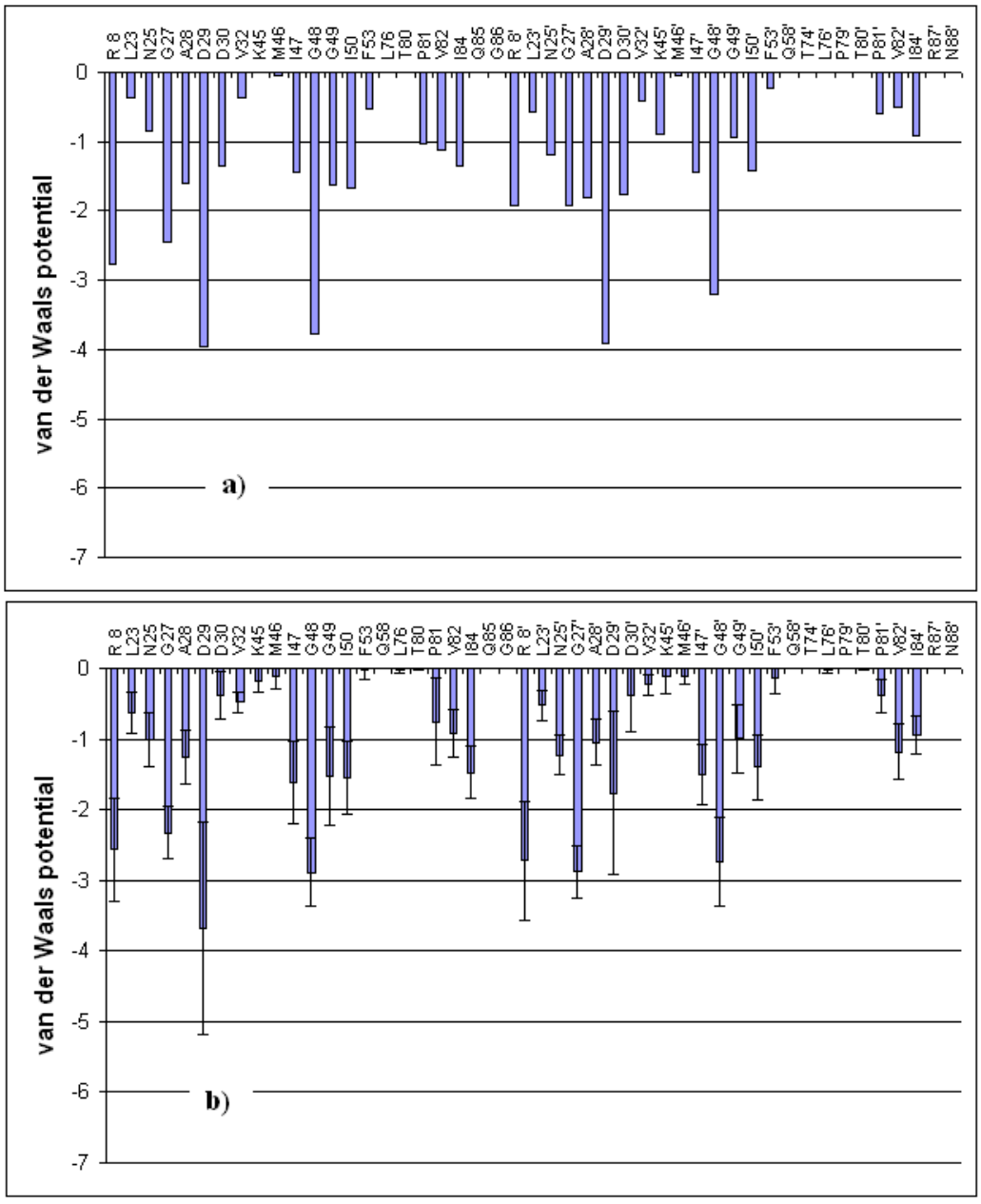


Figure B.5. a) Crystal and b) Dynamic van der Waals contacts of rt-rh substrate with the wild-type protease.

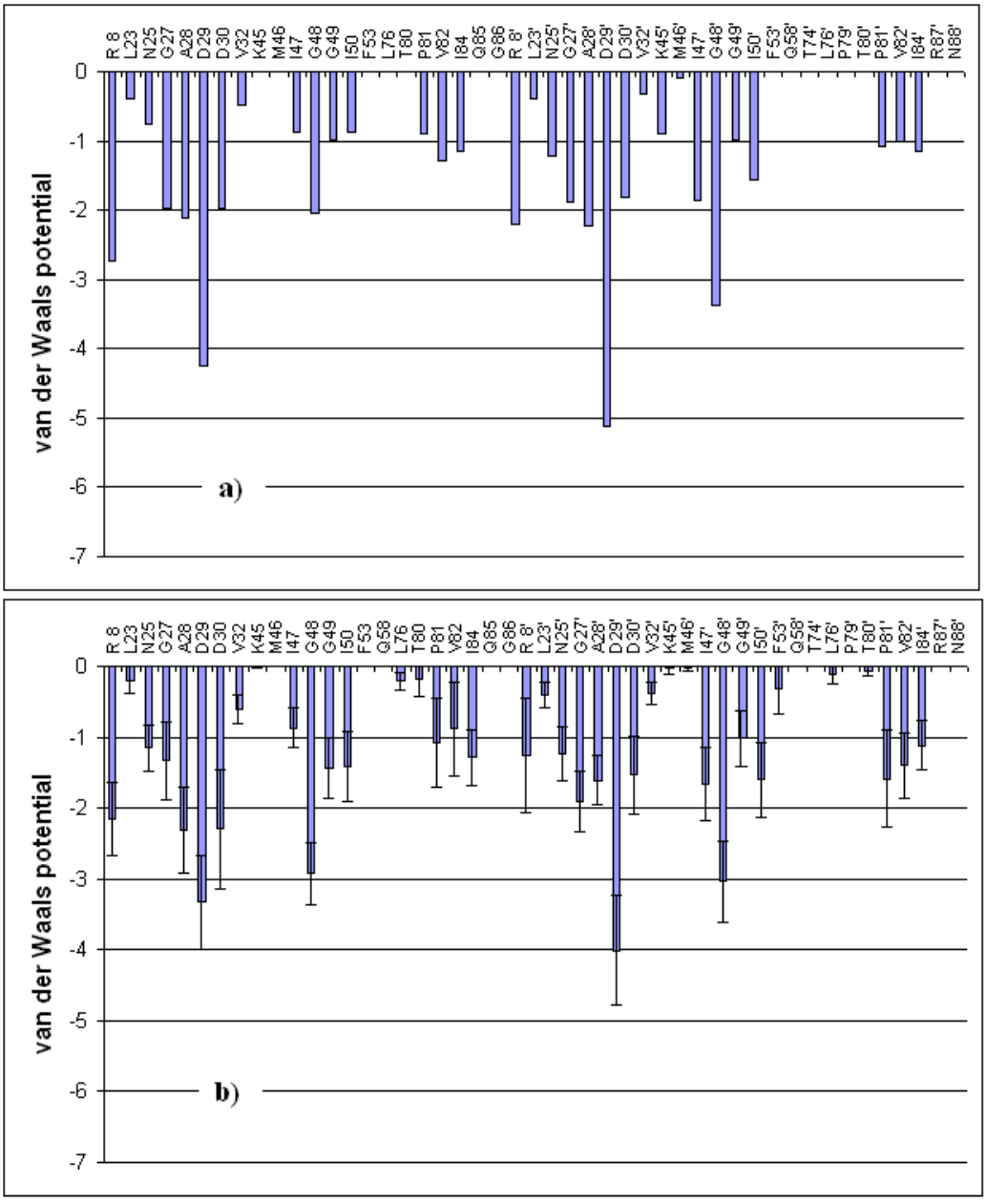


Figure B.6. a) Crystal and b) Dynamic van der Waals contacts of rh-in substrate with the wild-type protease.

APPENDIX C: ATOMIC POSITIONAL FLUCTATIONS

This chapter includes the time evolution of the cross-correlations of the atomic positional fluctuations of nonhydrogen protein atoms. The cross-correlations of the substrate residues at the sites from P4 to P4' are given.

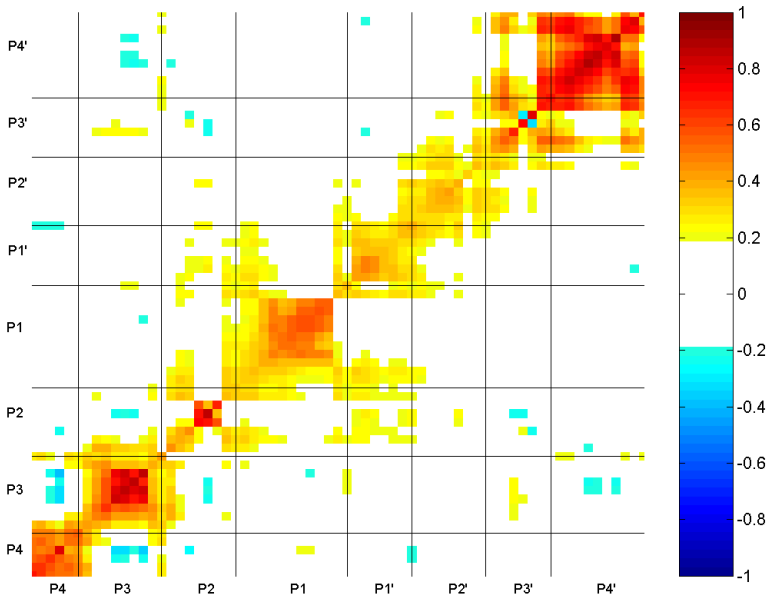


Figure C.1. Ma-ca substrate nonhydrogen atoms ($\tau = 20$ ps).

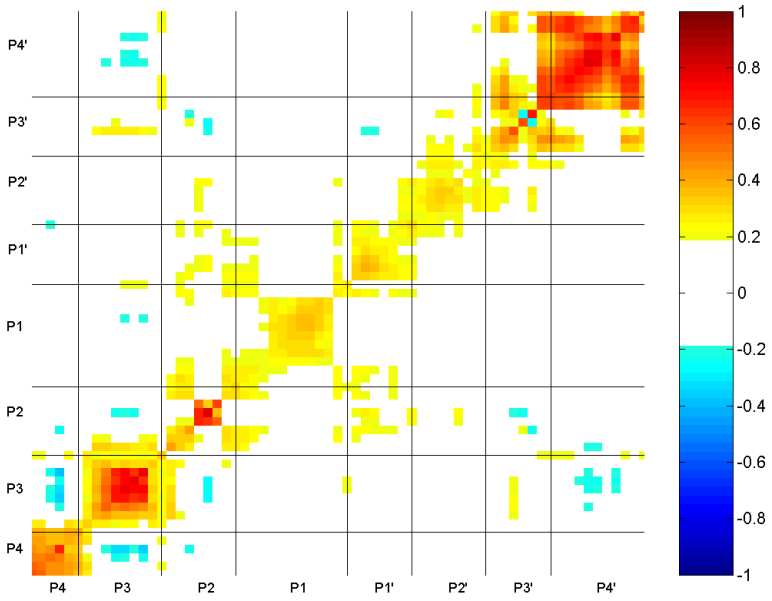


Figure C.2. Ma-ca substrate nonhydrogen atoms ($\tau = 100$ ps).

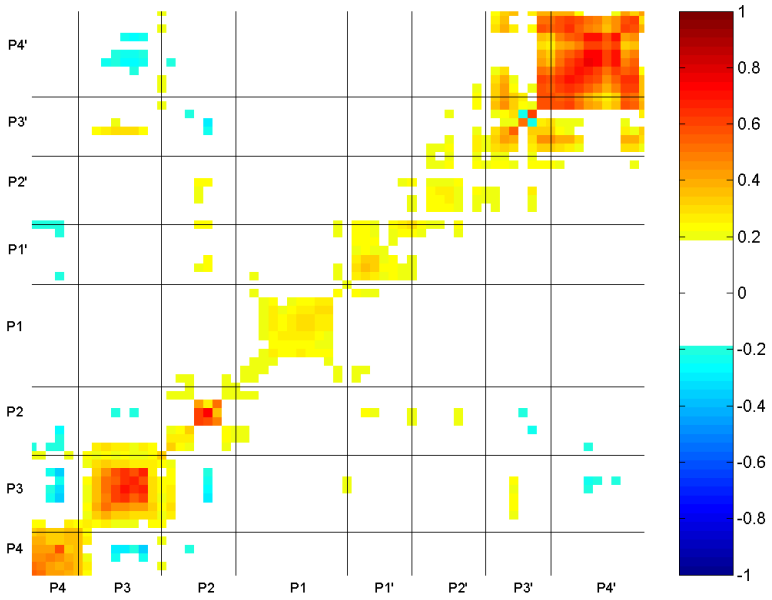


Figure C.3. Ma-ca substrate nonhydrogen atoms ($\tau = 200$ ps).

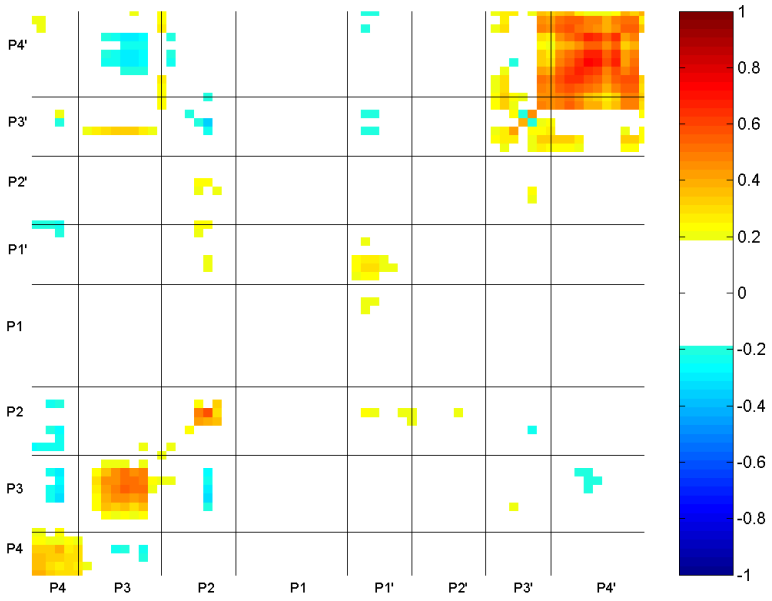


Figure C.4. Ma-ca substrate nonhydrogen atoms ($\tau = 500$ ps).

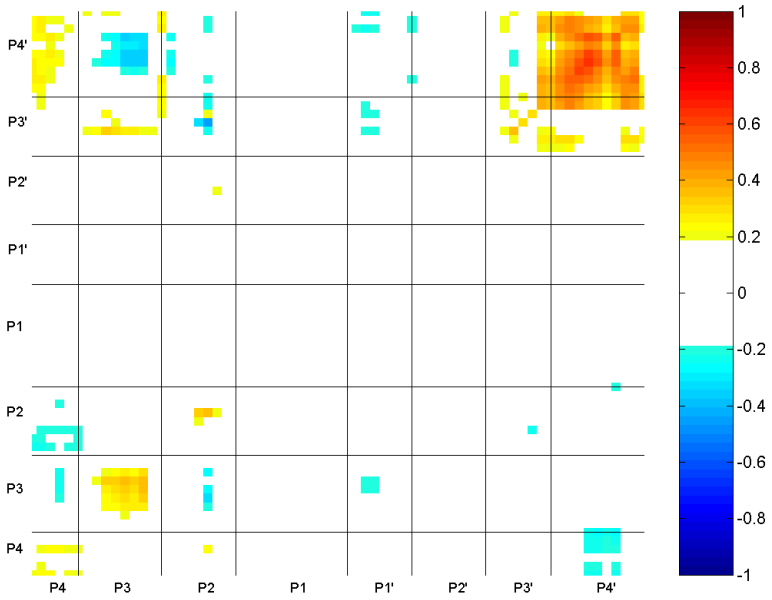


Figure C.5. Ma-ca substrate nonhydrogen atoms ($\tau = 1100$ ps).

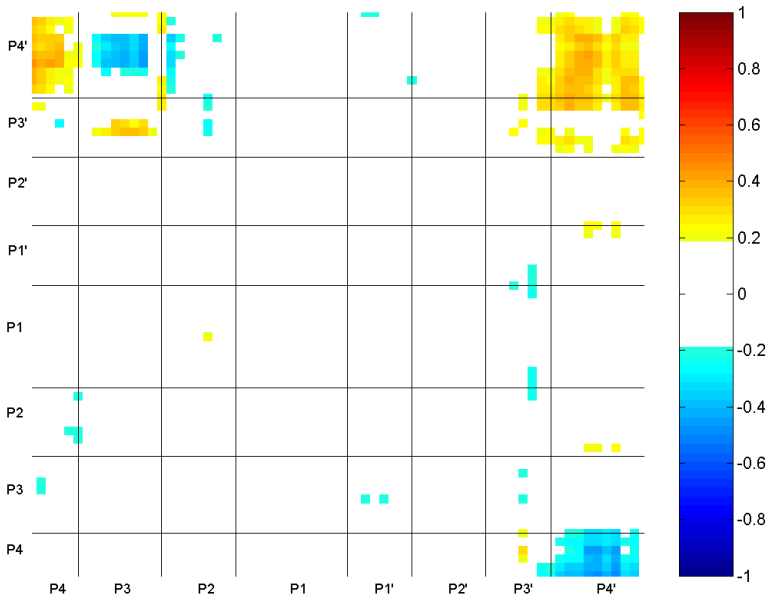
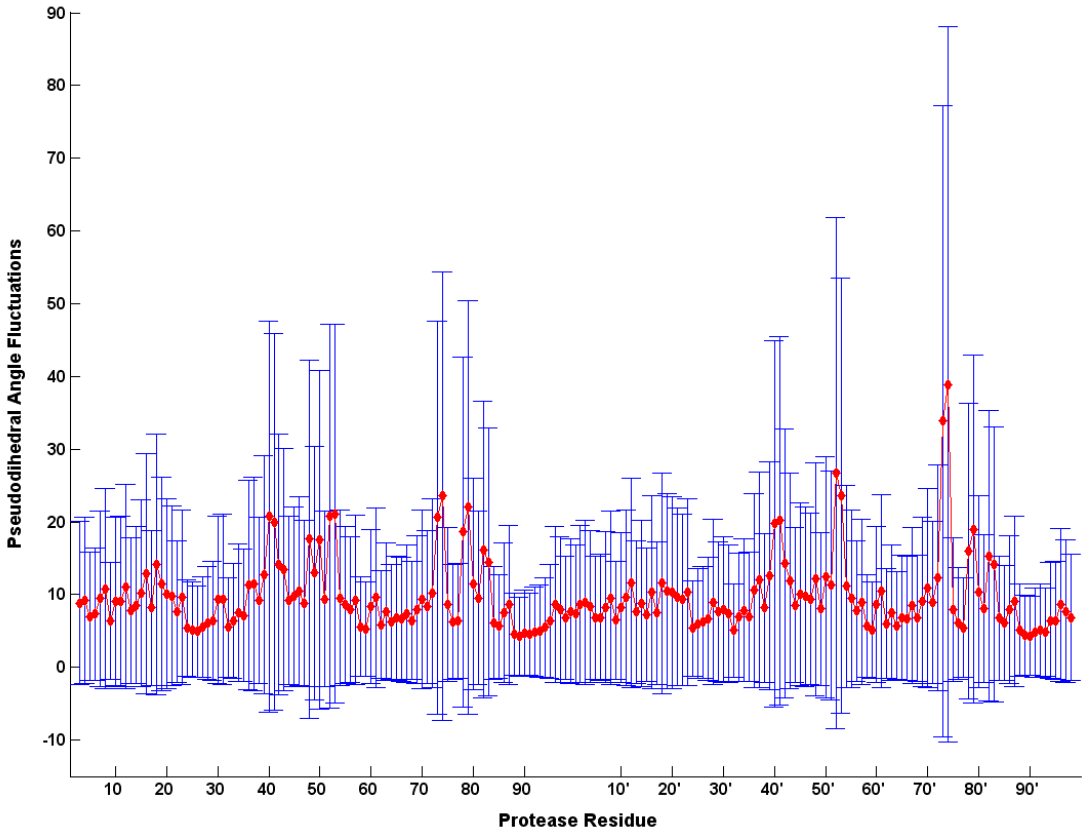


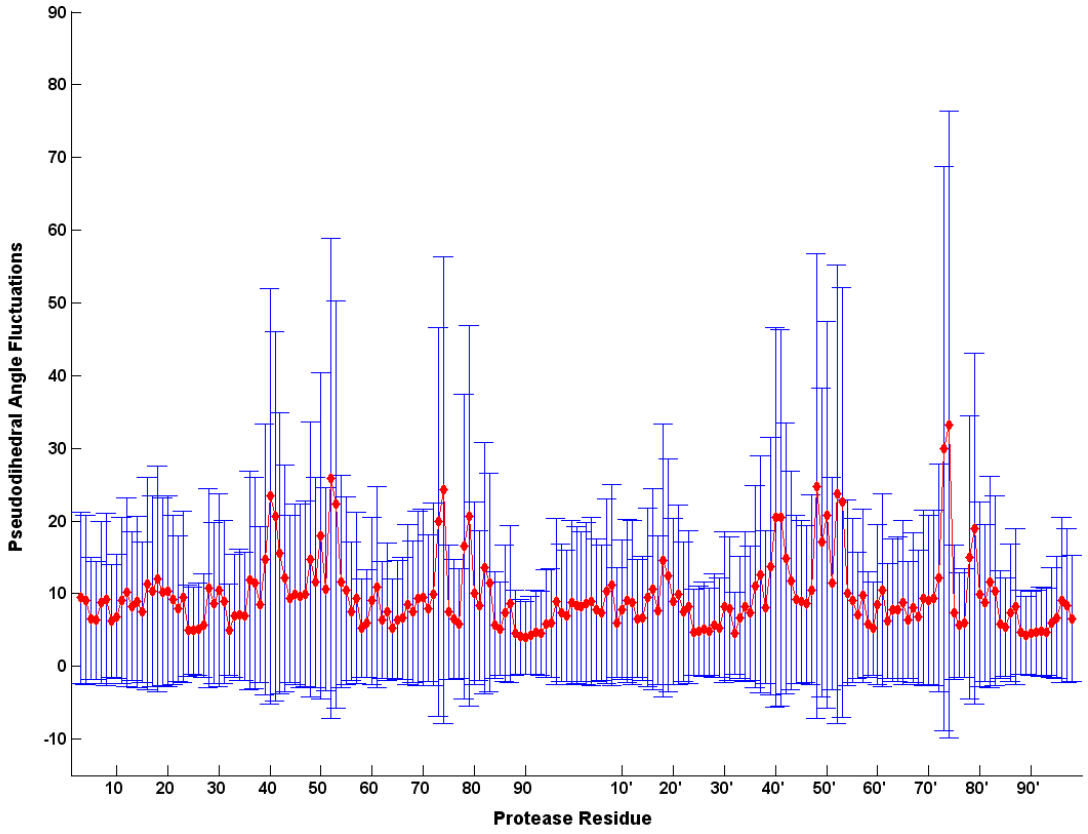
Figure C.6. Ma-ca substrate nonhydrogen atoms ($\tau = 2200$ ps).

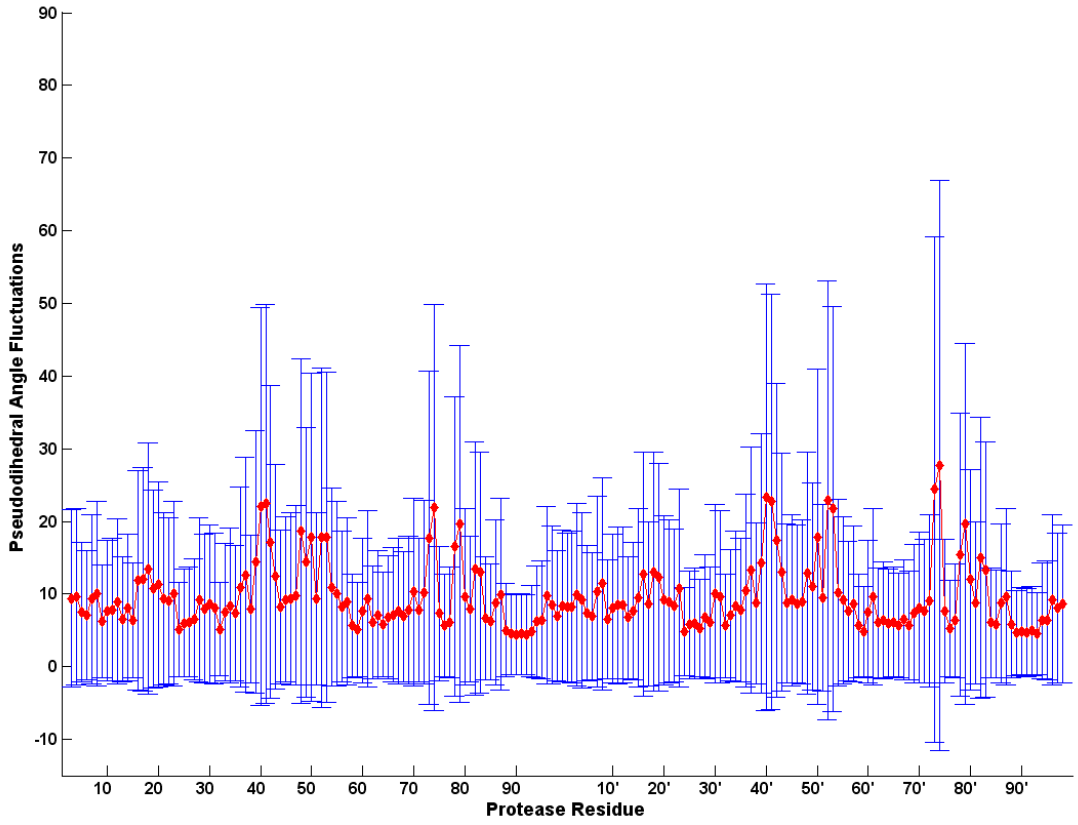
APPENDIX D: PSEUDODIHEDRAL ANGLE FLUCTUATIONS

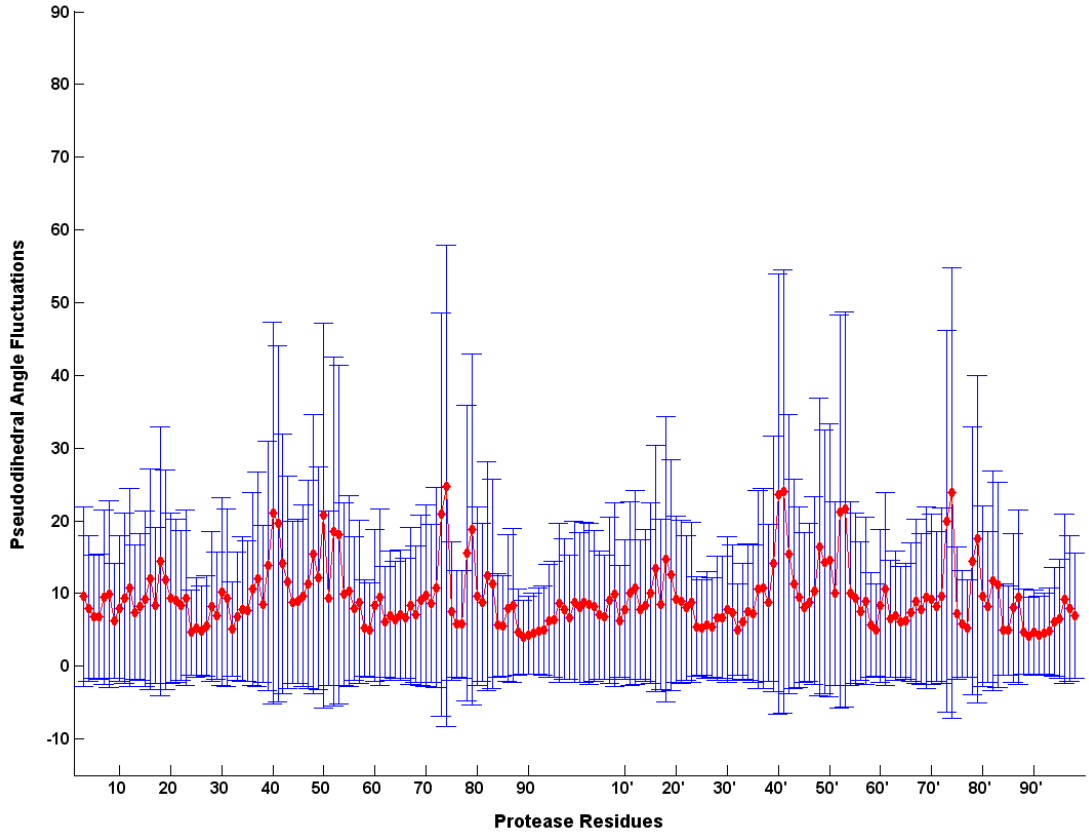
This chapter contains the pseudodihedral angle fluctuations of each protease residue in different substrate complexes.

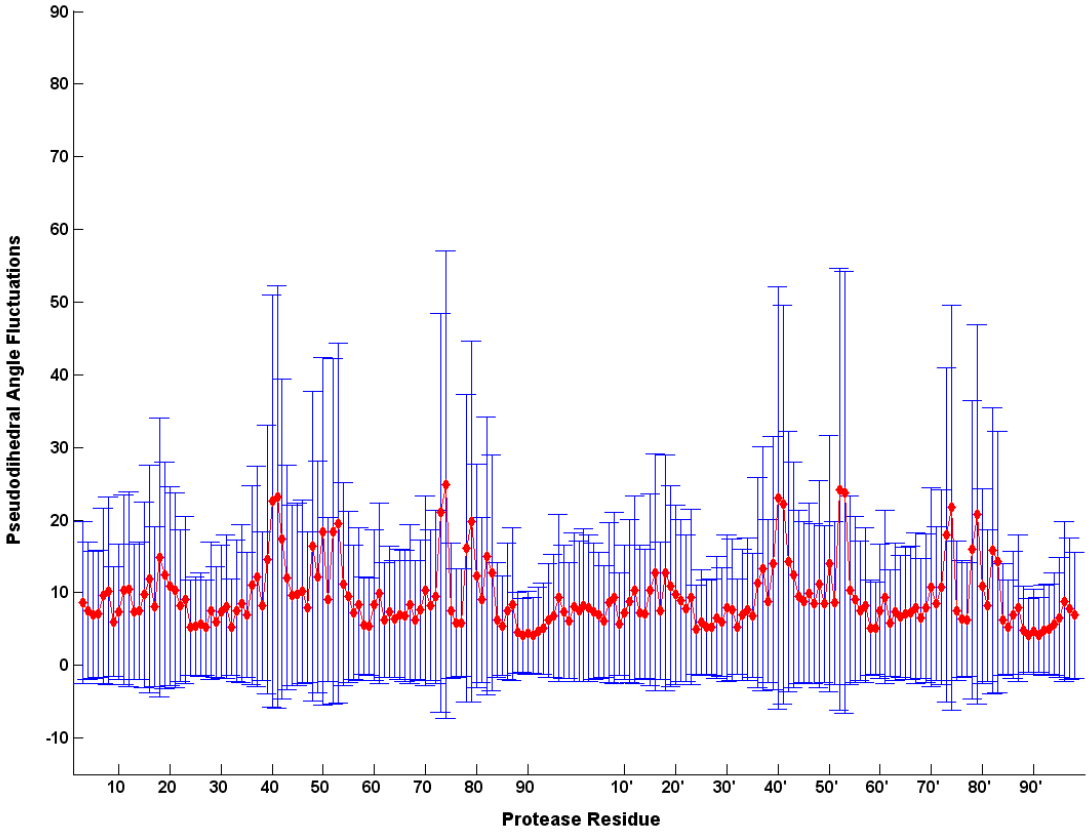
red

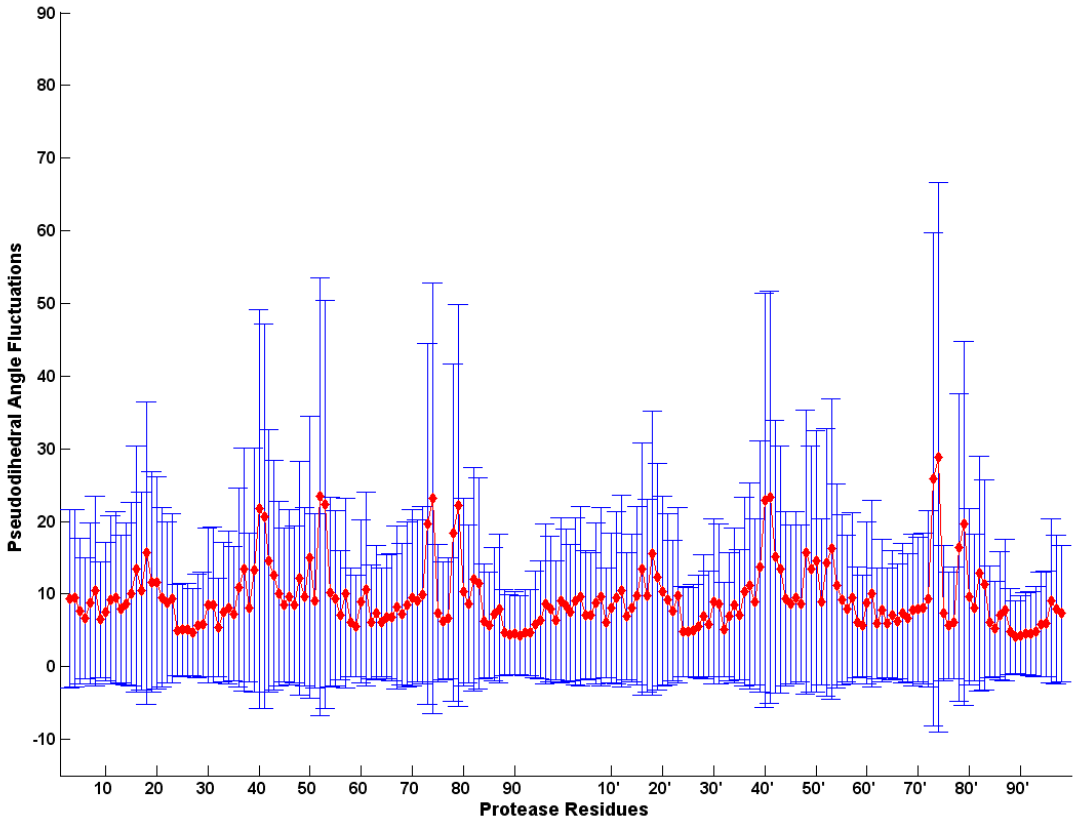



Pseudodihedral Angle Fluctuations
90
80
70
60
50
40
30
20
10
0
-10
10 20 30 40 50 60 70 80 90 10' 20' 30' 40' 50' 60' 70' 80' 90'
Protease Residue


Pseudodihedral Angle Fluctuations
90
80
70
60
50
40
30
20
10
0
-10
10 20 30 40 50 60 70 80 90 10' 20' 30' 40' 50' 60' 70' 80' 90'
Protease Residue


Pseudodihedral Angle Fluctuations
90
80
70
60
50
40
30
20
10
0
-10
10 20 30 40 50 60 70 80 90 10' 20' 30' 40' 50' 60' 70' 80' 90'
Protease Residues


Pseudodihedral Angle Fluctuations
90
80
70
60
50
40
30
20
10
0
-10
10 20 30 40 50 60 70 80 90 10' 20' 30' 40' 50' 60' 70' 80' 90'
Protease Residue


Pseudodihedral Angle Fluctuations
90
80
70
60
50
40
30
20
10
0
-10
10
20
30
40
50
60
70
80
90
10'
20'
30'
40'
50'
60'
70'
80'
90'
Protease Residues

REFERENCES

AIDS Epidemic Update: December 2007, UNAIDS and The World Health Organization, 2007.

Berendsen, H. J. C., J.P.M. Postma, W.F. Van Gunsteren, A. DiNola and J.R. Haak , 1984, "Molecular dynamics with coupling to an external bath", *Journal of Chemical Physics*, Vol. 81, pp. 3684-3690.

Barre-Sinoussi, F., J. C. Chermann, F. Rey, M. T. Nugeyre, S. Chamaret, J. Gruest, C. Dauguet, C. Axler-Blin, F. Vezinet-Brun, C. Rouzioux, W. Rozenbaum, and L. Montagnierand , 1983, "Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS)", *Science*, Vol. 220, pp. 868-871.

Case, D.A., T.A. Darden, T.E.I. Cheatham, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, K.M. Merz, B. Wang, D.A. Pearlman et al., 2004, "AMBER 8", *Science*, University of California, San Francisco, 2004.

Case, D.A., T.E. Cheatham, T. Darden, H. Gohlke, R. Luo, K.M. Jr. Merz, A. Onufriev, C. Simmerling, B. Wang and R. Woods, 2005, "The Amber biomolecular simulation programs", *Journal of Computational Chemistry*, Vol. 26, pp. 1668-1688.

Chellappan, S., V. Kairys, M. X. Fernandes, C. Schiffer, and M. K. Gilson, 2007, "Evaluation of the substrate envelope hypothesis for inhibitors of HIV-1 protease", *PROTEINS: Structure, Function, and Bioinformatics*, Vol. 68, pp. 561-567.

Coffin, J. M., 1995, "HIV population dynamics in vivo: implications for genetic variation, pathogenesis and therapy", *Science*, Vol. 257, pp. 483-489.

Cooper, J. B., 2002, "Aspartic Proteinases in Diseases: A Structural Perspective", *Current Drug Targets*, Vol. 3, pp. 155-173.

DeLano, W. L., 2002, *The PyMOL Molecular Graphics System*, http://www.pymol.org.

Essman, U., L. Perera, M.L. Berkowitz, T.A. Darden, H. Lee and L.G. Pedersen, 1995, "A smooth Particle Mesh Ewald method", *Journal of Chemical Physics*, Vol. 103, pp. 8577-8593.

Feher, A., I. T. Weber, P. Bagossi, P. Baross, B., Mahalingam, J. M. Louis, T. D. Copeland, I. Y. Yorshin, R. W. Harrison, and J. Tozser, 2002, "Effect of sequence polymorphism and drug-resistance on two HIV-1 Gag processing sites", *Journal of Biochemistry*, Vol. 269, pp. 4114-4120.

Flexner, C., 1998, "HIV-protease inhibitors", *The New England Journal of Medicine*, Vol. 338, No. 18, pp. 1281-1292.

Flory, P. J., 1969, "*Statistical mechanics of chain molecules*", John Wiley & Sons Inc.

Frenkel, D. and B. Smit, 1996, "Understanding molecular simulations: from algorithms to applications", *Academic Press.*

Hoggs R. S., K. V. Heath, B. Yip, K. J. P. Craib, M., V., O'Shaughnessy, M. T. Schechter, and J. S. G. Montaner, 1998, "Improved survival among HIV-infected individuals following initiation of antiretroviral therapy ", *Journal of the American Medical Association*, Vol. 279, No. 6, pp. 450-454.

Humphrey, W., A. Dalke, and K. Schulten, 1996, "VMD - Visual Molecular Dynamics", *Journal of Molecular Graphics*, Vol. 14, No. 1, pp. 33-38.

Jorgensen, W. L., J. Chandrasekhar, J.D. Madura, R.W. Impey and M.L. Klein, 1983, "Comparison of simple potential functions for simulating liquid water", *Journal of Chemical Physics*, Vol. 79, pp. 926-935.

King N., M. Prabu-Jeyabalan, E. A. Nalivaika, and C. A. Schiffer, 2004, "Combating Susceptibility to Drug Resistance: Lessons from HIV-1 Protease", *Chemistry & Biology*, Vol. 11, pp. 1333-1338.

Kirkpatrick, P., 2004, "Inside the envelope", *Nature Reviews Drug Discovery*, Vol. 3, pp. 1000.

Kolli M., S. Lastere, and C. A. Schiffer, 2006, "Co-evolution of nelfinavir-resistant HIV-1 protease and the p1-p6 substrate", *Virology*, Vol. 347, pp. 405-409.

Kozisek M., J. Bray, P. Rezacova, K. Saskova, J. Brynda, J. Pokorna, F. Mammano, L. Rulisek, and J. Konvalinka, 2007, "Molecular analysis of the HIV-1 resistance development: enzymatic activites, crystal structures, and thermodynamics of nelfinavir-resistant HIV protease mutants", *Journal of Molecular Biology*, Vol. 374, pp. 1005-1016.

Leach, A. R., 2001, "Molecular modelling: principles and applications (Second edition)", *Prentice Hall.*

Li, J. and L. A. Loeb, 1992, "Fidelity of HIV-1 reverse transcriptase copying RNA in vitro", *Biochemistry*, Vol. 31, pp. 954-958.

McDonald, C. K. and D. R. Kuritkzkes, 1997, "Human immunodeficiency virus type 1 protease inhibitors", *Archives of Internal Medicine*, Vol. 157, No. 9, pp. 951-959.

Miller, M, J. Schneider, B. K. Sathyanarayana, M. V. Toth, G. R. Marshall, L. Clawson, L. Selk, S. B. Kent, and A. Wlodawer, 1989, "Structure of complex of synthetic HIV-1 protease with a substrate-based inhibitor at 2.3 Å resolution", *Science*, Vol. 246, pp. 1149-1152.

Navia, M. A., P. M. D. Fitzgerald, B. M. McKeever, C-T Leu, J. C. Heimbach, W. K. Herber, I. S. Sigal, P. L. Darke, and, J. P. Springer, 1989, " Three-dimensional structure of aspartyl protease from human immunodeficiency virus HIV-1", *Nature*, Vol. 337, pp. 615-620.

Prabu-Jeyabalan, M., E. Nalivaika, and C. A. Schiffer, 2000, "How does a symmetric dimer recognize an asymmetric substrate? A substrate complex of HIV-1 pro-

tease.", *Journal of Molecular Biology*, Vol. 301, pp. 1207-1220.

Prabu-Jeyabalan, M., E. Nalivaika, and C. A. Schiffer, 2002, "Substrate shape determines specificity of recognition for HIV-1 protease: analysis of crystal structures of six substrate complexes", *Structure*, Vol. 10, pp. 369-381.

Prabu-Jeyabalan, M., E. Nalivaika, N. M. King, and C. A. Schiffer, 2004, "Structural basis for coevolution of a human immunodeficiency virus type 1 nucleocapsid-p1 cleavage site with a V82A drug-resistant mutation in viral protease", *Journal of Virology*, Vol. 78, No. 22, pp. 12446-12454.

Roberts J. D., K. Bebenek, and T. A. Kunkel, 1988, "The accuracy of reverse transcriptase from HIV-1", *Science*, Vol. 242, pp. 1171-1173.

Roberts J. D., B. D. Preston, L. A. Johnston, A. Soni, L. A. Loeb, and T. A. Kunkel, 1989, "Fidelity of two retroviral reverse transcriptases during DNA-dependent DNA synthesis in vitro", *Molecular and Cellular Biology*, Vol. 9, pp. 469-476.

Ryckaert, J. P., G. Ciccotti and H.J.C. Berendsen, 1977, "Numerical integration of the Cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes", *Journal of Computational Physics*, Vol. 23, pp. 327-341.

Wlodawer, A., M. Miller, M. Jaskolski, B. K. Sathyanarayana, E. Baldwin, I. T. Weber, L. M. Selk, L. Clawson, J. Schneider, and S. B. Kent, 1989, "Conserved folding in retroviral proteases: crystal structure of a synthetic HIV-1 protease", *Science*, Vol. 245, pp. 616-621.

Wlodawer, A. and J. Erickson, 1993, "Structure-based inhibitors of HIV-1 protease", *Annual Review of Biochemistry*, Vol. 62, pp. 543-585.

Yao, N., P. Reichert, S. S. Taremi, W. W. Prosise, and P. C. Weber, 1999, "Molecular views of viral polyprotein processing revealed by the crystal structure of the hepatitis C virus bifunctional protease-helicase", *Structure*, Vol. 7, No. 11, pp. 1353-1363.

REFERENCES NOT CITED

Branden, C. and J. Tooze, 1998, "Introduction to Protein Structure (Second Edition)", *Garland Publishing, Inc.*

Chellappan, S., G. S. K. Kumar, A. Ali, M. N. Nalam, S. G. Anjum, H. Cao, V. Kairys, M. X. Fernandes, M. D. Altman, B. Tidor, T. M. Rana, C. A. Schiffer, and M. K. Gilson, 2007, "Design of mutation-resistant HIV protease inhibitors with the substrate envelope hypothesis", *Chemical Biology & Drug Design*, Vol. 69, pp. 298-313.

Foulkes-Murzycki, J. E., W. R. P. Scott, and C. A. Schiffer, 2007, "Hydrophobic sliding: a possible mechanism for drug resistance in human immunodeficiency virus type 1 protease", *Structure*, Vol. 15, No. 2, pp. 225-233.

King N., M. Prabu-Jeyabalan, E. A. Nalivaika, P. Wigerinck, M-P. de Bethune, and C. A. Schiffer, 2004, "Structural and thermodynamic basis for the binding of TMC114, a next-generation human immunodificiency virus type 1 protease inhibitor", *Journal of Virology*, Vol. 78, No. 21, pp. 12012-12021.

Prabu-Jeyabalan, M., E. A. Nalivaika, N. M. King, and C. A. Schiffer, 2003, "Viability of a drug-resistant human immunodeficiency virus type 1 protease variant: structural insights for better antiviral therapy", *Journal of Virology*, Vol. 77, No. 2, pp. 1306-1315.

Prabu-Jeyabalan, M., N. M. King, E. A. Nalivaika, G. Heilek-Snyder, N. Cammack, and C. A. Schiffer, 2006, "Substrate envelope and drug resistance: crystal structure of RO1 in complex with wild-type human immunodeficiency virus type 1 protease", *Antimicrobial Agents and Chemotherapy*, Vol. 50, No. 4, pp. 1518-1521.

Turner, B. G. and M. F. Summers, 1999, "Structural biology of HIV", *Journal of Molecular Biology*, Vol. 285, pp. 1-32.